Geoderma 259–260 (2015) 134–148

Contents lists available at ScienceDirect

Geoderma

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

On the application of Bayesian Networks in Digital Soil Mapping

K. Taalab [a], R. Corstanje [b], J. Zawadzka [b], T. Mayr [b,*], M.J. Whelan [c], J.A. Hannam [b], R. Creamer [d]

[a] University College London, London, Greater London WC1E 6BT, United Kingdom
[b] Cranfield University, Cranfield, Bedfordshire MK43 0AL, United Kingdom
[c] University of Leicester, Leicester, Leicestershire LE1 7RH, United Kingdom
[d] Teagasc, Johnstown Castle, Wexford, Co. Wexford, Ireland

ARTICLE INFO

Article history:
Received 5 June 2014
Received in revised form 15 May 2015
Accepted 26 May 2015
Available online 11 June 2015

Keywords:
Bayesian Networks
Soil
Bulk density
Expert knowledge
Mapping
Modelling

ABSTRACT

Two corresponding issues concerning Digital Soil Mapping are the demand for up-to-date, fine resolution soil data and the need to determine soil–landscape relationships. In this study, we propose a Bayesian Network framework as a suitable modelling approach to fulfil these requirements. Bayesian Networks are graphical probabilistic models in which predictions are obtained using prior probabilities derived from either measured data or expert opinion. They represent cause and effect relationships through connections in a network system. The advantage of the Bayesian Networks approach is that the models are easy to interpret and the uncertainty inherent in the relationships between variables can be expressed in terms of probability. In this study we will define the fundamentals of a Bayesian Network and the probability theory that underpins predictions. Then, using case studies, we demonstrate how they can be applied to predict soil properties (bulk density) and soil taxonomic class (associations).

1. Introduction

To satisfy the growing demand for up-to-date, fine resolution soil data, there is a call to fully explore the potential of current mapping and modelling software, and apply existing modelling techniques in novel and innovative ways (Hartemink and McBratney, 2008). Predictive modelling of the spatial pattern of soil types and properties is based on a quasi-mechanistic understanding of soil formation and the factors that drive soil variation in the landscape, namely the ClORPT factors (Climate, Organic activity, Relief, Parent material and Time; Jenny, 1941). The relationships between soil forming factors and soil properties are complex and several non-linear modelling techniques have been employed to represent them including Random Forests (RFs) (Liaw and Wiener, 2002; Grimm et al., 2008; Wiesmeier et al., 2011) and Artificial Neural Networks (ANNs) (Agyare et al., 2007; Zhao et al., 2010). A principal disadvantage of these methods is that they are 'black-box', meaning that it is often difficult to interpret the relationship between response and predictor variables in physical terms (Suuster et al., 2012). In Bayesian Networks (BNs) the relationship between soil forming factors and soil properties can be directly addressed (Tavares Wahren et al., 2012). Many significant soil processes are not particularly well understood at the landscape scale and would benefit from the clarity and insight provided by BN modelling (e.g. Braakhekke et al., 2012). Chen and Pollino (2012) stated that improving system understanding is a key motivation for using a BN.

BNs are graphical probabilistic models in which predictions are obtained using prior probabilities derived from either measured data or expert opinion. They represent cause and effect relationships via connections in a network system (Hough et al., 2010) but they differ from other network based methods, such as ANNs, in that the structure of the network and the interactions between nodes are defined by the user based on prevailing process understanding. BNs are a flexible way of structuring process understanding stochastically and, unlike purely deterministic models, reflect the uncertainty surrounding cause–effect relationships (one event leading to another) by expressing the relationships between soil classes/properties and the covariates as a probability (Dlamini, 2010). They are also ideal for addressing problems where data are limited (Kuhnert and Hayes, 2009). BNs have been applied to ecological systems (McCann et al., 2006), notably conservation (McCloskey et al., 2011), habitat mapping (Smith et al., 2007), and risk mapping of events such as wildfire (Dlamini, 2010) and peat erosion (Aalders et al., 2011). Bayesian modelling approaches have also been applied to modelling soil classes (Skidmore et al., 1996; Bui et al., 1999; Mayr et al., 2010) or soil attributes (Cook et al., 1996; Corner et al., 2002). Despite this, BNs are not yet established as a mainstream tool in Digital Soil Mapping (DSM).

BNs were developed from the branch of mathematics known as probability theory, in particular from probabilistic reasoning (Pearl, 1988). Unlike deterministic models, BNs offer a structured method of dealing with uncertainty that, as a rule, diminishes as more information

* Corresponding author.
E-mail address: t.mayr@cranfield.ac.uk (T. Mayr).

http://dx.doi.org/10.1016/j.geoderma.2015.05.014
0016-7061/

is gathered. In the case of predicting the spatial distribution of soil classes and properties, the relationships between variables are highly uncertain and data availability is often limited, so BNs have great potential as a predictive tool (Finke, 2012). Another appealing aspect of BNs is their ability to integrate expert knowledge into the model, which can be used to supplement measured data, or define relationships between variables directly. There has been a long-standing drive to formally introduce expert knowledge into soil mapping, usually focusing on fuzzy set theory or possibility theory (McBratney and Odeh, 1997). In contrast, BNs use probability theory, which can be seen to offer a more coherent structure to decision making problems (Degroot, 1988), although, there has been some debate as to which is the superior approach (Krueger et al., 2012). In this study, BNs are explored for two typical problems in DSM; i) the prediction of a soil property, soil bulk density, and ii) the prediction of a soil taxonomic class.

BNs have a number of advantages compared with other modelling techniques regularly employed in DSM:

- The real strength of BNs can be fully appreciated in situations where domain knowledge is crucial and availability of data is scarce such as in Case Study 1 on bulk density.
- While BNs did not improve predictive performance, they have the advantage of offering some process-based insight (Correa et al., 2009). This has been confirmed in this study where the results are very similar to (albeit slightly lower than) the ANN and RF black-box modelling techniques which have previously been used to predict topsoil D_b with the same dataset (Taalab et al., 2012).
- In recent years, the focus of DSM has moved away from straightforward classification of soils towards developing a better understanding of the spatial distribution of soils in relation to the wider environment (Grunwald, 2009). This is necessary in order to resolve challenges such as climate change, desertification, and food production which are putting increasing pressure on soils as a resource (Hartemink and McBratney, 2008).
- BNs provide an opportunity to assess the understanding of soil processes. In conjunction with expert knowledge, BNs can either confirm or contradict the opinions of the expert(s). If the BN contradicts what the expert believes, it can prompt further investigation into the process, indicating a knowledge gap or a problem with the model itself. If the latter is the case, it is easy to amend both the model structure and the probabilistic relationship between nodes. Identifying the source of predictive inaccuracies in a black-box model is much less straightforward.
- As BNs are based on process understanding they can be used to answer specific questions using predictive reasoning. For example, what is the probability of X, given certain information, a capability that black box models do not possess. In addition, BNs are also capable of diagnostic reasoning. For example, given an outcome, the favourable conditions likely to lead to this outcome can be predicted.

In summary, the major appeal of BNs is their clarity, which allows experts to judge whether the model makes pedogenic sense and to develop a better understanding of the soil processes.

2. Materials and methods

2.1. Theory

BNs are named after the Reverend Thomas Bayes who, in the 18th century, developed a theorem regarding changing probabilities given new information (Bayes, 1783). The basis of a BN is conditional probability, which can be explained using an example from Jensen (1996), where a statement of conditional probability reads

"Given an event B, the probability of event A is x."

In mathematical notation this would read

$$P(A|B) = x. \tag{1}$$

This statement holds true, only if all other information which could affect event A is known and has been accounted for. The basic rule of conditional probability is:

$$P(A|B)P(B) = P(A,B) \tag{2}$$

where $P(A,B)$ is the probability of the joint event A and B both being true ($A \wedge B$). From this, the Bayes Rule (Eq. (3)) can be derived.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3}$$

This rule forms the basis of BN modelling, as we can use Bayes' rule to inform us of the probability of event A given information about B. Referring to Eq. (1), the posterior probability $P(A|B)$ was an unknown x, we now see that it can be calculated using our prior belief in the occurrence of event A $P(A)$ and event B $P(B)$ and the probability of B given that A has occurred $P(B|A)$. This is known as Bayesian inference and to illustrate how this might work in practice for DSM applications, we adapt an example given by Aitkenhead and Aalders (2009).

From Eq. (3), $P(A|B)$ is the posterior probability of event A (e.g. high bulk density; D_b) given B (e.g. arable land use) (note that the class 'high bulk density' is an example of discretization of a continuous variable into a set of classes, the boundaries of which would need to be defined). $P(A)$ is the probability that bulk density is 'large' (a prior probability derived from either data i.e. the percentage of samples recorded as high or from expert opinion), $P(B)$ is the probability of the occurrence of arable land (proportion of the study area that is arable land) and $P(B|A)$ is the proportion of high bulk density samples that are found on arable land. For example, if 30% of the total number of D_b samples are classified as large, i.e. $P(A) = 0.3$, 40% of the terrain in the study area is classed as arable, i.e. $P(B) = 0.4$., and the proportion of high D_b samples found on arable land is 50% i.e. prior probability $P(B|A) = 0.5$. This probability can be generated either by expert knowledge or using observed data. Combined, these probabilities give the probability that if the land is arable, the bulk density will be high, known as the posterior probability $P(A|B)$. In this instance;

$$P(A|B) = \frac{0.5 * 0.3}{0.4} = 0.375 \tag{4}$$

hence, there is a 37.5% probability that D_b will be high on arable land.

In reality, when dealing with complex problems in soil mapping, there will be numerous factors that influence variables of interest. Hence BNs are designed to link large numbers of influencing variables and combine the conditional probabilities of each. BNs comprise two components; 1) a directed acyclic graph (DAG), where each node represents a variable in which the directed links between nodes represent the conditional dependencies of the model and 2) a quantitative component of a network consisting of conditional probability tables (CPT) that accompany each node, which define the dependencies of each variable. Each CPT contains a list of possible states that could be applied to the variable. Using an example adapted from Nadkarni and Shenoy (2004), Fig. 1 shows a BN comprised of four variables: Land Use (L), Soil Group (S), Organic Carbon Content (C) and Soil Bulk Density (D). The directional arrows between variables indicate causality. The variables with arrows leading into them are known as the 'child nodes' and the variables where the arrows originate are known as 'parent nodes'. Each state is mutually exclusive and the list is definitive; for clarity, we have kept the number of states in Fig. 1 to a minimum. It is acknowledged, however, that in complex natural systems the environmental