

Soft computing-based decision support tools for spatial data

Serge Guillaume¹ Brigitte Charnomordic², Bruno Tisseyre³, James Taylor⁴

¹ *Irstea, UMR ITAP, 361 rue J.F.Breton - BP 5095, F-34196 Montpellier, France*
E-mail: serge.guillaume@irstea.fr

² *INRA/SupAgro, UMR MISTEA 2 Place Viala, F-34060 Montpellier, France*
E-mail: bch@supagro.inra.fr

³ *SupAgro, UMR ITAP, 2 Place Viala, F-34060 Montpellier, France*
Email: tisseyre@supagro.inra.fr

⁴ *CLEREL, Dept. of Horticulture, CALS, Cornell University, USA*
Email: james.taylor@cornell.edu

Received 1 December 2012

Accepted 1 March 2013

Abstract

In many fields, due to the increasing number of automatic sensors and devices, there is an emerging need to integrate georeferenced and temporal data into decision support tools. Geographic Information Systems (GIS) and Geostatistics lack some functionalities for modelling and reasoning using georeferenced data. Soft computing techniques and software suited to these needs may be useful to implement new functionalities and use them for modelling and decision making. This work presents an open source framework designed for that purpose. It is based upon open source toolboxes, and its design is inspired by the fuzzy software capabilities developed in *FisPro* for ordinary non-georeferenced data. Two real world applications in Agronomy are included, and some perspectives are given to meet the challenge of using soft computing for georeferenced data.

Keywords: Fuzzy, zone, learning, geostatistics, FisPro, GeoFIS, agronomy, environment

1. Introduction

Management of complex systems cannot only rely on a thorough mathematical modelling. Decision support systems are necessary to assist the decision maker, and system design should benefit from all the available knowledge, including expert knowledge and data.

In many application fields, for instance in Agonomic and Environmental Sciences, the considered data are often georeferenced and temporal data. They come from measurements (satellite or aerial images, embedded sensors e.g. yield, harvest com-

pounds, etc.), manual sampling (soil analyzes) or may be given by experts (flood-risk area). There is a need for aggregating heterotopic data of various kinds (expert, measurements), from different sources, with various spatial resolutions, protocols and assessments. Imprecision, partial truth, and uncertainty are a recurring characteristic.

Much effort has been made to design dedicated software for spatial data management, mainly Geographic Information Systems (GIS) used to handle and display georeferenced data, and geostatistical methods for data processing and estimation. Nevertheless, there have been relatively few soft com-

puting developments to address the specific characteristics of georeferenced data. Even if some GIS propose fuzzy methods, like the popular fuzzy clustering algorithm, fuzzy *k-means*, these methods are not designed specifically for georeferenced data.

Soft computing techniques, especially fuzzy logic and fuzzy inference systems, proved to be efficient to cope with imprecise data and uncertainty attached to expert judgment and have already been used in Agronomic and Environmental Sciences^{1,2,3,4,5,6,7}. Spatial data specificities are likely to open novel research topics in soft computing. For instance, the notion of zone is not clearly defined in GIS, it is often mistaken for a projection of a classification achieved in the attribute space without considering geographic continuity. This concept is central in spatial reasoning and essential in decision making, particularly in these fields. As in practice, decisions need to be applied to management zones, satisfying geographical contiguity and shape criteria. For realistic decision support, zones must be defined with respect to the imprecision and uncertainty of available data and knowledge.

This work presents the outline of a decision support system framework for spatial data. It is freely available and based upon open source tool-boxes as well as on the authors' experience in soft computing software, through the former development of *FisPro*^a, that offers a high level of semantics and human-machine interaction. It could be part, as a spatial package, of a wider project like the GNU Fuzzy one proposed in the 2007 Fuzz'IEEE Conference⁸.

The paper organization is as follows. The next section presents a state of the art of the available software environments for spatial data. The proposed architecture, including *FisPro* and the *GeoFIS*^b framework, is introduced in Section 3. Section 4 presents a soft computing-based distance, available in *FisPro* and *GeoFIS*. The framework functionalities are illustrated with two real world applications in Section 5. Finally, Section 6 summarizes the main conclusions and the open challenges.

^a<http://www7.inra.fr/mia/M/fispro/>

^b<https://mulcyber.toulouse.inra.fr/projects/geofis/>

^c<http://geotools.org/>

2. State of the art and need for specialized software

GIS are powerful systems designed to capture, store, manipulate, analyze, manage, and display geographically referenced data. They are used in many application areas, archaeology, resource management, agriculture, etc.

The most popular GIS include commercial software such as ArcGIS, JMap, MapInfo, Small-World, or open source library and software, such as GeoServer, GRASS, gvSIG, GeoTools^c, OpenMap, Quantum GIS, Udig or SAGA.

GIS use digital data and a spatio-temporal (space-time) location as the key index variable for all information, allowing information from different sources to be related by accurate spatial information. They include a vast range of spatial analysis techniques, among them contour lines, topological and hydrological modelling, map overlay, geocoding, geostatistics and classification. In a GIS, geographical features are often expressed as vectors, by considering those features as geometrical shapes: points, lines or polygons. A spatial data set with a given geometry constitutes a layer. Alternatively, a layer can also be constituted by a raster data set. Map overlay uses the combination of several of these layers to create a new output, visually similar to stacking several maps of the same region. Elementary operators are available, such as union, intersection and symmetric difference.

Geostatistics relies on statistical models based on random variable theory to produce field estimations from data points, by modelling the uncertainty associated with spatial estimation and simulation. It includes interpolation methods to complete the input data collected at a number of sample points.

Despite these powerful tools, GIS lack some functionalities for modelling and reasoning using georeferenced data. Geographic information is displayed for informing decision making, but there is neither a clear definition nor handling of some concepts, for instance the zone concept, which is often

confused with the class concept. GIS focus on providing tools for multi-criteria decision-making, for instance for site selection and suitability. However the concept of learning from data is not explicit. To our knowledge, zone learning, zone operators and features for dynamic evolution of zones seem not to be available.

Another notable point is the limited use of soft computing techniques in GIS, though reasoning about space often has to deal with some form of uncertainty or imprecision. Recent add-ons to ArcGIS include fuzzy operators for map overlay and fuzzy classification. The concept of a linguistic variable is used to model the inaccuracies in attributes and in the geometry of spatial data. Data are fuzzified through membership functions and overlay operators are applied on membership values instead of raw data. An add-on to GRASS provides fuzzy membership functions, fuzzy operators and fuzzy rules to implement fuzzy inference systems for classification tasks.

Fuzzy *k-means* clustering may be used for mining GIS data. In recent work⁹ the authors propose an extended fuzzy *k-means* method for GIS, that allows cluster centers to be hyperspheres, and apply it to find fire-point event hotspots from georeferenced data. Recent publications, for instance¹⁰ that uses a fuzzy GIS-based spatial multi-criteria framework for irrigated agriculture, take place in the application fields of Agronomic and Environmental Sciences.

On a different note, several advanced packages (*spatial*, *geoR*, *gstat*...), are available for the open source R¹¹ software. They provide multivariate geostatistical functions for kriging, analysis and simulation, and often include GIS support (GRASS for *gstat*) for querying data and executing scripts. They are intended for researchers or engineers having a good background in Statistics. SAGA (System for Automated Geoscientific Analyses)^d offers an open source comprehensive set of geoscientific methods.

The need for modelling using georeferenced data is increasing, in many application fields, but particularly so in Life Sciences. The large amount of available spatial data has begun to open new avenues of scientific inquiry into behaviors and patterns of pre-

viously considered unrelated information. However, the software tools presented above, including GIS and R, are complex and require lengthy training and specialised skills to be taken over. This is a limiting factor for the practical use of spatial modelling in some domains, such as Agronomic and Environmental Sciences where the stakeholders are not specialists of spatial data. Moreover, the available software products lack an easy way to introduce expert knowledge, and are poor in soft computing tools.

Zadeh proposed the concept of a linguistic variable¹² to implement approximate concepts and reasoning. A fuzzy partition carries semantics and knowledge about the variable behavior. Recent work¹³ makes it possible to take advantage of the fuzzy set formalism to add a semi-supervised aspect to distance-based statistical procedures such as clustering. The semi-supervision is done by using available expert knowledge to superimpose linguistic concepts onto numerical data. This pretreatment is followed by the design of a pseudo metric based on the fuzzy partitions corresponding to the concepts. The procedure is thus an approach to combine numerical data and imperfect data resulting from a human judgment, and its software implementation can help to promote soft computing.

New software, designed to facilitate modelling using expertise as well as georeferenced data, would be most useful to stakeholders intervening at different levels of decision. Ideally it should provide some of the basic viewing functionalities of GIS and interaction with maps. Expertise and data are available, and Decision Support Systems (DSS) must integrate them. The software should be easy to use with a quick and progressive learning, and a friendly interface so that decisions can be made and updated from map viewing, learning using expert knowledge and data, and map evolution. The concept of management zones, not limited to classes, is required. To limit the necessary work, the DSS software must be open, be based on existing GIS components through available libraries, include elementary geostatistical techniques through calls to R.

It can then become an open platform for adding new soft computing developments, adapted to spa-

^d<http://www.saga-gis.org/>

tial data. Targeted users include researchers in modelling tasks, counselors in Agronomic and Environmental Sciences and also teachers in those fields.

3. Proposed architecture

The DSS architecture is shown in Figure 1.

The figure is divided by a dashed line: the left part includes the components involved in the GeoFIS design while the right one illustrates how they are used.

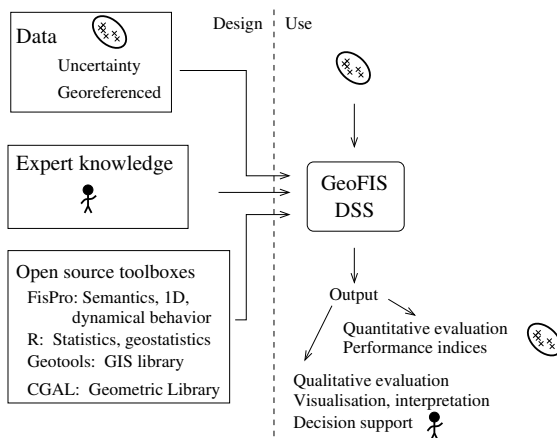


Fig. 1. GeoFIS architecture.

The data under consideration are georeferenced data. Another characteristic of the data available for the decision maker, especially in life sciences like Agronomic and Environmental Sciences, is their uncertainty. This is due to biological variability but also to the necessity of using poorly defined concepts, such as flood-risk area.

Expert knowledge is central in decision making. The DSS should be oriented towards the service of the decision maker, his/her knowledge being given the leading part.


In the proposed architecture, various open source toolboxes and libraries are used for the cooperation between expert knowledge and data. Statistical and geostatistical functions are implemented in the R project^e and, among the available GIS libraries, GeoTools is chosen because it includes

all of the necessary concepts and is written in Java, a good language to design friendly interfaces. CGAL (Computational Geometry Algorithms Library)^f provides efficient and reliable geometric algorithms in the form of a C++ library.

The FisPro environment offers a high level of interaction between expertise and data for designing and optimizing fuzzy inference systems. It is not designed to handle geographic data, but can be useful for instance to build composite variables by approximate reasoning or to design fuzzy partitions in the attribute space. Available on the Web since 2002, it is widely used in different fields and for various purposes (education, research, commercial).

FisPro's main functionalities, which are detailed below, inspired the GeoFIS framework. The goal is to provide the decision maker not only with useful indices for a quantitative evaluation but with a user-friendly interface to make a qualitative evaluation of the whole model. Interactive modelling capabilities are a must. Specific tools needed for spatial data visualization, spatial reasoning and to investigate the spatial system behavior are under development and introduced in the GeoFIS section.

3.1. FisPro (Fuzzy Inference System Design and Optimization)

FisPro  has a C++ core and a Java interface. It allows Fuzzy Inference System (FIS) design from expert knowledge or data. Among the available fuzzy software toolboxes, FisPro stands out for system interpretability, which is a necessary condition for cooperation between expert knowledge and data.

FIS can be completely, and automatically, designed from data¹⁴. In the latter case, semantics is guaranteed at each step. The necessary conditions for fuzzy partitions to be interpretable and to implement the linguistic variable concepts have been studied by several authors¹⁵. The main points are distinguishability, a justifiable number of fuzzy sets, normalization, sufficient overlapping and coverage. These conditions are met by so-called strong fuzzy

^e<http://www.R-project.org>

^f<http://www.cgal.org/>

partitions (SFPs), such as the one shown in Figure 2. A SFP described by f membership functions (MFs) on the universe U fulfills the following condition:

$$\forall x \in U, \quad \sum_{i=1}^f \mu_i(x) = 1 \quad (1)$$

where $\mu_i(x)$ is the membership degree of x in the i th MF.

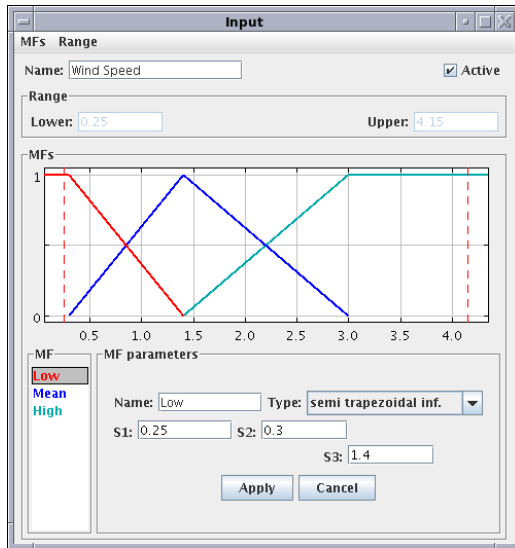


Fig. 2. A strong fuzzy partition (Fispro snapshot).

The rules share the same linguistic terms and the optimization module does not modify the FIS structure and semantics are preserved after parameter tuning.

FisPro's efficient approach in exploratory analysis and system modelling has been used to deal with agricultural applications³. Special attention has been put on the dynamical behavior of a FIS following user modifications. Each variable, rule or data item can be activated/deactivated. The system parameters (operators, partitions, rule description) can be edited. All changes are dynamically handled and all current windows are updated, including the inference result ones. Response surfaces are also available for an analysis of the system behavior.

To help the user to assess the rule representativeness, an option that evaluates the *links between*

rules and examples is available. A detailed cross-summary is given for each rule, the samples that fire this rule above a given matching degree, and for each sample, the rules that are fired.

Inference can be done manually or on the current data file, with evaluation criteria that take into account the numerical accuracy as well as the significance of data items regarding the FIS.

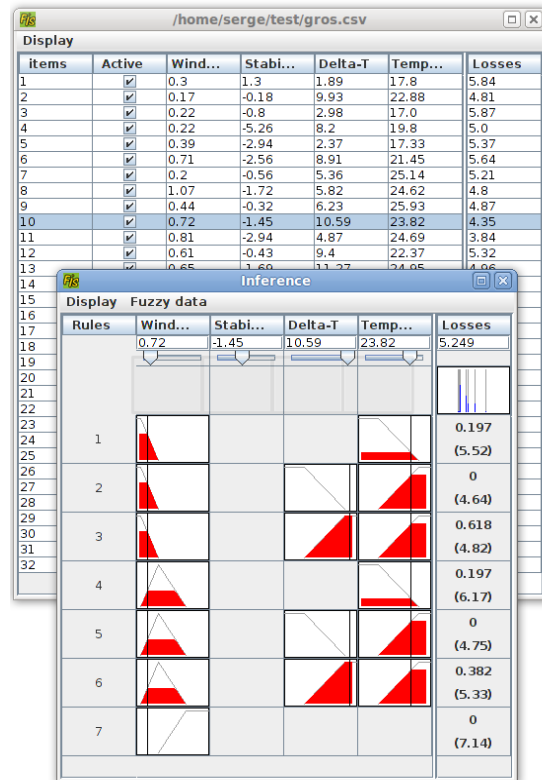



Fig. 3. FisPro Inference from the data table.

Figure 3 shows two distinct windows. The upper one shows the data as a table: a row corresponds to a data item, a column to a variable. The output variable is in the last column. A double-click on a given row opens the inference window with the corresponding input values, as shown in the bottom part of the figure. Each row corresponds to a rule. For each rule, the four first columns correspond to the input variables. The fuzzy set is shaded up to the corresponding membership degree for the given input value. The second input variable is not involved

in any rule. The last column displays the rule outputs. This being a Sugeno FIS, the rule conclusion is given in parenthesis below the rule matching degree for the current input data. The inferred output value, which results from rule output aggregation, appears in the top right corner (5.249). Modifying any FIS element would update this window.

Fuzzy inference systems are useful for building composite variables to be used in DSS. Fuzzy partitioning can be used to model uncertainties through linguistic variables, and an example will be given in Section 5.

3.2. GeoFIS (Geographic Fuzzy Reasoning)

GeoFIS  provides a simple evolutive framework to visualize and analyze spatial data. Based on open source libraries, it is written in Java and uses GeoTools to display existing data layers or generate them from raw text files. It includes calls to R to provide one-dimensional spatial analysis. It is relatively easy to implement more geostatistical techniques through calls to R spatial packages. *GeoFIS* also includes an elementary zone learning module, written in C++. Add-ons will allow to introduce new learning methods into the framework, in particular soft computing ones.

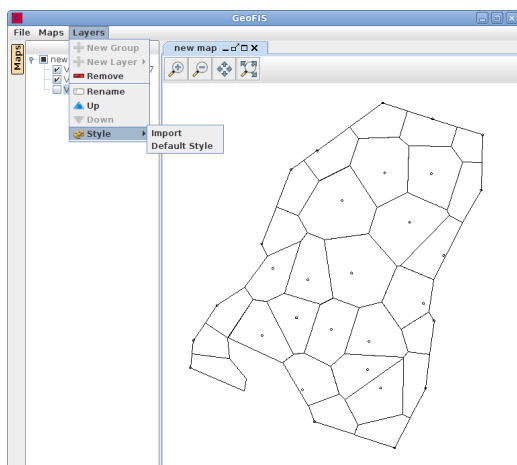


Fig. 4. The GeoFIS main window.

^g<http://www.gstat.org/>

^h<http://www.rforge.net/Rserve/>

Figure 4 shows an example of a two layer map. The first layer displays the data points while the second one corresponds to their Voronoi tessellation. The Voronoi tessellation for a set of points S in the plane is a partition of the plane into convex polygons, each of which consists of all the area in the plane closer to one particular point of S than to any other.

3.2.1. One-dimensional statistical analysis

All these functionalities are implemented using the R software¹¹ with the *gstat*^g package. The R functions are used by a large research community and are well tested. The interface implemented here uses the *Rserver*^h developments, which allow the direct transfer of objects between R and Java.

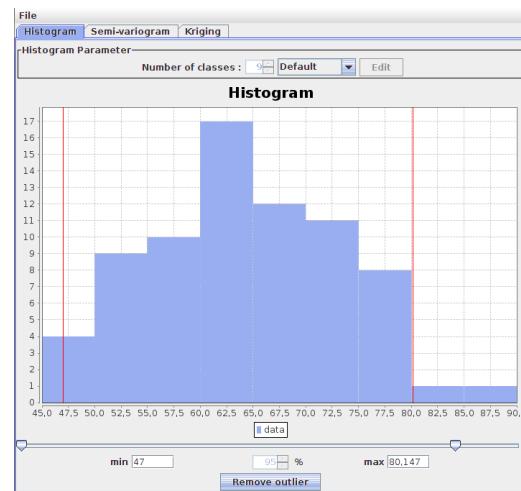


Fig. 5. GeoFIS histogram window.

The histogram window shows the distribution of data values for the selected variable. The number of classes and the class bounds can be customized. Different choices are possible, including equally spaced containers, bins with an equal number of elements, or Sturges¹⁶ algorithm for selecting the best number of classes.

Given the distribution, data can be automatically or manually filtered, to define a validity range, for instance one that holds 95% of the data, or by selecting the bounds, and so remove outliers.

The histogram window and the map viewing one are dynamically linked, so that the valid and removed data points are plotted in the latter window in two distinct colors, and updated according to the user edits in the former one.

In the case of spatial data, it is important to model the degree of spatial dependence. This is done using a semi-variogram. In *GeoFIS*, the variogram window prepares for kriging, i.e. interpolation using a defined model. The variogram model often needs expert tuning to fit the model taking into account the data set specificities (spatial resolution, shape and size of the area under study ...). All of the model parameters can be adjusted and the theoretical model (exponential, Gaussian, linear with sill and spherical), as well as the data fit, are updated accordingly.

The variogram model can be saved in standard format (xml) for reuse on new data or exported to other software.

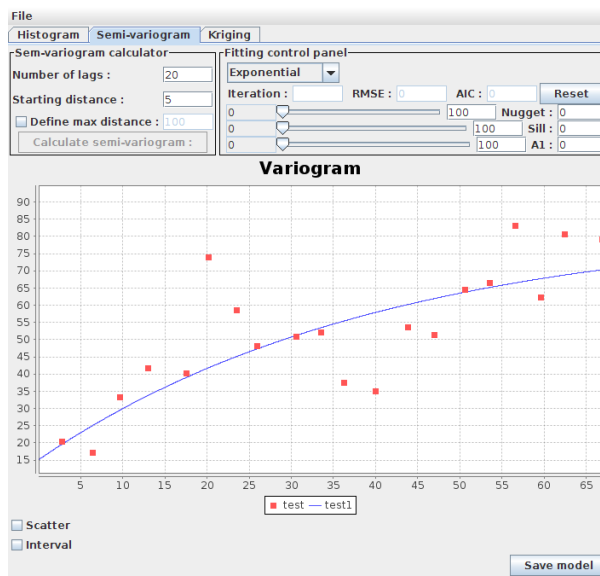


Fig. 6. GeoFIS variogram window.

3.2.2. Learning module

The zone learning module is based on a segmentation algorithm, inspired from an image-processing region merging algorithm. It allows the delineation of discrete contiguous management zones. Management in agricultural systems is dependent on both the magnitude of variation and how it is partitioned¹⁷. Segmentation algorithms differ from classification algorithms in that they are object-oriented (note: the term *object-oriented* here is used in its image analysis context, not a software engineering context). This focus leads to the production of discrete zones rather than classes and the output is spatially structured.

Fig. 7. GeoFIS zone learning parameters.

One of the disadvantages with many object-oriented segmentation algorithms is a reliance on regular grid data for determining segment morphology. This is probably an artefact from their primary application in image analysis and has restricted the use of these algorithms on irregular agro-environmental data sets. The zone learning algorithm implemented in *GeoFIS* is able to process low or high resolution data, on a regular grid or not. It is inspired from a region-merging algorithm and all details can be found in ¹⁸. A fundamental point is the way the spatial coordinates are used here. They are not involved in any distance calculation, but are only used to define point and zone neighbourhood. The algorithm works on two spaces simultaneously (attribute space and geographic space). The proximity criterion used for zone merging is based on a distance in the attribute space, and it is only calculated within a given neighbourhood. Spatial interpolation of data is not necessary for the algorithm to run. This is an asset, as interpolation generates synthetic data, whose artificial nature is often forgotten in the interpretation of the results.

Figure 7 shows the main parameters of the zone learning algorithm. It presently works on a single dimension in the attribute space, which is referred to by *Attribute column number*. Stop criteria include the number of zones to generate and a zone spatial heterogeneity based criterion. Intermediate maps may be required to allow users to see the evolution of the zone merging process. An auxiliary variable can be specified to recursively re-run the algorithm on a zone, using that auxiliary feature to guide the new zoning.

As for all segmentation or classification methods, the algorithm is sensitive to the choice of the distance in the attribute space. Options include the Euclidean distance, as well as a fuzzy partition based distance.

ⁱhttp://www.cecill.info/licences/Licence_CeCILL_V2-en.html

^j<http://mulcyber.toulouse.inra.fr/projects/geofis/>

3.2.3. Implementation details

GeoFIS currently includes the two previously described modules, one-dimensional statistical analysis and zone learning. Both modules are implemented into the Java-based interfaced general framework, and the second one is also available as a stand alone C++ program. This makes it possible to use it independently for computationally intensive automated learning tasks. *GeoFIS* is protected by a CeCILL open source licenseⁱ.

The programs are hosted on the collaborative development environment MULCYBER^j, in binary and source form.

GeoFIS handles various input and output formats: csv files, shapefiles, raster files.

4. Introduction of soft computing

In the present work, soft computing is introduced using fuzzy partition-based distances in the segmentation algorithm, instead of the classical Euclidean distance. A fuzzy partition-based distance, also called FP-based distance, allows the introduction of expert knowledge in the algorithm¹⁹. The FP-based distance combines numerical and symbolic elements. Its numerical part allows it to handle multiple membership in transition zones, while the symbolic one takes into account the granularity of the concepts associated with the fuzzy sets (see ¹⁹ for details).

The proposal applies to data in the unit interval $U = [0, 1]$ and relies on Fuzzy Partitions (FPs). The distance is called d_P . We only recall here the definition and some properties. All details can be found in ¹³.

4.1. Mono-dimensional FP-based distance

Although the pseudo-metric, d_P , is defined for general FPs¹³, its expression becomes quite simple for Strong Fuzzy Partitions (SFPs), as used in this paper.

Let $S_i = [S_i, \overline{S}_i]$ the i th MF support, defined by $\{x | \mu_i(x) > 0\}$.

Let $K_i = [\underline{K}_i, \overline{K}_i]$ the i th MF kernel, defined by $\{x | \mu_i(x) = 1\}$.

Let $X_i = [\underline{K}_i, \underline{K}_{i+1}[$, for $0 \leq i \leq f$.

We denote by $I(x)$ the function such that:

$$\forall i \in [0, f], x \in X_i \Leftrightarrow I(x) = i$$

Let us introduce the function P:

$$P(x) = I(x) - \mu_{I(x)}(x) \quad (2)$$

P is a positive non-decreasing function of x and is increasing in overlapping zones.

$d_P(x, y)$ can then be written as:

$$d_P(x, y) = \frac{|P(y) - P(x)|}{f - 1} \quad (3)$$

4.1.1. Rank inversion vs Euclidean distance

Figure 8 shows an example of rank inversion of the fuzzy partition based distance results compared with the Euclidean distance ones. With the univariate fuzzy partition-based distance d_P , y and z are further apart than x and y , while they would be closer than x and y , were the Euclidean distance used. This rank inversion is due to the fact that all elements within a given fuzzy set kernel have a null distance.

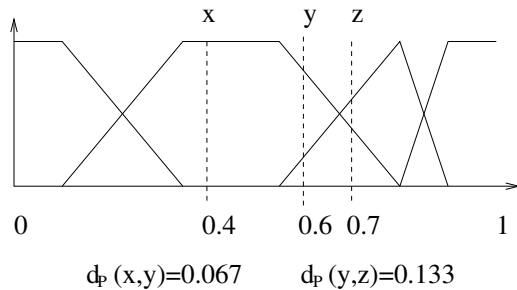


Fig. 8. Example of FP-based distance (d_P) behavior.

4.1.2. Particular case of regular SFPs

A regular SFP is composed of triangular membership functions, with equidistributed kernel centers $K_1 \dots K_f$, with $K_i = \underline{S}_{i-1} = \underline{S}_{i+1}$.

For this particular case, it was shown in ¹³ that the proposed pseudo-metric is a metric and that it

yields the same result as the Euclidean distance, regardless of the number of terms in the partition. The reason is that the distance proposed in Eq. (3) distorts the Euclidean distance according to the two following points: the symbolic distance between concepts and the indistinguishability of the kernel elements. In the case of a regular SFP, these two characteristics disappear because all kernels are reduced to single points and are equidistant.

4.2. Multi-dimensional FP-based distance

A simple and efficient way to obtain a multidimensional pseudo-metric is to perform a Minkowski-like combination of the univariate pseudo-metrics. Let two multidimensional points $x = (x_1, \dots, x_M)$ and $y = (y_1, \dots, y_M)$ with $x_i, y_i \in [0, 1]$, $\forall i \in 1, \dots, M$.

We have the following definition for the multi-dimensional distance, which is also a pseudo-metric:

$$\forall x, y \quad d(x, y) = \left[\sum_{j=1}^M (d_j(x_j, y_j))^k \right]^{\frac{1}{k}} \quad (4)$$

where k is a scalar positive value, corresponding to the Minkowski exponent. The advantage of this definition is that one can use different sub-distances in the various dimensions, for instance a FP-based distance in dimension a if expert knowledge is available for the corresponding feature, and on the contrary, the Euclidean one in dimension b .

5. Case studies

This section presents two real world agronomic applications involving spatial data and expert knowledge. The first one is monodimensional, and the second one is bidimensional.

5.1. Defining management zones in a wine growing application

The georeferenced data are yield data²⁰, coming from an embedded sensor on a grape-harvesting machine. The 1.4 ha field is planted with the Bourboulenc variety and was harvested in 2001 in Provence (France). The average sampling rate is

about 2400 measurements per ha. But, due to a data acquisition problem, some records are missing.

The objective of the study is to find suitable management zones from the information found in the yield data and the domain knowledge. Several operations could then be adapted including, for example, fertilization, winter pruning and inter row management. In this case, the grower was considering the establishment of grass in the rows located in zones of high production to introduce a competition with the vines and reduce their vigour and the resulting yield.

Let us discuss the different modelling steps made possible by the software framework.

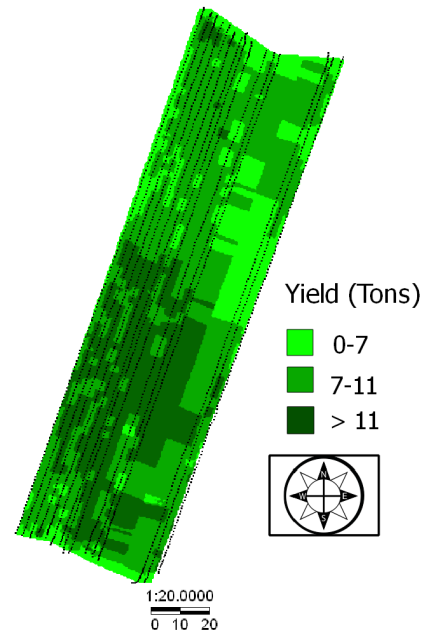


Fig. 9. Clustering vine data - three expert groups.

5.1.1. Viewing the spatial distribution

The first stage is to view the spatial distribution of the yield attribute, by splitting it into classes, and projecting it into a two dimensional map. Various methods can be used: expert definition of classes or automatic definition from data. We present here three different choices for clustering in the attribute space: a) *crisp* clustering using expert boundaries, b) *automatic k-means* with three groups and c) clustering into three *equi populated* groups. Figures 9, 10 and 11 show the respective clustered maps.

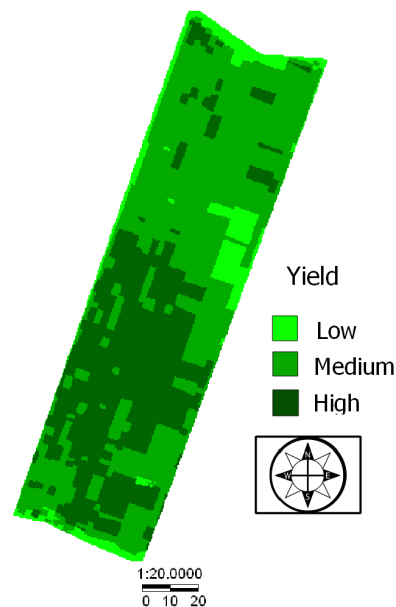


Fig. 10. Clustering vine data - three *K-means* clusters.

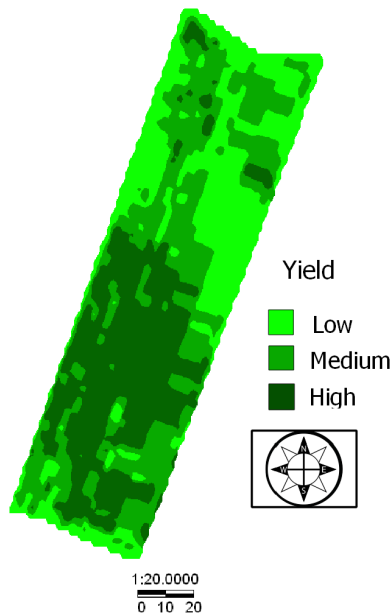


Fig. 11. Clustering vine data - three equipopulated groups.

All three maps are derived from interpolated data. Interpolation is used to represent a continuous map, so even if the sampling is irregular and/or there are gaps in the data (see Figure 9), it is possible to visualize the main spatial patterns of the field.

Each of the different types of maps is important for operational data analysis.

The map displayed on Figure 9 provides expert classes. It displays the response of the field in relation to the technical goals of the grower. The central class corresponds to the yield target, the lower and upper classes are the yields for which the vineyard operations (pruning, fertilization, etc.) are probably not appropriate. Figure 9 shows a northern zone that matches the yield goal and a southern zone for which the vine management does not seem appropriate because the yield is too high.

Other representations are necessary for operational purposes. The *k-means* classification (Figure 10) helps to identify whether there is a particular distribution of data in the plot. Equiprobable classification (Figure 11) allows to visualize the data variability, showing for example that the northern zone con-

sists of medium and very low yields. This map may be useful to highlight the effects of environmental factors (soil, elevation, etc..) which explain the observed spatial variability. In all examples, regardless of the classification methods used, the maps show discontinuous spatial patterns. Although classification is interesting for analysis purposes, the resulting maps can hardly be taken into account to propose site-specific management of the field.

5.1.2. Zoning with a Euclidean distance

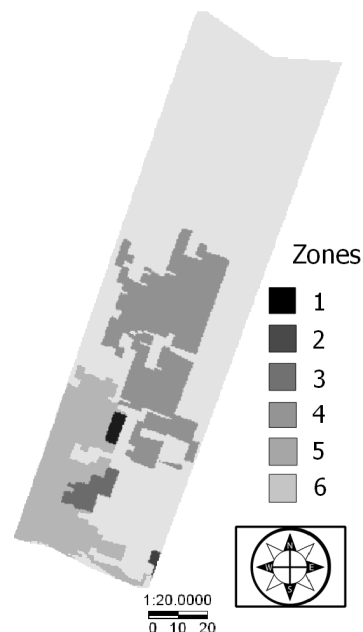


Fig. 12. Zoning vine data using the Euclidean distance.

The second stage consists in a spatial zoning of the yield data, using a Euclidean distance in the attribute space. The merging algorithm mentioned in Section 3.2 is used. It yields a series of maps with a decreasing number of zones. The six zone map is presented in Figure 12, that highlights the usefulness of zoning. It shows zones where site-specific management may be considered. However, from a practical point of view, that map remains difficult to use.

This zoning method yields zones with complex borders and does not allow a simple view of the field.

5.1.3. Zoning with a FP-based distance

The third stage improves the spatial zoning of the yield data by incorporating expert knowledge through a fuzzy partition-based distance (see Section 3.2). The fuzzy set breakpoints are 7,9,11, which are related to the choice made previously for the crisp classification. A *FisPro* snapshot is shown in Figure 13, that displays the fuzzy partition together with the data distribution.

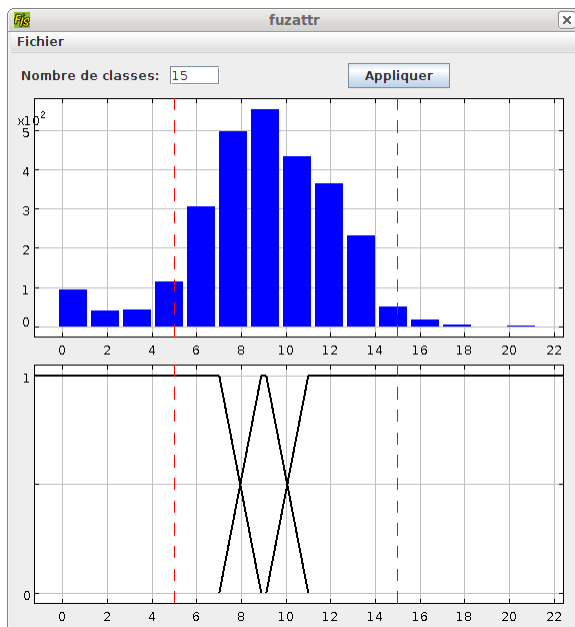


Fig. 13. Histogram and fuzzy partition for vine data.

The six zone map obtained by running the zoning algorithm, guided by the fuzzy partition based distance, is shown in Figure 14.

The introduction of fuzzy logic in the zoning method provides a map that simplifies the representation of the field. Two main management zones are highlighted, one corresponding to the northern low yield area, the other one to the southern high yield area. Note that a few specific zones of small size are also identified. They correspond to i) a zone of very high yield in the center of the plot and ii) two low yield zones located along the southern edge of

the field which are due to border effects (beginning of the rows). Depending on the goal and the machinery of the grower, these small zones may not be considered sensible for site-specific management.

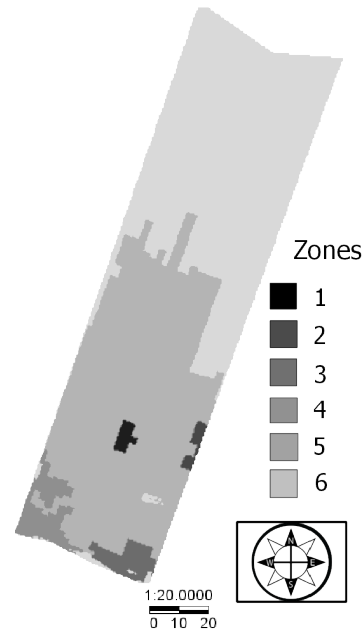


Fig. 14. Zoning vine data using a FP-based distance.

5.2. Bivariate study of yield-protein interaction in cereal production

The interaction between yield and protein is key to understanding yield potential and Nitrogen (N) budgets in cereal production systems. With on-harvester cereal yield and protein sensors now commercially available it is possible to acquire high-density yield and protein information. However, the application of agronomic decisions with co-joined yield and protein data has not been well examined to date, in part because effective measures of interrogating the data for decision-making have not been proposed.

Here we propose a two-step process using an adapted version of the univariate zoning algorithm presented in Section 3.2.2. The alternative would be to use a bidimensional distance, see Eq. (4). The advantage of the solution proposed here is the in-

interpretability of the zoning results for the decision makers.

The yield and protein data were collected in 2004 from a 80 ha field in north-west NSW, Australia, using on-harvesters sensors connected to a DGPS unit on a commercial grain combine harvester. The data were trimmed of outliers. The two data sets were at different spatial resolutions; the protein data was at a density of 65 points/ha while the yield data was at 725 points/ha.”

5.2.1. FP-based bivariate segmentation

The two-step process is as follows:

1. The trimmed (uninterpolated) yield and protein data were independently run through the zoning algorithm to give two independent results (y yield and p protein zones). The membership functions used in the process are shown in Figure 15 and 16, together with the data distribution. The Protein MFs are based on the general rules for hard bread wheat²¹.

Zones are restricted to a minimum size, which for yield was set at 350 points and for protein at 30 points. Both equate to approximately 0.5 ha. Multiple outputs [5,10,15] are generated for each univariate analysis (p/y).

2. The two results (y -zones and p -zones) from the segmentation were intersected to make a polygon layer of z segmentation-derived Yield-Protein zones.

All possible combinations of the derived p -zones and y -zones were considered. The zoning was labeled as $p05y05$ for the intersection of the 5-zone protein and 5-zone yield maps; $p05y10$ for 5-zone protein and 10-zone yield maps etc.

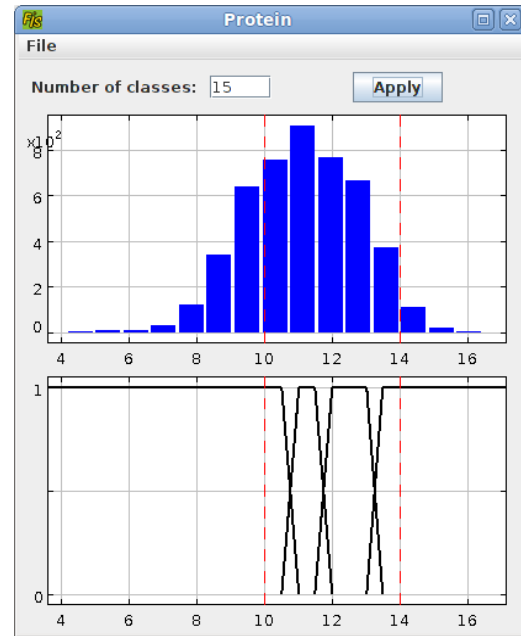


Fig. 15. Histogram and fuzzy partition for cereal protein.

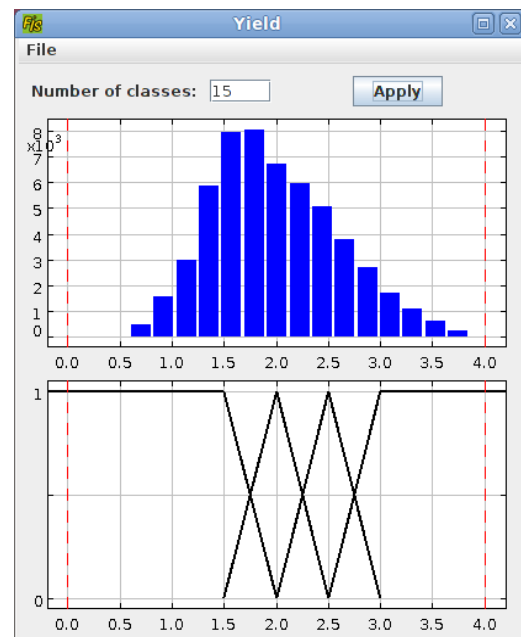


Fig. 16. Histogram and fuzzy partition for cereal yield.

5.2.2. Data clean up and results

In the segmentation approach, the univariate segmentation is constrained to ensure that the y yield zones and the p protein zones are large enough to be managed, as explained in the previous case study.

However, when these are intersected to produce Yield-Protein zones, some of the zones may be too small to be managed. Therefore the Yield-Protein zone maps need to be cleaned to produce a map that is suitable for agronomic use (and decision-making). Polygons that were less than a desired threshold (in this example 0.5 ha) were identified and removed, creating a clean polygon map that contains holes. The 5-m field grid (that was used for the interpolation) was intersected with the cleaned polygon maps. Grid points located outside a polygon were re-associated with the nearest polygon using a nearest neighbor interpolation. Grid points located inside a polygon retain the polygon features. The grid data was then transformed into a final polygon map, which is now a continuous surface.

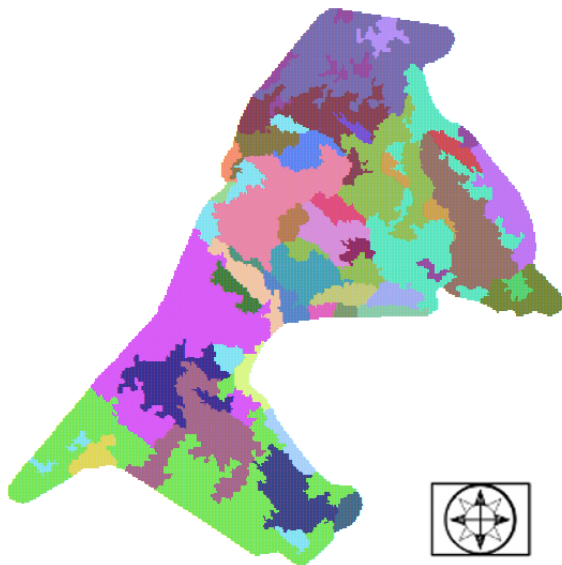


Fig. 17. Map of the cleaned 43-zone segmentation results (cereal yield-protein interaction).

5.3. Perspectives

These two case studies show the interest of incorporating expert knowledge into zoning algorithms to guide the zoning and generate more interpretable maps. This is important for decision making in many application fields. Even when dealing with multidimensional cases, the expert knowledge often remains monodimensional, as interactions are complex and difficult to grasp. Therefore, the FP-based distance is a useful tool that must be associated with aggregation operators. These operators should use soft computing to relax constraints on zone delineation and allow more flexibility in the aggregation process, while preserving the result interpretability.

6. Conclusion

Cooperation between knowledge and data is still an open challenge in system modelling. Among soft computing methods, fuzzy logic provides original efficient solutions. Its success stems from the ability to express the system behavior in a linguistic, highly interpretable way. An emerging ambitious challenge is the development of methods and software suitable for cooperation between domain knowledge and georeferenced data, also called spatial data, which are now becoming available in great quantities.

In this paper, we proposed an open source framework, based on specialized toolboxes and software, to be used for modelling and decision support. It also aims to answer some educational needs of students in these application domains, including advanced programs for developing countries where the use of open source software is an asset. We introduced a soft computing tool allowing the users to define distances based on expert knowledge by means of fuzzy partitions and to incorporate them in a segmentation algorithm.

The software functionalities are illustrated on two case studies in Agronomy, that show how they can help practitioners.

This is only a first step. For instance, it is necessary to develop specific visualization tools, in order to represent a fuzzy zone, with uncertainties in two

different spaces, the geographical space and the attribute space.

The interpretability constraints which have been implemented in fuzzy software for ordinary data, such as *FisPro*, are not so easy to define for geo-referenced data. There is no trivial extension of strong fuzzy partitions to a two-dimensional space. The development of approximate map comparison techniques and suitable aggregation operators, in order to monitor the temporal evolution of zones on a map, or to compare maps for different attributes, constitutes another topic of interest. Image analysis techniques have to be extended to include irregularly spaced data, coming from manual measurements, and domain knowledge.

Applying fuzzy logic tools, or more generally soft computing tools, to spatial data is an attractive perspective that opens new research topics, both methodological and software related.

Acknowledgments

The authors are thankful to Mr. James Hassall and Dr Brett Whelan for making the data (and their knowledge) available for this work. The cereal data was collected as part of the GRDC SIP09 project. This work was funded in part by the Agropolis Foundation.

References

1. F. Colin, S. Guillaume, B. Tisseyre, "Small catchment agricultural management using decision variables defined at catchment scale and a fuzzy rule-based system: a mediterranean vineyard case study", *Water Resources Management*, **25**, 2649–2668 (2011).
2. M. El Hajj, A. Bégué, S. Guillaume, "Integrating spot-5 time series, crop growth modeling and expert knowledge for monitoring agricultural practices - the case of sugarcane harvest on reunion island", *Remote Sensing of Environment*, **113** (10), 2052–2061 (2009).
3. S. Guillaume, B. Charnomordic, "Interpretable fuzzy inference systems for cooperation of expert knowledge and data in agricultural applications using fispro", I. C. N. CFP10FUZ-DVD (Ed.), *IEEE International Conference on Fuzzy Systems*, Barcelona, Spain, 2019–2026 (2010).
4. S. Matreata, M. Matreata, "Application of fuzzy logic systems for the elaboration of an operational hydrological warning system in ungauged basins", L. Pfister, L. Hoffmann (Eds.), Uncertainties in the 'monitoring-conceptualisation-modelling' sequence of catchment research. *11th Conference of the Euromediterranean Network of Experimental and Representative Basins (ERB)*, Luxembourg, 57–162 (2006).
5. J.-N. Paoli, O. Strauss, B. Tisseyre, J.-M. Roger, S. Guillaume, "Spatial data fusion for qualitative estimation of fuzzy request zones: Application on precision viticulture", *Fuzzy Sets and Systems*, **158** (5), 535–554 (2007).
6. T. Rajaram, A. Das, "Modeling of interactions among sustainability components of an agro-ecosystem using local knowledge through cognitive mapping and fuzzy inference system", *Expert Systems with Applications*, **37** (2), 1734–1744 (2010).
7. N. Tremblay, M. Bouroubi, B. Panneton, S. Guillaume, P. Vigneault, C. Bélec, "Development and validation of fuzzy logic inference to determine optimum rates of n for corn on the basis of field and crop features", *Precision Agriculture*, **11**, 621–635 (2010).
8. D. D. Nauck, "Gnu fuzzy", *IEEE International Conference on Fuzzy Systems*, 1019–1024 (2007).
9. F. Di Martino, S. Sessa, "The extended fuzzy c-means algorithm for hotspots in spatio-temporal gis", *Expert Systems with Applications*, **38**, 11829–11836 (2011).
10. Y. Chen, S. Khan, Z. Paydar, "To retire or expand? a fuzzy gis-based spatial multi-criteria evaluation framework for irrigated agriculture", *Irrigation and Drainage*, **59** (2), 174–188 (2010).
11. R Development Core Team, "R: A Language and Environment for Statistical Computing", *R Foundation for Statistical Computing*, Vienna, Austria, ISBN 3-900051-07-0 (2008).
12. L. A. Zadeh, "The concept of linguistic variable and its application to approximate reasoning - parts i, ii and iii", *Information Sciences*, **8-9**, 199–249, 301–357, 43–80 (1975).
13. S. Guillaume, B. Charnomordic, P. Loisel, "Fuzzy partitions: a way to integrate expert knowledge into distance calculation", *Information Sciences*, **in press** (2013), doi:10.1016/j.ins.2012.07.045.
14. S. Guillaume, B. Charnomordic, "Learning interpretable fuzzy inference systems with fispro", *Information Sciences*, **181**, 4409–4427 (2011).
15. J. V. de Oliveira, "Semantic constraints for membership functions optimization", *IEEE Transactions on Systems, Man and Cybernetics. Part A*, **29** (1), 128–138 (1999).
16. H. A. Sturges, "The Choice of a Class Interval", *Journal of the American Statistical Association*, **21** (153), 65–66 (1926).
17. M. Pringle, A. McBratney, B. Whelan, J. Taylor, "A preliminary approach to assessing the opportunity for

- site-specific crop management in a field, using yield monitor data”, *Agricultural Systems*, **76** (1), (2003) 273 – 292.
18. M. Pedroso, J. Taylor, B. Tisseyre, B. Charnomordic, S. Guillaume, “A segmentation algorithm for the delineation of management zones”, *Computer and Electronics in Agriculture*, **70**, 199–208 (2010).
 19. S. Guillaume, B. Charnomordic, P. Loisel, “A numerical distance based on fuzzy partitions”, S. Galichet, J. Montero, G. Mauris (Eds.), *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-2011) and LFA-2011*, Annecy, France, 1000–1006 (2011).
 20. B. Tisseyre, A. B. McBratney, “A technical opportunity index based on mathematical morphology for site-specific management: an application to viticulture”, *Precision Agriculture*, **9**, 101–113 (2008).
 21. W. M. Strong, I. C. Holford, “Fertilisers and manures, Sustainable crop production in the subtropics”, AL Clarke, PB Wylie, 214–234 (1997).