Classification of Soil Contamination

Ms. Pavagada Keerthana

Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Soil pollution is can be considered to be an imbalance of chemicals in the soil at a particular site. Such unnatural contamination must be addressed to avoid hazard to the environment and inhabitants of a polluted site. Soil quality varies and soils respond differently, depending on the management inputs. Soil quality has both inherent and dynamic characteristics. The moisture stored in or flowing through the soil affects soil formation, structure, stability and erosion and is of primary concern with respect to plant growth. Soil Classification concerns the grouping of soils with a similar range of properties (chemical, physical and biological) into units that can be geo-referenced and mapped. Soils are a very complex natural resource, much more so than air and water. However, it is important to first be able to identify whether a site is contaminated before determining a solution. This paper explores the classification of soil samples at a particular site (McConnell Air Force Base) to investigate the natural or unnatural contamination of soil. The samples are addressed using Data analysis, Naïve Bayes and Support Vector Machine (SVM) Algorithm. From these evaluations, contamination at the site of interest can be considered.*

I. INTRODUCTION

Soil contamination is characterized by solid or liquid hazardous substances mixed with naturally occurring soil. Soil pollution can arise from a number of sources, which could be both naturally occurring in soil and man-made. In other words, the ratio of chemicals in the soil of a given site may be attributed to both natural and unnatural accumulation or production of compounds due to specific environmental conditions. These contaminants can adversely impact the health of plants, animals, and humans when directly or indirectly coming into contact with contaminated soil. Due to the detrimental nature of contamination and the multiple methods to address soil pollution, it is of key interest to be able to determine whether specific sites have contaminated soil. Soil contaminants may vary in both location and type. In this report, the scope of the soil contamination is limited to soil samples taken from the McConnell Air Force Base. The components investigated in the soil samples are fixed as nonbiological concentrations. The goal to be accomplished is to be able to discern between naturally contaminated soil samples and unnaturally contaminated soil samples. If the samples can be consistently classified, a geographical mapping of the unnatural (i.e. man-made) contamination locations may help to determine the source(s) of contamination. The traditional assessment of soil contamination is based on the regular routine of comparison of allowable threshold values with the results of monitoring. This approach is even a required action in environmental agencies, agricultural administration, and managing organization. Very often, solving a particular problem concerning the soil contamination or respective decision making is based solely on single results and not on a more generalized model about the state of the soil contamination in a certain region. The application of multivariate statistical approaches to the problem allows a better classification, modeling, and interpretation of the soil monitoring data. This environmetric strategy makes it possible to detect relationships between the chemical pollutants and specific soil parameters, between sampling sites and, therefore, to achieve a stratification of the pollution. Further, it becomes possible to identify possible pollution sources and to construct apportioning models allowing the determination of the contribution of each identified source to the formation of the total pollutant mass (Stanimirova et al. 2006, 2009; Einax and Soldt 1995; Singh et al. 2008; Andrade et al. 2007; Buszewski and Kowalkowski 2006; Kemper and Sommer 2002; Terrado et al. 2007; Perez Pavon et al. 2008). The aim of the present study is to assess the soil quality in the region of Burgas, Bulgaria by the application of two already classical multivariate statistical methods (cluster analysis and principal components analysis) in order to get information about some spatial distribution of the soil pollutants in the region (by comparing the linkage between the different sampling sites) and to identify possible pollution sources (by determining an appropriate number of latent factors undergoing logic interpretation). The region of Burgas is located close to the Bulgarian Black Sea costal line and is characterized by high industrial and agricultural activity.

II. LITERATURE SURVEY

Studies on Soil Contamination Due To Used Motor Oil and Its Remediation

S. K. Singh, R. K. Srivastava, and Siby John 2018

An experimental program was undertaken to evaluate the changes in behaviour of soils due to interaction with used motor oil (U.M.O) followed by their remediation. Different types of soils classified as clay with low plasticity (CL), clay with high plasticity (CH), and poorly graded sand (SP) were used for the study. Laboratory studies were conducted on virgin (uncontaminated) soil samples and soil samples simulated to varying degrees of contamination (i.e., 3%, 6%, and 9% by dry weight of soil) to compare the geotechnical properties before and after contamination. The engineering properties altered due to contamination. Surfactant (sodium dodecyl sulphate (SDS)) enhanced washing was employed to decontaminate the soils. It was observed that the original geotechnical properties of soils could be almost restored (variation ranging from 0 to 12%) upon decontamination with SDS at an optimum dosage.

Studying the Effects of Contamination on the Geotechnical Properties Of Clayey Soil

M.O. Karkush - 2018

The present study describes the geotechnical behavior of synthetically contaminated soil. Physical, chemical, and mechanical properties were compared with the geotechnical properties of intact soil. The soil samples, disturbed and undisturbed, were obtained from Al-Khadymiya district, which is located at the north west of Baghdad city in Iraq. Four different types of contaminants were used: kerosene, ammonium hydroxide, lead nitrate and copper sulphate. The natural soil samples were contaminated synthetically by soaking in isolated pans containing a solute of water and contaminant for a period of 30 days. The contaminants were mixed with distilled water in two percentages 10 and 25% of the dry weight of the clay soil sample. The results showed that these contaminants have significant effects on the geotechnical and chemical properties of the soil. The contaminants causing an increase in Atterberg's limits (except samples contaminated with lead nitrate), maximum dry unit weigh ($\gamma d,max$), initial void ratio (eo) (except samples contaminated with lead nitrate), compression index (Cc), swelling index (Cr), and collapse potential (CP). Also, the contaminants caused a decrease in specific gravity (Gs) and optimum moisture content (ωopt) (except samples contaminated with ammonium hydroxide), coefficient of vertical consolidation (Cv), and cohesion between soil particles (c).

Effect of Soil Contamination with Azadirachtin on Dehydrogenase and Catalase Activity Of Soil

Rıdvan Kizilkaya, Izzet Akça, Tayfun Aşkin - 2019

Insecticides are used in modern agriculture in large quantities to control pests and increase crop yield. Their use, however, has resulted in the disruption of ecosystems because of the effects on non-target soil microorganisms, some environmental problems, and decreasing soil fertility. These negative effects of synthetic pesticides on the environment have led to the search for alternative means of pest control. One such alternative is use of natural plant products such as azadirachtin that have pesticidal activity.

The aim of this experiment was to study the effect of soil contamination by azadirachtin (C35H44O16) on dehydrogenase (DHA) and catalase activity (CA) of soil under field conditions in Perm, Russia. The tests were conducted on loamy soil (pHH2O 6.7, ECH2O 0.213 dSm-1, organic carbon 0.99%), to which the following quantities of azadirachtin were added: 0, 15, 30 and 60 mL da-1 of soil. Experimental design was randomized plot design with three replications. The DHA and CA analyses were performed 7, 14 and 21 days after the field experiment was established.

The results of field experiment showed that azadirachtin had a positive influence on the DHA and CA at different soil sampling times. The increased doses of azadirachtin applied resulted in the higher level of DHA and CA in soil. The soil DHA and CA showed the highest activity on the 21th day after 60 mL azadirachtin da-1 application doses.

Problem Statement

- Although soil is a non-renewable natural resource, human has increasingly used it as a contaminant sink since industrial Revolution.

- It is getting polluted in a number of ways and there is urgency in controlling the soil pollution in order to preserve the soil fertility and increase the productivity.
- The soil pollution occurs when amounts of some soil elements and other substances may exceed levels recommended for the health of humans, animals, or plants.
- We propose a model to classify the soil type whether it is erosion or not.
- In agricultural soils, however, the concentration of one or more of these elements may be significantly increased in several ways, like through applications of chemicals, sewage sludge, farm slurries, etc.
- Increased doses of fertilizers, pesticides or agricultural chemicals, over a period, add heavy metals to soils which may contaminate them.
- In Existing system to classify the soil contamination using KNN algorithm to be used.

Disadvantages

- Less accuracy
- Poor performance
- Feature Extraction is complex

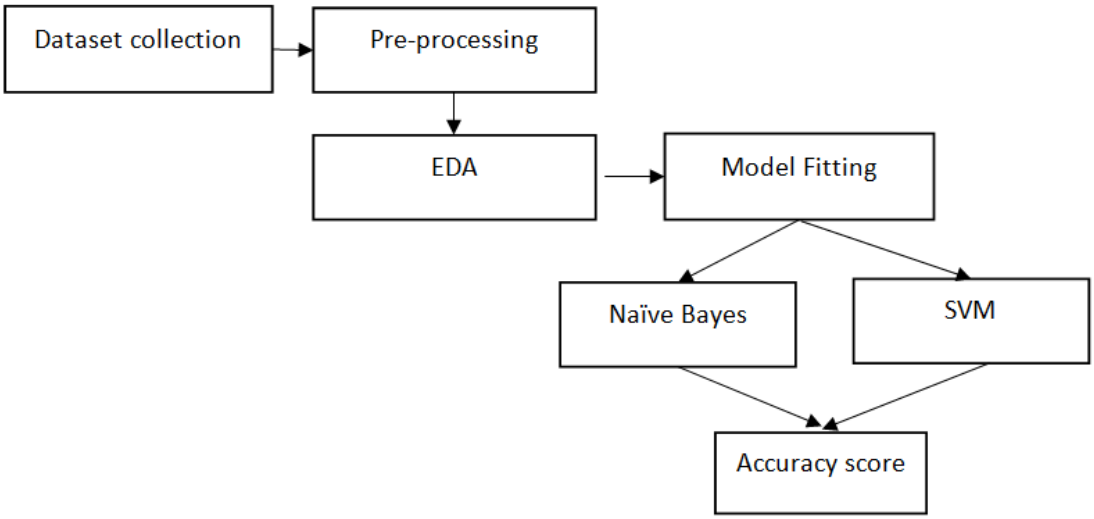
Proposed Work

- Environmental pollution is being the burning challenge of current living organisms on the earth.
- Pollution is the introduction of contaminants into an environment, and undesirable change in air, water and soil which affect human life.
- We propose a model to classify the soil contamination based on PH Level, calcium content, Phosphorus content and sand value
- By using Naïve bayes and Support Vector Machine (SVM) in Machine learning.
- Normalization of the raw input data to dimensionless units in order to avoid the influence of the different range of chemical dimensions (concentration);
- Determination of the distance between the objects of classification by application of some similarity measure, e.g., Euclidean distance or correlation coefficient;
- Performing appropriate linkage between the objects by Naïve bayes and SVM algorithms like single, average or centroid linkage;
- Plotting the results as dendrogram
- Determination of the clustering pattern
- Predict the soil type

Advantages

- High accuracy is obtained and time consumption for detecting the Soil type.
- More datasets are included.
- We can find the all types of soils on different field in application also.
- Easily Extract feature values.
- More performance.
- High accuracy given.

III. METHODOLOGY



3.1 Data collection

The samples data sets are collected from Kaggle website. There are six features in dataset they are: Calcium content, Phosphorus content, pH value, SOC value, Sand value, Soil type The 1230 samples were also mapped to geographic locations around the site, allowing for postliminary contamination assignments.

- Datasets are in Data frame format, Data frame means two dimensional- Two dimensional: x axis and y axis,
- Dataset having input and output column
- Inputs are PH Level, calcium content, Phosphorus content and sand value.
- Output is Soil Type.
- Dataset is in .csv file format.

3.2 pre-processing

Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.

Text pre-processing is an essential a part of any NLP method and the significance of the NLP pre-processing are

- To minimize indexing (or knowledge) records dimension of the textual content records
    1. Stop words bills 20-30% of total phrase counts in a special textual content record
    2. Stemming may just diminish indexing size as much as forty- 50%
- To make stronger the efficiency and effectiveness of the IR method
    1. Stop words aren't valuable for shopping or textual content mining
    2. Stemming used for matching the similar words in a text record

3.3 Exploratory Data Analysis

- Exploratory Data analysis (EDA) is used for visualize the datasets
- To visualize the dataset like pie chart, bar chart, box plot, histogram graph etc.,

3.4 Model fitting

In this proposed system we are using five machine learning algorithms named as Support Vector Machine (SVM) and naïve Bayes algorithms.

- We could able to train the system using these two algorithms and evaluate training score calculated.
- To classify the soil type – Erosion or not.

3.5 Evaluation

- The system will predict the soil type with 98% accuracy
- To evaluate the accuracy score using confusion matrix method

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

Confusion matrix include:

- Precision
- Recall
- Support
- F1score
- Accuracy

IV. IMPLEMENTATION

TABLE 1
DATASET

	Calcium_content	Phosphorus_content	pH_value	SOC_value	Sand_value	Soil_type
0	-0.3	0.0	-1.1	0.4	1.3	1
1	-0.2	-0.2	-0.3	0.1	2.1	1
2	-0.4	0.0	-0.7	-0.3	2.2	1
3	-0.3	-0.2	-0.8	0.1	1.4	1
4	0.0	0.2	0.0	0.6	0.6	1
...	...	...	...	...	...	...
1224	-0.3	-0.3	-0.9	3.1	-1.0	1
1225	-0.3	-0.4	-0.7	2.8	-1.0	0
1226	-0.4	-0.4	-0.4	4.4	-0.4	1
1227	-0.5	-0.4	-0.2	2.5	-0.6	0
1228	-0.2	-0.2	-0.5	4.0	-0.4	1

1229 rows × 6 columns

TABLE 2
CONTENT VALUE

	Calcium_content	Phosphorus_content	pH_value	SOC_value	Sand_value	Soil_type
count	1229.000000	1229.000000	1229.000000	1229.000000	1229.000000	1229.000000
mean	-0.009439	-0.029780	-0.038649	0.080228	-0.001383	0.502034
std	1.044214	0.969263	0.909404	1.148980	1.008727	0.500199
min	-0.500000	-0.400000	-1.900000	-0.900000	-1.500000	0.000000
25%	-0.500000	-0.300000	-0.700000	-0.600000	-0.900000	0.000000
50%	-0.400000	-0.300000	-0.200000	-0.400000	-0.100000	1.000000
75%	-0.100000	-0.100000	0.400000	0.300000	0.800000	1.000000
max	9.600000	13.300000	3.400000	7.600000	2.300000	1.000000

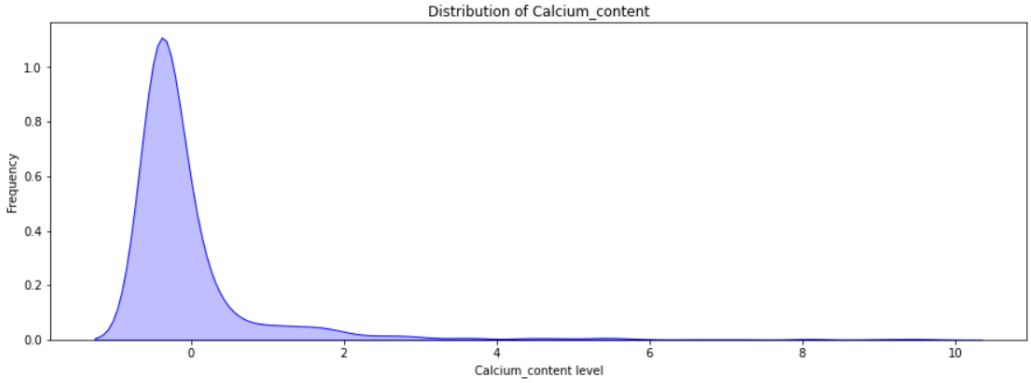


FIGURE 1: Calcium Content Value

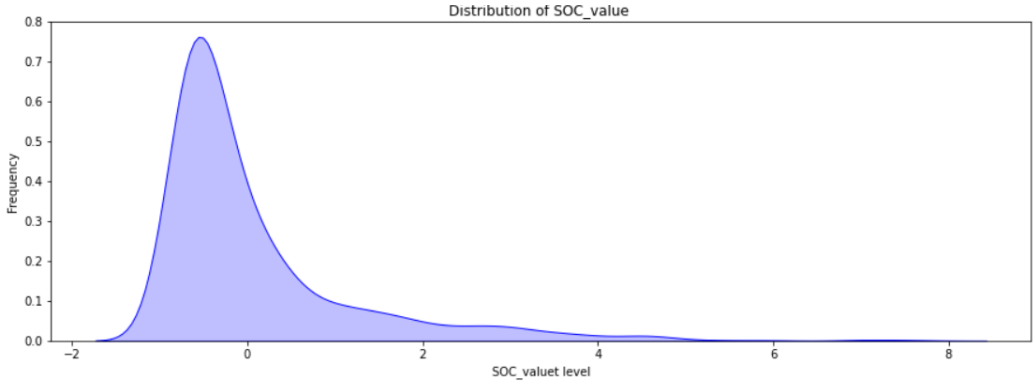


FIGURE 2: SOC Content Value

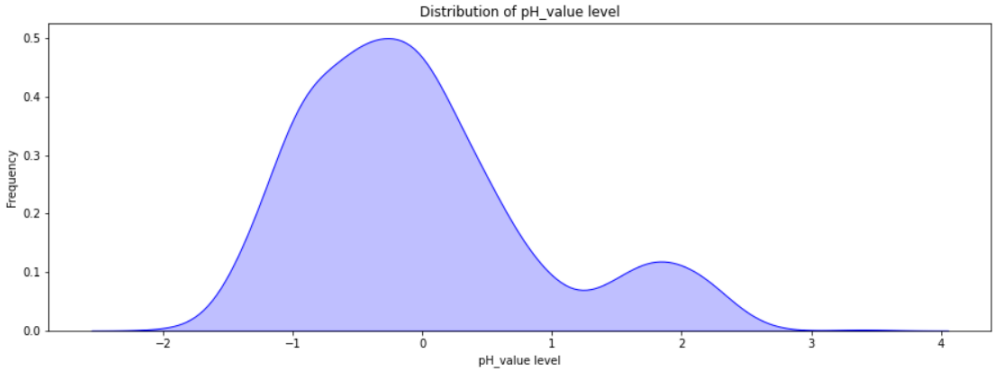


FIGURE 3: pH Value

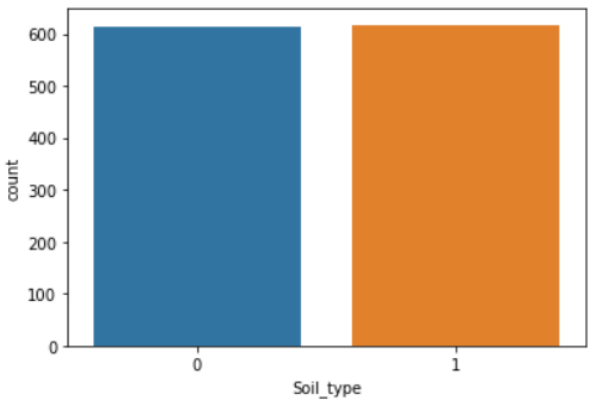


FIGURE 4: Soil Type and Content Count

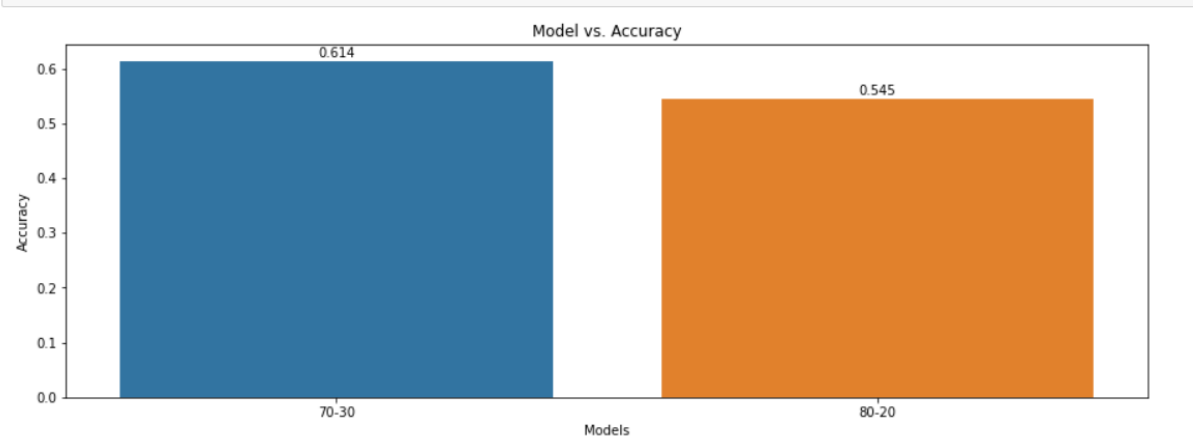


FIGURE 5: Model Vs Accuracy

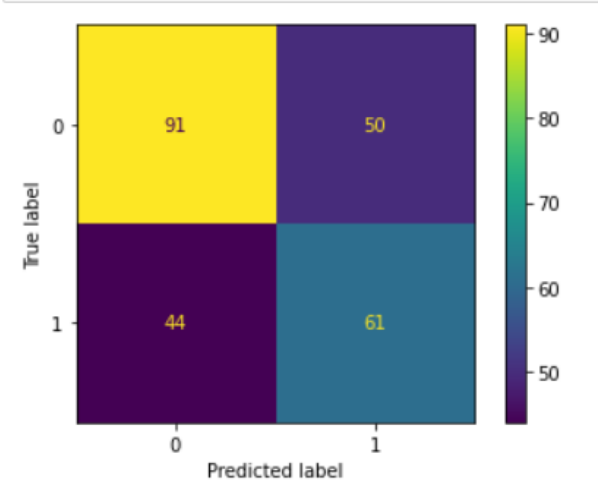


FIGURE 6: Predictions

V. CONCLUSION

From the analyses of the data with respect to classification, it can be stated with high confidence that the soil contamination. This contamination is characterized by especially high concentrations of Sodium, phosphorous and PH level. The source of the contamination is yet unknown, given the provided data, but may be related to depth of soil sample or specific site operations.

Classification of soil sample contamination is one that is constantly undergoing change. Most available data use hierarchical classification to determine clusters of samples, along with principal component analysis. In future work, I would like to investigate the accuracy of hierarchical classifications, using either principal component analysis (PCA) or independent component analysis (ICA). These reductions primarily differ in the assumption of Gaussian features, or lack thereof. I would like to compare the classifications from those strategies and models to those achieved from k-means, as done in this report.

REFERENCES

[1] Singh, S. K., R. K. Srivastava, and Siby John. "Studies on soil contamination due to used motor oil and its remediation." *Canadian Geotechnical Journal* 46, no. 9 (2009): 1077-1083.

[2] Karkush, M. O., A. T. Zaboon, and H. M. Hussien. "Studying the effects of contamination on the geotechnical properties of clayey soil." *Coupled Phenomena in Environmental Geotechnics, Taylor & Francis Group, London* (2013): 599-607.

[3] KIZILKAYA, Rıdvan, A. K. Ç. A. İzzet, Tayfun AŞKIN, Rezan YILMAZ, Vladimir Olekhov, İraida SAMOFALOVA, and Natalya Mudrykh. "Effect of soil contamination with azadirachtin on dehydrogenase and catalase activity of soil." *Eurasian Journal of Soil Science* 1, no. 2 (2012): 98-103.

[4] Weissmannová, Helena Doležalová, and Jiří Pavlovský. "Indices of soil contamination by heavy metals–methodology of calculation for pollution assessment (minireview)." *Environmental monitoring and assessment* 189, no. 12 (2017): 1-25.

[5] Kang, Seong Seung, Kyungho Park, and Daehyeon Kim. "Potential soil contamination in areas where ferronickel slag is used for reclamation work." *Materials* 7, no. 10 (2014): 7157-7172.

[6] Winding, Anne, Kerstin Hund-Rinke, and Michiel Rutgers. "The use of microorganisms in ecological soil classification and assessment concepts." *Ecotoxicology and Environmental Safety* 62, no. 2 (2005): 230-248.

[7] Beyer, W. Nelson. *Evaluating soil contamination*. Vol. 90, no. 2. US Department of the Interior, Fish and Wildlife Service, 1990.

[8] Ji-xi, G. A. O., D. U. A. N. Fei-zhou, and Xiang Bao. "The application of principal component analysis to agriculture soil contamination assessment." 25, no. 5 (2006): 836-842.

[9] Wcisło, E. "Soil contamination with polycyclic aromatic hydrocarbons (PAHs) in Poland-a review." *Polish Journal of Environmental Studies* 7, no. 5 (1998): 267-272.

[10] Bradham, Karen D., Elizabeth A. Dayton, Nicholas T. Basta, Jackie Schroder, Mark Payton, and Roman P. Lanno. "Effect of soil properties on lead bioavailability and toxicity to earthworms." *Environmental Toxicology and Chemistry: An International Journal* 25, no. 3 (2006): 769-775.