Science of the Total Environment 485–486 (2014) 71–82

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

The use of cluster analysis for plant grouping by their tolerance to soil contamination with hydrocarbons at the germination stage

Konstantin Potashev, Natalia Sharonova, Irina Breus *

Kazan Federal University, 18 Kremlevskaja Str., Kazan 420008, Russian Federation

HIGHLIGHTS

- Cluster analysis is promising for plant grouping by seed tolerance to hydrocarbons.
- Reduction of the dimension of input matrix in a dose–response format was performed.
- Clustering using two independent parameters with practical meaning was proposed.
- In contrast to the manual ranking the generalized seed tolerance was estimated.
- Plant features governing germination of 42 plants upon contamination were revealed.

ARTICLE INFO

Article history:
Received 3 December 2013
Received in revised form 15 March 2014
Accepted 16 March 2014
Available online 3 April 2014

Editor: Mark Hanson

Keywords:
Soil contamination
Hydrocarbons
Germination
Cluster analysis
Plant tolerance

ABSTRACT

Clustering was employed for the analysis of obtained experimental data set (42 plants in total) on seed germination in leached chernozem contaminated with kerosene. Among investigated plants were 31 cultivated plants from 11 families (27 species and 20 varieties) and 11 wild plant species from 7 families, 23 annual and 19 perennial/biannual plant species, 11 monocotyledonous and 31 dicotyledonous plants. Two-dimensional (two-parameter) clustering approach, allowing the estimation of tolerance of germinating seeds using a pair of independent parameters ($C_{75\%}$, $V_{7\%}$) was found to be most effective. These parameters characterized the ability of seeds to both withstand high concentrations of contaminants without the significant reduction of the germination, and maintain high germination rate within certain contaminant concentrations. The performed clustering revealed a number of plant features, which define the relation of a particular plant to a particular tolerance cluster; it has also demonstrated the possibility of generalizing the kerosene results for n-tridecane, which is one of the typical kerosene components. In contrast to the "manual" plant ranking based on the assessment of germination at discrete concentrations of the contaminant, the proposed clustering approach allowed a generalized characterization of the seed tolerance/sensitivity to hydrocarbon contaminants.

1. Introduction

Modern systems of computer data processing enable the analysis of arrays of poorly structured data in various fields. Such problem is common for biological objects in which data matrices do not always meet the requirements of completeness, statistical reproducibility, etc. The methods of cluster analysis have a high potential to solve the problem of organizing such data into meaningful structures. Cluster analysis is often applied as a starting point, called exploratory data analysis, for solving classification problems based on unsupervised learning (Massart and Kaufman, 1983). This term was first used by Tryon (1939) and it encompasses a number of different algorithms and methods for grouping objects of similar kind into respective groups that are meaningful (SAS Institute Inc., 2008; Hanrahan, 2010). It reduces the number of observations by grouping the objects into a smaller set of groups of relatively homogeneous cases in such a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Thus, it spontaneously discovers structures in data without explaining their existence.

In the current literature, the application potential of cluster analysis for biological objects is undervalued. The main reason for this is the necessity to have a wide range of study objects. Clustering was used for plants classification (identification of similarities between species and hybrids) based on morphometric and germination data for *Amaranthus* (Lanta et al., 2003) and *Pennisetum purpureum* (Zhang et al., 2010) plants, as well as for the interpretation of soil quality monitoring data in order to determine the relationships between the chemical contaminants and specific soil parameters (Astel et al., 2011). However, this method was very rarely employed to study the effect of contaminants.

We propose that cluster analysis may be promising for primary processing of data on plant tolerance to soil contamination with petroleum hydrocarbons (HC). HC are constituents of engine fuels, industrial

* Corresponding author.

http://dx.doi.org/10.1016/j.scitotenv.2014.03.067
0048-9697/