

RESEARCH ON SOIL ENVIRONMENTAL
AND HEALTH RISK ANALYSIS BASED
ON MACHINE LEARNING.

ZHAO ZHIHAN

UNIVERSITI TEKNOLOGI MALAYSIA

1.Introduction

Due to the rapid development of global industry and agriculture, the degree of soil pollution worldwide has increased significantly. Heavy metals, pesticide residues, and industrial wastewater accumulate gradually in the soil, which may lead to the contamination of crops. These contaminated crops pose a huge threat to public health. However, traditional statistical survey methods for soil assessment have limitations and lack the ability for multi - source integration.

In recent years, machine - learning models have witnessed rapid development and have been successfully applied in various environmental fields. These techniques can quickly and effectively capture some in - depth data. Such research methods can efficiently analyze a large amount of isolated data and conduct a relatively comprehensive analysis of datasets. Based on the above - mentioned developments, the aim of this study is to analyze the in - depth relationships between different environments and datasets through cutting - edge machine - learning algorithms, and to predict soil pollution and issue early warnings for environmental pollution.

2.Problem Background

Heavy metals and other organic pollutants contaminate crops through the soil and thus enter the food chain. These pollutants can cause chronic diseases such as neurological disorders and cancer. In addition, weather conditions and industrial activities around the soil can also exacerbate soil pollution. Traditional pollution assessment techniques fail to incorporate all factors. Moreover, heterogeneous datasets are isolated from each other, making data analysis more difficult.

3.Problem Statement

Although there are individual studies indicating associations between specific soil pollutants and health problems, currently, there is no robust and scalable framework to integrate various environmental and clinical datasets for spatial risk modeling. As a result, relevant departments lack precise, data - driven, and efficient methods. They are unable to efficiently identify high - risk areas and it is also very difficult to establish an effective high - risk early - warning system.

4. Research Questions

How can machine learning techniques be applied to integrate multi-source soil, environmental, and health data for spatial risk assessment?

What spatial patterns emerge between soil contaminant levels, meteorological conditions, and disease incidence?

Which ML models yield the highest predictive accuracy for identifying high-risk regions?

5.Research Objectives

To develop a data integration pipeline for harmonizing soil pollutant, weather, agricultural, industrial, health, and demographic data.

To implement and compare multiple ML algorithms for predicting spatial health risks based on soil contamination.

To design a prototype early warning dashboard.

6 Scope of Study

The scope of the project encompasses a diverse range of data. Environmental data includes soil pollutant concentrations, soil types, and weather conditions. Industrial and agricultural data covers the types of industries and agriculture as well as industry distribution. Health data involves disease types, severity, symptoms, and health reports. Additionally, demographic data focuses on aspects such as the gender and age of the affected population.

7. Significance of the Study

Use machine - learning methods to build models, make full use of various types of data to identify the impacts of factors such as soil pollution and climate on human health, and construct a supervision and evaluation system. Provide a reliable scientific basis for people in different regions to supervise and prevent soil pollution. While promoting industrial and agricultural development, offer an effective soil protection plan, thereby reducing the harm of soil pollution to humans.