DOCTORAL (PhD) THESES

FRIDERIKA MIKE - HEGEDŰS

Mosonmagyaróvár
2006

DOCTORAL (PhD) THESES

UNIVERSITY OF WEST-HUNGARY
Faculty of Agricultural and Food Sciences
Institute of Biosystems Engineering

PhD School for Precision Crop Production Methods
Head of the PhD School:
Prof. Dr. Géza Kuroli DSc.

Program:
Technical condition system of habitat – specific precision crop production
Program Director and Dissertation Adviser:
Prof. Dr. Miklós Neményi DSc.

Applying fuzzy logic and neural networks to database evaluation in precision agriculture

by
FRIDERIKA MIKE - HEGEDŰS

Mosonmagyaróvár
2006

Introduction

Location specific precision plant cultivation applies accurately prescribed technology and treatments that are appropriate for specified locations within a field. The analysis of variability within fields and the decision-making process based upon this analysis call for an interdisciplinary approach. This approach must account for the collection of data from different fields and resources, the storage, transmission and analysis of these data as well as the establishment of decision-making techniques. A co-operation between these fields of research leads to a better understanding and handling of the causes of within field variability.

Yield monitoring provides producers with a direct method to measure the spatial variability of yield. Yield maps based on collected yield data differentiate between areas with different yield levels and, as a result, radically change the decision-making process. Variability is caused by several factors, including soil type, physical and chemical soil properties, field location, previous production and the availability of nutrients. The accurate mapping of soil properties is critical for the success of location specific cultivation. The number of soil samples is also decisive. Moreover, the method of interpolation that is used for converting discrete sample data into continuous maps is one of the most important factors for the success of soil mapping.

Multifactor models and prediction techniques contribute to sound decision-making. Consequently, analytical and predictive tools as well as production models are significant elements of the technology of precision agriculture. A computer-supported space information database enables us to create a comprehensive picture of the agricultural fields under examination. With knowledge of spatial variability, we can intervene in any given location in an adequate way. The objective of the author's research was to analyse and eliminate errors and uncertainties in yield maps, and to describe the heterogeneity and spatial structure of cultivated plants and soil properties. As a second step, the author elaborated prediction models that describe the relationship between some of the more important soil properties and the yield.

Data and methods

The field experiment on which the dissertation is based started in 2001 in the framework of the National Development Research Programme and under the management of the Institute of Biosystems Engineering. Field data were collected on field No. 80/1 (size: 15.3 ha) of the education estate of the Faculty of Agricultural and Food Sciences of West Hungarian University. The data were collected in the period between 2001 and 2005.

Precision agriculture must begin by gathering information about the factors that determine the conditions of plant growth. Parameters influencing the productiveness of soil are most important among these.

The characterisation of soil starts with location-specific soil sampling based on a systematic grid. Soil samples were collected by DGPS navigation along a 50x50 grid in 63 treatment units with an average size of 0.25 ha.

The collected soil samples were subjected to a comprehensive laboratory analysis. Soil properties necessary for determining the amount of artificial fertiliser to be distributed and exploring the relationship with the yield were analysed in detail.

Data on yield and grain moisture were collected by yield monitoring from 2001 to 2005.

Yield was measured by an Agrocom ACT yield measurement system, installed on a Deutz Fahr M 35.80 harvesting machine. On the monitored field, the following crops were produced: maize in 2001 and 2002, spring barley in 2003, autumn wheat in 2004, and maize in 2005. Data on grain moisture and average yield were collected during harvest.

Soil compactness and soil resistance are important for assessing energy requirements. They were measured by a system measuring continuous soil draft.

Meteorological data were based on daily minimum and maximum temperatures and rainfall data collected between 1995 and 2004 at the Mosonmagyaróvár station of the Meteorological Group of the Department of Mathematics, Physics and Informatics.

Multi-year data series of maize yield were provided by the Szigetköz Research Centre of the Centre of Agricultural Sciences.

The data on yield and soil properties for five years were analysed with both statistical and geostatistical methods after technical preparations.

Descriptive statistics were used to determine location, variance and the characteristic parameters of distribution. Afterwards, filtering conditions based on these statistics were used to eliminate outliers and extreme values resulting from sampling. Data that had the same coordinates due to positioning errors were also excluded from further analysis, which was thus implemented on the filtered data base.

These statistical examinations did not provide answers to questions on the spatial location or the spatial variability and distribution of the sample. Geostatistical analysis is based on the theory of regional variables. The spatial distribution of variables was described by a variogram function. It is a useful aggregating statistic, which shows how data are *dispersed* in the field.

Two interpolating techniques were used to draw maps: the method of inverse distance weighting and kriging. Their correct application demands a precise definition of the spatial structure, which was given by geostatistical characteristics: experimental variograms and functions. The goodness of predictions was checked by cross-checking and the method of independent testing.

The number of soil samples is insufficient for drawing a reliable variogram. Therefore, data were supplemented with numbers drawn from random distributions of the variables.

Methods based on traditional Boolean algebra do not take into account the uncertainty of data or the continuous nature of ecological parameters.

The tools of *artificial intelligence* are better suited to the description of the continuous character of soil properties and the uncertainty of spatial variability. First, imprecise (no ‘sharp’) data were represented and treated as fuzzy numbers. Second, uncertain information was represented and treated as a set of fuzzy rules.

Traditional kriging was extended with the help of fuzzy kriging, which enabled treating together precisely measured data and imprecise predictions defined as fuzzy numbers in the spatial interpolation.

Fuzzy modelling is especially useful in a field of research where the relationships between components are not precisely known or the incompleteness or uncertainty of data renders statistical analysis unreliable. The essence of the fuzzy control system is a rule base model consisting of rules of the type '*if the input is A, then the output is B'*, where the values of variables are given by fuzzy membership functions. Our calculations were based on the Mamdani inference procedure.

The relationships between spatial variables were established with the help of regression and classificatory neural networks.

The optimisation of fuzzy rules was based on a bacterial algorithm, whereas the selection of inputs used for yield prediction was based on a genetic algorithm.

Results

The key question of the realisation of site specific production is whether sufficiently accurate information is gathered from the field. Measurement and sampling are two well-known sources of uncertainty. A measurement error occurs at every instance of measurement because the measuring device operates with limited precision. Sampling errors occur because we are able to observe only a small part of the object under examination. Two significant sources of the errors in collected yield data are the variable harvest breadth and the time lag between observing the harvesting position and the yield. Diverse techniques have been developed to reduce these errors. Yield data points are discrete in space, each point representing the average yield of the surrounding area.

The traditional statistical analyses that were executed detected outlying and extreme values. These errors were identified and filtered by the method of interquartile error analysis. Filtered yield data sets had a normal distribution in each year. The variability of yield significantly decreased during the five-year period. The three-year long data series for maize clearly shows this levelling: The yield in 2001 showed great variability (*CV=42%);* the yield in 2002 showed medium variability (*CV=35%);* and the yield in 2005 was characterised by low variability (*CV=12%).* The yield variability of both barley and wheat was *18 %.*

A variance analysis of yield data for maize showed that yield data vary significantly (p =0.1 %) over the years discussed. The average value increased by *1.5 t/ha* from 2001 to 2005, and by *4.37 t/ha* from 2002 to 2005. Between 2001 and 2005, the aggregate increase was thus *5.88 t/ha*.

The characteristics that appeared most important on the basis of the evaluation of the soil samples showed small and medium variability (1% ≤ CV ≤ 28%). In 2005, the variability of most characteristics was lower than in 2001: a levelling took place for some of the soil parameters. The comparison of soil properties made clear that there was a significant difference between the soil samples of the two respective years. The negative or positive changes in individual soil properties call for further investigation.

In order to analyse the spatial variability of yield, we prepared experimental variograms, and fitted (exponential and spherical) functions to the variograms.

Among the parameters of variograms, the nugget effect/sill (N/S) ratio was used to define small scale variability in order to characterise spatial structure. The spatial correlation interval shows the distance within which the values of the soil property correlate with one another. In the field analysed, the maximum interval is 240 m.

A small N/S and a large correlation interval usually mean that a soil property can be mapped with relatively great precision. According to the terminology used in the relevant literature, the years *2003, 2004 and 2005* can be described as having a weak spatial structure (*N/S ≥ 0.6*). This means that 60% of the variability of data cannot be explained and have a small-distance random variation. In 2001, the data had a structure of medium strength (N/S≈0.3). In 2002, the spatial structure could be considered strong (N/S<0.2).

The accuracy of interpolation was defined statistically, and the differences between real and interpolated values were calculated in the test set. For each interpolation, differences between estimated and measured were analysed. For the selected test sample, the errors based on the quadratic sums of differences showed that kriging and inverse distance weighting produced nearly identical values: $RMSE_{kriging}$= 0,449; $RMSE_{inverse}$=0,491. We chose a kriging that closely followed the spatial structure according to the variogram. The goodness of kriged values was calculated for each year: $RMSE_{2001}$ = 0.449; $RMSE_{2002}$ = 0.458; $RMSE_{2003}$ = 0.214; $RMSE_{2004}$ = 0.219; $RMSE_{2005}$ = 0.329.

The estimated data that were the results of kriging were used to draw yield maps. The yield maps for the years discussed showed a similar pattern concerning the variability of the yield. Thus, a relationship was assumed with the spatial structure of the important soil properties. This raises the possibility that yield maps can be used as a basis for creating an optimal sample structure for soil properties. We produced the variogram functions that describe the spatial dispersion of soil parameters. The intervals defined by the variogram functions show the average extent of the spatial structure for individual variables The picture is very similar for all the analysed variables. The greatest distance is 250 m. Sample density must be adjusted to the interval of spatial dependence so as to avoid both under- and over sampling. If sampling density is based on variograms of auxiliary data, the interval of the proposed sample is between one third and one half of the average variogram interval. In the analysed field, *the sample distance is* thus 55-*65 m*.

The variograms based on the original measured data cannot be considered as entirely stable because there are too few points of measurement. The relevant literature considers a variogram reliable if it is based on 50-100 data points. That is, the number of sample points in our analysis (63) is sufficient from the viewpoint of reliability.

In order to increase accuracy and diminish uncertainty, soil data were supplemented by ‘measurements’ simulated by relying on the distribution of soil parameters. The ‘inaccuracy’ and uncertainty of the data so compiled were managed by fuzzy sets. The membership functions describing the value intervals of individual variables were defined. The data were described by fuzzy numbers, given by a triangle-shaped membership function, $T(x| a, b, c)$, where $a \leq b \leq c$ represent the intervals of the variables at the examined site. The membership functions assigned to the parameters enable us to manage the continuity of soil properties and yield value as well as the overlapping of value intervals. Measured (sharp) data were embedded in the set of fuzzy numbers. They are special cases of fuzzy numbers, with a membership function of $\mu = 1$. On the data set constructed in this way, a fuzzy variogram was defined, whose parameters were characterised by fuzzy numbers, similarly to the representation of measured data by fuzzy numbers. The constructed variogram was given by the membership function $\mu_V=min(T_{co}, T_c, T_h)$. The experimental fuzzy variogram was then used for the interactive fitting of sharp theoretical variograms. Subsequently, the estimation was carried out by fuzzy kriging on the basis of the spatial structure. The spatial structure is described in agreement with the sharp measured data so the method is suitable for handling uncertain, ‘soft’, data. Using fuzzy numbers has several advantages. One possibility is that, for areas with few accurate measured data, values are given by relying on expert knowledge. The reduction in the density of measured points leads to a decrease in kriging variance, and the result contains more information because indeterminate and uncertain information that is ignored by traditional methods is also taken into account. The result was illustrated by isolines, which could be transferred to an ASCII data set. The result of kriging was exported to the mapping programme, and the usual contour soil map was drawn. On a map that is based on fuzzy kriging, the transition of values follows real trends more closely as it reflects the continuity of soil data. The advantages of using fuzzy sets become even more pronounced when a number of variables are treated together. Variables with different units of measurement and of different orders of magnitude can be transformed to a common scale with the help of their membership functions. Therefore, the number of variables can be reduced and the variables can be aggregated in a special way.

Mathematically, a multivariate spatial interpolation can be represented in real space by a function $Z = f(x, y, v_1, \dots v_n)$, where (x, y) denote local coordinates and $v_1 \dots v_n$ denote the set of examined variables.

The above problem can be solved by a number of different models of interpolation. They include co-kriging and the structural models of direct and cross-variograms. Two deficiencies of these methods are noteworthy: it is difficult to prepare a model that corresponds to the empirical variogram, and data vectors with higher dimensions require more variograms.

The fuzzy set of soil parameters enables us to define the possible values of soil properties in a unified way. The individual values of soil properties correspond to their membership in the fuzzy set, as defined by the membership function.

The relationships between variables were determined with the aid of rule based systems. The function was given by the fuzzy relationship.

The system was constructed from factual data and rules, and processed based on inference mechanisms. The rules employed were of the ‘if-then’ type. The following type of fuzzy rules was used:
R^i*: IF (*x_1 *is* A_{i1}*) AND (*x_2 *is* A_{i2}*) AND …AND (*x_n *is* A_{in}*) THEN (*y *is* B_i*)*, where A_{ij} and B_i are fuzzy sets, x_i and y are fuzzy input and output variables.

Antecedents were linked through a fuzzy AND operator and our calculations were based on the Mamdani inference method. The Mamdani regulator facilitated a special interpolation procedure. The fuzzy membership function at the output was transformed: it was defuzzified by the centre-of-gravity method.

An important goal of our research was to establish a relationship between soil data and crop yield, and thus construct an optimal rule base. A bacterial algorithm optimised the fuzzy rule base.

The relationship between soil parameters and yield data was examined by a fuzzy rule based system with six and eight variables. Trapezoid shaped membership functions were employed to describe the rules. The functions were described by the trapezoid’s four break points. These membership functions are suitable for handling the variables’ special intervals. The membership functions contained in the rules cover the variables’ entire intervals and thus ensure continuity. Each membership function was identified by two indexes: the j^{th} input variable of the i^{th} rule was identified by $A_{ij}= (a_{ij}, b_{ij}, c_{ij}, d_{ij})$ while the membership function of the output in the i^{th} rule was $B_i = (a_i, b_i, c_i, d_i)$, where $(a \le b \le c \le d)$.

In the fuzzy interpolation procedure, input variables included local coordinates (x, y), organic matter in soil (%), soil pH value, soil texture, P, K, and soil resistance. The output variable was crop yield.

In the interpolation, all rules of the rule base were evaluated, with the help of the membership functions and the truth values obtained from the inputs.

The obtained rules correspond to a special fuzzy-like function definition. The simulation result: "predicted yield value" on the site (*4.184, 4.396, 5.264, 6.010*) *if 1st rule, (3.892, 5.734, 7.217, 7.932), if 2nd rule*"… does not contain sharp boundaries but fuzzy overlaps.

We examined how well the fuzzy rule base fitted the training samples. The difference between the output desirable according to the samples and the output calculated by the fuzzy system was examined. An important note must be made on the intervals of the variables: it is worth setting the narrowest intervals possible for the variables, which are still valid for the entire area under examination.

With regard to estimating maize yield in 2001, ten rules jointly produced the following result: the error of outputs after interpolation, compared to the inputs, was *12.3%.* By fine-tuning the parameters of the simulation, the rules were trained further, which resulted in a more precise rule base. By slightly changing the lower and upper bounds for the data (i.e. their intervals), new simulations were carried out by analysing data on maize yield in 2002. The new rule base resulted in a better approximation. Each simulation produced an error below 10%.

In the next simulation, the rule base was constructed using data on winter wheat from 2004. The simulation runs now produced a very good approximation in both dimensions. The interval of the level of wheat yield is described by 15 rules. The error of simulated values versus measured values was 2.8% for 8 variables, and 2.4% for 6 variables. We obtained results of similar quality for data on spring barley yield in 2003: the error of the best rule base was 4.9%.

Our analysis showed that the goodness of the model is influenced by the characteristics of the input database. Yield variability was significantly lower for spring barley and winter wheat (CV=18%) than for maize yields in 2001 and 2002 (CV=42% and 35%). The rule base constructed through simulation was employed to evaluate additional, new, samples. The existing, trained, rule base for a new sample carried out the interpolation. The crop yield was estimated with an accuracy of 87-93%.

There is a trade-off between the complexity of the model (in our case, the number of rules or the number of membership functions) and the accuracy of approximation. The model that we have developed can accommodate 8-9 variables. For a larger number of variables, hierarchical rule bases are to be preferred.

The factors affecting crop yields, such as soil, weather and management are so complex that traditional statistics cannot give accurate results. Therefore, a more complete understanding of relationships between yield and soil-site properties is of critical importance to precision farming. A necessary first step in this process is the search for techniques suitable for identifying a functional relationship between measured soil and site characteristics and crop yield. Five site-years of crop yields and corresponding site and soil characteristics were studied. As an automatic learning tool, an artificial neural network is a useful alternative for processing the massive data set generated by precision farming production and research.

Our first goal was to evaluate the predictive ability of supervised feed-forward neural networks on a multiple site-year data set of crop yields and soil characteristics. We wanted to identify those techniques that are most capable of generalisation within each site-year on a point-by-point basis. Second, we analysed data sets concatenated with appropriate meteorological data in order to evaluate the ability of these methods to generalize over multiple site-years.

The predictive goodness of the network changes over the years: our regression function approximates the samples with an accuracy of 80 % in 2001, 85 % in 2002, and 97 % in 2005. This change over time may be due to differences in the heterogeneity of initial samples (32 %, 27% and 6 %, respectively).

We examined 94 samples, and included the following factors in our analysis: soil pH, organic matter, soil texture, total heat (GDD), May rainfall, rainfall in all critical periods, rainfall in the vegetation period, N fertilisation, and crop rotation.

These models show sufficient flexibility to use the additional meteorological variables to achieve a quality fit between maize yield and site and soil parameters.

Summary

Within a field, crop yield and the conditions of cultivation (e.g. soil fertility) vary in space and time. Precision agriculture necessitates information about the relationship between the spatial variability of soil properties and the spatial variability of crop yield. In our field experiment, we have analysed crop yield and site and soil properties based on data collected in a period of five years.

We identified the spatial variability and correlation of yield and soil properties, which enabled us to make the spatial interpolation of data more accurate. We employed fuzzy sets to manage imprecise, uncertain and missing data and information, analyse continuous soil properties and make spatial estimations.

The method of fuzzy rule based modelling was used for estimating crop yield. This method is especially suited for handling cases in which the relationships among components are not known with certainty or the incompleteness of available data precludes traditional analysis.

Using neural networks and the available samples, we constructed a complex model of the relationships between crop yield and important soil properties.

Our results contribute to the understanding of spatial variability, the management of spatial estimation and the treatment of uncertainty.

After ensuring an adequate preparation of data and drawing on expert knowledge, the methods developed in our PhD thesis may facilitate a more precise modelling of site-specific plant production.

New Research Results (Theses)

1. The constructed variograms are suitable for tracking spatial changes in yield data within a field. Moreover, the variograms are able to reflect changes and corrections due to filtering errors from the data sets.

2. Although crop yields differ significantly within the three-year data series for maize, the interval of their spatial correlation is of similar size. Yield maps serve as a good basis for developing a strategy of soil sampling.

3. The fuzzy set is an appropriate tool for analysing and modelling the examined variables (soil parameters, crop yield). The 'inaccuracy' and uncertainty of data can be handled by the introduced membership functions. They account for the overlapping of individual value intervals and the continuity of soil properties and crop yield.

4. Variables with different units of measurement and of different orders of magnitude can be transformed to a common scale with the help of their membership functions. Therefore, the number of variables can be reduced and the variables can be aggregated in a special way.

5. The author employed the inference system based on a fuzzy rule-base to identify the spatial interpolation of crop yield with 6 and 8 variables. The crop yield in the field was approximated with high levels of accuracy (88% and 97%).

6. The author trained the constructed regression neural networks with the data in the multi-year sample and thus managed to estimate the measured crop yields with an accuracy of 90%.

List of Publications

Varga Haszonits Z. – **Mike-Hegedűs F**.(1993) Az éghajlati változékonyság és a növénytermesztés. Crop Production 42, 361-373.

Varga Haszonits Z. – Schmidt R. **- Mike-Hegedűs F**.(1994) Az éghajlati változékonyság és a gazdasági növények. Crop Production 43, 485-497.

Mike-Hegedűs F.- Varga Haszonits Z. – Schmidt R (1994: A meteorológiai tényezők hatása a kukorica fejlődésének ütemére. Acta Agronomica Óváriensis 36, 51-66.

Koltai G. – **Mike-Hegedűs F**. –Palkovits G.-Schummel P. (2002) Az őszi búza terméseredményei a talajvízszint és a tápanyagellátás függvényében a Szigetközben. Növénytermelés 51, 61-69.

Koltai G. –**Mike-Hegedűs F**. –Palkovits G.- Schummel P. (2002) A kukorica terméseredményei a talajvízszint és a tápanyagellátás függvényében a Szigetközben. 1, 581-593.

Mike-Hegedűs F. (2006**)** The Estimation of Maize Yield by Neural Network. Acta Agronomica Óváriensis (accepted for publication)

Proceedings

Mike-Hegedűs F.-Mesterházi P.-Neményi M. (2005) A fuzzy logic analysis of soil properties for site specific crop production. Informatika a Felsőoktatásban Konferencia kiadvány

Mike-Hegedűs F.-Mesterházi P.-Neményi M. (2005) A termőhely specifikus növénytermesztés adatbázisának elemzése a fuzzy logika módszereivel. Magyar Biometriai és Biomatematikai Konferencia

Czimber Gy.-Koltai G.- **Mike-Hegedűs F**.-Palkovits G.-Pinke Gy.-Schummel P. Szabó P.(2006) Vegetation over rudeal areas and changes in the moisture content of soils. Danube Monitoring Scientific Conference