Neth. J. agric. Sci. 24 (1976): 179-186

Computer-aided interactive soil suitability classification - a simple Bayesian approach

Stein W. Bie[1], Joris R. E. Lieftinck[2], Karel R. van Lynden[1] and Anton W. Waenink[1]

[1] Netherlands Soil Survey Institute, P.O. Box 98, Wageningen, the Netherlands
[2] IWIS-TNO, P.O. Box 297, The Hague, the Netherlands

Accepted: 20 May 1976

Key words: soil suitability classification, interactive computer identification, Bayesian statistics

Summary

Soil suitability classification may be approached as the allocation of a soil individual to a suitability class on the basis of its values for a number of assessment factors (soil attributes). The study reported here uses a simple Bayesian algorithm to calculate for each suitability class the probability that a soil individual is a member thereof. This is achieved by comparing the values of the assessment factors to an existing set of already classified individuals. The system is implemented as a user-friendly interactive computer program, and an application to soil suitability classification for forestry in the Netherlands illustrates its use.

Introduction

Soil suitability classification is a reinterpretation of existing soil information for a particular land use. This interpretation considers the values of the soil attributes (assessment factors) in the light of the requirements for the land use. For simplicity the interpretation is usually confined to the central concept of a soil profile class, a legend unit or a mapping unit. It gives rise to a correspondence table where one or more profile classes (or legend units or mapping units) correspond to one suitability class.

Yet it is widely appreciated that ideally every soil individual should be considered for soil suitability, on the basis of the values of the assessment factors actually occurring for that individual. In practice the construction of correspondence tables for every soil individual (i.e. every foreseeable combination of attribute-value pairs) for each potential land use becomes impossible.

This paper suggests how this difficulty may be overcome, and illustrates the use of Bayesian techniques in computer-aided interactive soil suitability classification for forestry in the Netherlands.

The principle

The approach used by us is a simple attempt to mimic human thought processes for soil suitability classification.

For one particular land use a soil scientist has through experiments, observation or experience concluded that a soil individual belongs to a suitability class. Faced with a new individual, so far unclassified, he or she will review the value of the assessment factors of the new individual, one by one in a subjectively chosen order. He or she will begin to form an opinion of the class to which the individual may belong. The preliminary conclusion will be reinforced, weakened and changed (updated) for each of the following attribute-value pairs considered. The sequence of assessment factors used may depend on the preliminary conclusions drawn. Some or all attribute-value pairs will be used for this classification. Once classified, the result may or may not be incorporated in the scientist's 'experience', depending on the certainty the scientist attaches to the final classification.

In our approach this thought process is simulated by the use of:

1) a simple Bayesian algorithm,
2) interactive computer-aided classification,
3) automated construction of the correspondence table for later use (report file).

In another application in soil science, Giltrap et al. (1974) demonstrated the use of Bayes' theorem for the evaluation of the goodness of parts of the New Zealand soil classification.

Simple Bayesian algorithm[1]

Assume that we have a number of soil individuals for which we have constructed a correspondence table linking the name of the soil individual to a suitability class for a particular purpose. Further assume that the suitability classes are obtained by considering the values of a number of pertinent soil attributes.

We can call this set of soil individuals for a *trial population*, divided into K suitability classes. The attribute-value pairs are recorded within an *incidence matrix* H, where H_{ijk} is the observed frequency in group k of value j of attribute i.

We can now derive from H the probability matrix P, where

$$P_{ijk} = H_{ijk}/H_{i.k.}$$

The *prior* probabilities (p_1, $p_2 \ldots p_k$) describe the relative probabilities that an individual belongs to each of the K groups on the basis of the trial population before any value of any attribute is investigated. Assume we now know the attribute-value pair ij, we may then calculate the *posterior* probabilities (p'_1, p'_2, $\ldots p'_k$).

Bayes' theorem on the probabilities that A and B occur

$$P(A|B) = P(A.B)/P(B)$$

may in this context be rewritten as

[1] We here follow closely the terminology of Wishart (1973).

$$p'_k = p_k P_{ijk} / \sum_m (p_m P_{ijm}),$$

where P_{ijm} is the probability that attribute i has the value j for class m.

WIACLAS

WIACLAS (*W*ishart *I*nter-*A*ctive *CLA*ssification *S*ystem) implements the simple Bayesian algorithm in the context of soil suitability classification.[1]

A number of soil individuals, with their associated attribute-value pairs and suitability classes form the trial population, from which the initial incidence matrix H and the derived probability matrix P are created. H and P may be created from a set of original observations, created by adding the individuals one by one, or as synthetic values.

The soil scientist is hooked up to WIACLAS by a remote terminal and is asked to give the soil individual an identifier, to be used in the report file. The system then produces the prior probabilities $p_1, p_2 \ldots p_k$ given that it knows nothing about the values of the attributes.

It then poses the first question:

Attribute 1 = ?

to which the response is j.

The posterior probabilities $p'_1, p'_2 \ldots p'_k$ are calculated and displayed. The system then optimises the questioning by posing a further question so that, on the basis of H, the greatest chance exists that the highest value of p' will be reinforced.

For the following question $p'_1, p'_2, \ldots p'_k$ will be regarded as new prior probabilities.

The questioning is broken off either when the posterior probability for one class reaches a set minimal value or when a stated number of questions have been asked (this could be all possible questions).

The system then invites the scientist to accept or reject the result of the classification. If he or she accepts the result, the matrices H and P are updated to take account of the new individual, if rejected the matrices remain unaltered.

The system generates automatically a report of each classification exercise, containing the identifier, the values of the attributes, the date of classification and the time as well as the result of the classification and the final probabilities for classes 1, 2 . . . k. This report file can be used for subsequent manipulation, calculations and retrievals for documentation purposes.

An example

Fig. 1 illustrates the incidence matrix H for the tree species ash (*Fraxinus excelsior*) generated from a trial population of 683 individuals for 4 attributes.

[1] A substantially enlarged and amended version of INTERCOM 1 originally written by David Wishart, Civil Service Department, London (Wishart, 1972). WIACLAS consists mainly of FORTRAN IV modules and is currently implemented on a CDC Cyber 72 (64K words under Scope 3.4.3.).

```
TEXTREE   <- Name question file
INCES     <- Name incidence file
3         <- No. possible suitability classes
378  137  168  <- No. examples in classes 1, 2 and 3

          K1  K2  Z1  Z2  Z3  Z4  Z5  V1.1 V1.2 V2  V3  V4  V5
VT    3   1   1   2   2   9   9   0   4   57  154 92  44  <- Class 1
      13  5   2   13  7   0   0   0   23  74  0   0   0   <- Class 2
      56  31  11  0   0   0   0   0   70  0   0   0   0   <- Class 3

          1.1 1.2 2   3.1 3.2 3.3
LV    111 72  84  58  38  15
      8   29  43  28  29  0
      19  44  61  44  0   0

          1.1 1.2 2   3.1 3.2
VL    115 89  93  48  33
      69  51  17  0   0
      78  90  0   0   0

          1   2   3   unknown
pH    12  52  287 27
      12  14  71  40
      8   8   54  98
```

Fig. 1. Annotated print-out of incidence file for ash based on 683 individuals. Similar incidence files exist for 11 more tree species grown commercially in the Netherlands. The matrix illustrates the frequency by which each combination of class and value of an assessment factor occurs in the trial population.

Fig. 2 are two classifications completed with the system, with minimum automatic update probability set at 1.0, to ensure that the soil scientist must himself decide on whether to accept the classification or not.

Fig. 3 is the report file of the two classifications, illustrating, in the second example, how the rejection of the classification leads to an altered report. The system also has an option where the soil scientist may insert his or her subjective classification in the report file (it will be flagged accordingly) although the incidence matrix remains unchanged.

Discussion

WIACLAS may serve at least 5 purposes:

1. WIACLAS may be used as a class allocation program allocating unclassified individuals to a class in an existing classification that is not being updated. No knowledge is required as to the relative importance of each attribute to the classifi-

```
TYPE IDENTIFICATIØNkpz21          } Identification
11                                }
3.2                               }

PRØB.: .5510 .2012 .2478   <— Begin probabilities
V.L.=?   1.1               <— Moisture supply ?   1.1

PRØB.: .4365 .2638 .2996   <— Aeration ?   3.2
L.V.=?   3.2

PRØB.: .4399 .5557 .0044   <— Fertility ?   Z1
V.T.=?   z1

PRØB.: .2070 .7816 .0114   <— pH ?   2
PH  =?   2

PRØB.: .2628 .7317 .0056
TEST ENDS AT STEP  4 TARGET GRØUP 2 PRØBABILITY  .731669   <— Group 2 = suitable
UPDATE?y   <— Yes

STØPPEN? n   <— Stop ? No

TYPE IDENTIFICATIØNtzd21   ↓ Following classification
vii
1.1

PRØB.: .5502 .2023 .2475   <— Note begin probabilities updated
V.L.=?   3.2

PRØB.: .9866 .0060 .0074   <— Minimum 2 questions must be asked
L.V.=?   1.1

PRØB.: .9960 .0012 .0028                                          } After 2 questions
TEST ENDS AT STEP  2 TARGET GRØUP 1 PRØBABILITY  .995961          } probability exceeded
UPDATE?y                                                          } 0.85, so questioning
                                                                  } discontinued.
STØPPEN? n                                                        } Group 1 = unsuitable

TYPE IDENTIFICATIØNzn21
11
3.2

PRØB.: .5509 .2020 .2471
V.L.=?   1.2

PRØB.: .3846 .2230 .3925
PH  =?   2

PRØB.: .5412 .2462 .2126
L.V.=?   3.2

PRØB.: .5038 .4933 .0029
V.T.=?   z2

PRØB.: .2795 .7204 .0001
TEST ENDS AT STEP  4 TARGET GRØUP 2 PRØBABILITY  .720403
UPDATE?n   <— No, disagree with suggested classification

AAN WELKE GRØEP DACHT U?1   <— Think it should be Group 1
```

Fig. 2. Annotated print-out of an interactive session involving 3 classifications. The first example illustrates a complete run where the result is accepted, the second example a short run where an adequate probability is reached after the set minimum of two questions has been asked, in the third example the classification suggested by WIACLAS is rejected by the scientist, who thinks that class 1 is more likely to be correct, on the basis of information other than the assessment factors used.

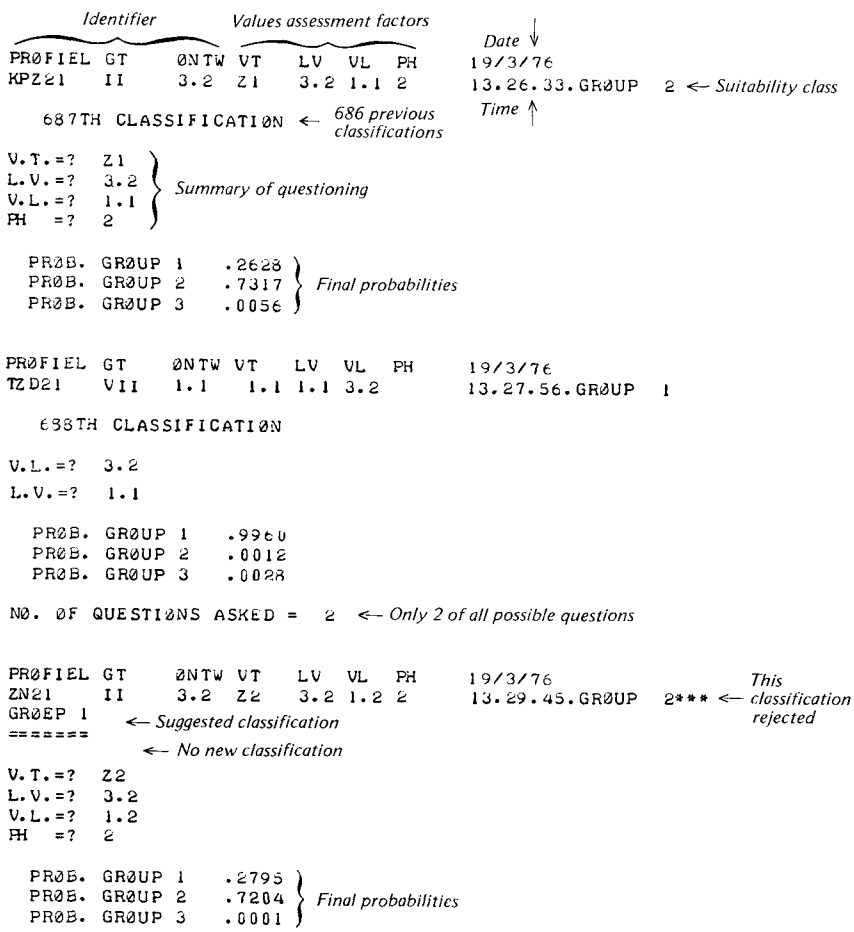

Fig. 3. The automatically generated report file with annotations. This report file forms the archival material that can later be used for further data manipulation (statistics, retrievals).

cation, or their interdependence. The system calculates the relations in the form of probabilities. If the real relations were known, more powerful class allocation algorithms could be devised, but such algorithms are of little use when, as seems often to be the case with soil suitability classification, the relations are poorly known or only local in extent. WIACLAS may thus serve the conventional function of a fixed (static) classification system.

2. WIACLAS may also be used to construct a static system, since the trial population need not be larger than one individual in each class (ideally it will of course be larger). Once the number of individuals is large enough in the eyes of the soil scientist, the incidence matrix may be 'frozen', and will go on functioning as a static system as under 1 above.

3. The ability to build up a file, as in 2, also allows for the interesting possibility of doing soil suitability classification not by referring to a static file (where a given set of attribute-value pairs will lead to the same result, however many times repeated) but to a dynamic file where each accepted classification will lead to an update of the incidence matrix. This makes use of the self-learning capacity of WIACLAS. Any soil individual can be classified on the basis of the incidence file existing at that moment, thus increasing the chance that the classification will be optimal. It does therefore include that a soil individual entered for a second time will be allocated to a different class from the first classification, as the incidence file may have been changed in the meantime. This approach is clearly a departure from current working methods, which make use of a static file. It overcomes the shortcoming of a static file in respect to optimal classification, for whilst the static file may be optimal at its creation, it can take no account of experience subsequently gained. It should be noted that the use of this dynamic file contravenes the commonly held notion that a classification of an individual should be reproducible.
4. During the process of interactive classification the soil scientist will get an impression of the impact of individual attribute-value pairs on the class allocation. It constitutes thus a form of sensitivity analysis.
5. As constructed, the report file generated by the system can be used for other manipulation in a computerised information system. The automated addition of date and time, as well as the sequence number of the classification, give further possibilities for retrieval.

A number of large soil survey organizations are currently (1975) known to be constructing soil interpretation files (e.g. U.S. Department of Agriculture, Canada Department of Agriculture, Netherlands Soil Survey Institute). The files proposed appear to be based on the principle of static files, with a correspondence table to link named soil classes to suitability classes. Whilst the system outlined in this article also may be used towards this purpose, it also opens the opportunity for continuous optimalization of the suitability classification through a dynamic file. Most importantly, it allows each soil individual (as distinct from class) to be classified easily and reliably, with an estimate of the goodness of the classification.

Conclusion

When soil suitability classification is based on the allocation of a soil individual to a class on the basis of the values of a number of attribute-value pairs, the workability of the classification depends on the ability to construct correspondence tables linking the various combinations of attribute-value pairs to the suitability classes. Since the number of possible combinations is usually large, correspondence tables are commonly linked to central concepts of soil classes or mapping units. To aid the ease of application, the correspondence tables are usually static, i.e. they remain fixed once decided upon.

With increasing emphasis on quantification also in soil suitability classification is likely to come a demand that the classification remains optimal through time.

This requires dynamic correspondence tables, where experience gained is continuously integrated in the classification. The use of a simple Bayesian algorithm comes some way to meet this request for a self-learning soil suitability classification.

We have implemented this algorithm in a user-friendly conversational computer program, which now shows promise in its application to soil suitability classification in the Netherlands.

Acknowledgments

We are very grateful for the assistance given by David Wishart during the initial stages of the development of WIACLAS. The work reported here is part of a project by Werkgemeenschap Informatiesysteem Aardwetenschappen (Working Community Information System for the Earth Sciences, the Netherlands).

References

Giltrap, D. J., L. C. Blakemore & M. L. Leamy, 1974. A probability approach to soil classification. 1. Application of chemical data to the assignment of soils to Category III in the New Zealand genetic soil classification. *N. Z. Jl Sci.* 17: 451-461.

Wishart, D., 1972. Computers in personnel management. *Personnel Rev.* 1 (2).

Wishart, D., 1973. A Bayesian method for computer-aided medical diagnosis. In: H.-J. Lange & G. Wagner (Eds), Computerunterstützte ärtztliche Diagnostik. Schattauer, Stuttgart, p. 305-310.