Precision Agriculture

An Optimisation-Based Approach to Generate Interpretable within field zones

--Manuscript Draft--

Manuscript Number:	PRAG-D-17-00134R2
Full Title:	An Optimisation-Based Approach to Generate Interpretable within field zones
Article Type:	Manuscript
Keywords:	management classes; zoning criteria; contrast indicator; segmentation
Corresponding Author:	patrice loisel INRA Centre de Montpellier FRANCE
Corresponding Author's Institution:	INRA Centre de Montpellier
First Author:	patrice loisel
Order of Authors:	patrice loisel
	Brigitte Charnomordic
	Hazaël Jones
	Bruno Tisseyre
Funding Information:	

PRAG1759_17-134

An Optimisation-Based Approach to Generate Interpretable within-field zones

Patrice Loisel[a*], Brigitte Charnomordic[a], Hazaël Jones[b], Bruno Tisseyre[b].

Abstract

The paper proposes a numerical criterion to evaluate zoning quality for a given number of classes. The originality of the criterion is to simultaneously quantify how zones are heterogeneous on the whole field under study and how neighbouring zones are similar. This approach allows comparison between maps either with different zones or different labels, which is of importance for zone delineation algorithms aiming at maximizing inter-zone variability. In addition, this study also proposes an optimisation procedure that yields interpretable within-field zones in which each zone is assigned a clear label. The zoning procedure involves contour delineation based on quantile values. The key point of the paper is to use the proposed numerical zoning quality criterion to guide the optimisation procedure showing the complementarity of both proposals in delineating relevant within-field zones.

In order to demonstrate the relevancy of the criterion, the zoning procedure and the implementation of both together, the method was tested on 50 theoretical fields with known variability and known spatial structure. A real plot with yield monitoring data was also used to demonstrate the value of the approach on a real case. Results show the relevancy of the methodology to compare maps with different zones and to sort them. Results also demonstrate the interest of the optimisation procedure to provide a ranked set of possible maps with different within-field zones**.** This set of relevant maps may constitute a decision support for practitioners who may consider additional expert information to choose the most appropriate map in the specific conditions under consideration.

Keywords: management classes; zoning criteria; contrast indicator; segmentation

Introduction

Delineation of management zones is of paramount importance for many applications in precision agriculture. Indeed, zoning makes it possible to split a block presenting a significant variability into more homogeneous spatial units (zones) in order to perform a more adapted operation. This interest justifies zoning as a support for site-specific application for many authors in the field such as for fertilization (Peralta et al., 2015, Shaddad et al., 2016), irrigation (Hedley and Yule, 2009; Meron et al., 2010), differential harvest (Urretavizcaya et al., 2014 ; Bramley et al, 2005). In this context, the automatic generation of interpretable maps (called zonings in the following) is a challenge (Pal and Pal, 1993, Pedroso et al., 2010, Peralta et al., 2015, Zane et al., 2013).

In the literature, there are two main types of methods aiming at generating zones from spatial data: i) data classification based on *a priori* classes (Lark and Stafford, 1997), followed by manual aggregation to facilitate interpretation, ii) data segmentation (Pedroso et al, 2010) inspired from image analysis.

Classification algorithms do have some limitations. Dealing with zoning, the most important is that classification algorithms do not consider spatial constraints to define contiguous classes i.e. zones. Zones delineation should be dependent on both the magnitude of variation and how the variability is spatially organised (or the morphology of the classes/zones) (Pringle et al., 2003). Classification algorithms generally address the first issue but not the second one. There have been a few attempts among others (McBratney et al., 2000, Shatar and McBratney, 2001) to spatially constrain the clustering algorithm to produce management zones instead of classes, but these have not been widely adopted. As a result, most approaches still rely on unconstrained clustering in precision agriculture (Fridgen et al., 2004 and Taylor et al., 2007).

Another major issue with the use of classification algorithms for management zone delineation is an absence of a clear statistic or indication of the optimum number of classes (Cupitt and Whelan, 2001). This usually requires some expert intervention and often leads to having small, discrete, unmanageable zones.

Segmentation algorithms that are widely used in image processing constitute a second family of methods to generate within-field zones. In these methods, the spatial properties are of major concern. This overcomes the issues of class fragmentation (split into multiples zones) usually observed with classification algorithms. Despite the use of imagery in precision agriculture, segmentation algorithms have not yet been widely applied (Roudier et al.., 2008, Pedroso et al, 2010). Segmentation algorithms either work by splitting the image into regions (Coquerez and Philipp, 1995) which all exhibit a certain level of homogeneity, or growing regions from seed points (Chassery and Garbay, 1984) until a threshold level is reached, and merging them. Drawbacks of the segmentation methods are the sensitivity to the initial choice of seeds and to the order of merging operations.

Zoning is not an easy process. First the process must create a full partition of the data space. Second it must reach a compromise to satisfy contradictory objectives on the plot: i) to minimize the intra-zone variability, ii) to maximize the inter-zone contrast between all neighbouring zones (Barghout and Lawrence, 2003, Shapiro and Stockman, 2001), iii) it must be flexible enough to take into account operational constraints (for instance minimal zone area) (Roudier et al., 2008, Tisseyre and McBratney 2008).

A zoning approach based on the evaluation of an overall numerical zoning quality criterion can provide a way to consider simultaneously all these objectives and constraints.

The aims of this paper are therefore a) to propose a numerical criterion to evaluate zoning quality for a given number of classes, b) to present an optimisation-based approach to generate interpretable maps by using this criterion c) to study the algorithm behaviour on hypothetical data and, d) to illustrate the relevancy of the approach on real data from precision viticulture.

The quality criterion

The criterion is designed in order to make the resulting zoning easy to interpret. The aim of the quality criterion proposed hereafter is two-fold: i) quantify how zones are heterogeneous on the whole field under study, which is of importance for zone delineation algorithms aiming at maximizing inter-zone variability, ii) quantify how neighbouring zones are similar, which is of importance considering most segmentation algorithms involve a fusion step aiming at merging similar contiguous zones. The novelty of the proposed approach is to consider these two aspects simultaneously and for the whole field. The key idea is to maximize the contrasts in all parts of the map.

Consider n georeferenced data points $\left(s_i, F(s_i)\right), i = 1 \dots n$ where $s_i = (\mathrm{x}, \mathrm{y})$ represents the Cartesian co-ordinates. The attribute $F(s_i)$ is a one-dimensional numerical value. A zoning is composed of p zones, and is denoted $Z = (z_l)_{l\in\{1,2\dots p\}}$.

Neighbourhood based on spatial co-ordinates

A fundamental point is the way the spatial co-ordinates are used here. They are not involved in any distance calculation, but are only used to define point and zone neighbourhood. The method can be applied either on irregular data or on grid data. That point is important as it allows using data measured at various resolutions and not necessarily at regularly spaced locations (i.e. not necessarily interpolated data). Based on the Voronoi tessellation (Okabe et al., 2000) of s_i locations, a neighbourhood is defined between data points, with $V(s_i)$ being the Voronoi polygon associated with data points s_i. Two data points are neighbours if their Voronoi polygons have one edge in common. Zones are defined as a set of contiguous points. Zone z_I is considered as a neighbour of zone z_J if at least one data point in z_I has a neighbour in zone z_J. The criterion is based on the following indicators:

Heterogeneity between neighbouring zones z_I, z_J.

This indicator considers all differences between attribute values in the two zones.

$$M_{IJ} = \frac{1}{A_I A_J} \sum_{s_i \in z_I} \sum_{s_j \in z_J} (F(s_j) - F(s_i))^2 A(s_i) A(s_j) \qquad (1)$$

In Eq. 1, $A(s_i)$ is the area of $V(s_i)$, $A_I(= \sum_{s_i \in z_I} A(s_i))$ is the area of zone z_I. M_{IJ} can be decomposed into three parts (see Appendix 1):

$$M_{IJ} = \sigma_I^2 + \sigma_J^2 + (\mu_I - \mu_J)^2 \qquad (2)$$

With μ_I (resp. μ_J) being the weighted mean of attribute F in zone z_I (resp. z_J) and σ_I^2 (resp. σ_J^2) being the corresponding weighted variance. The weights are the areas of the Voronoi polygons.

Heterogeneity within zone z_I.

The same formulae are used to define M_{II} and obtain its decomposition $M_{II} = 2\sigma_I^2$.

Contrast indicator

A scaling is done to build the contrast indicator between neighbouring zones z_I, z_J. The contrast indicator C_{IJ} is the ratio of the inter-zone heterogeneity M_{IJ} and of the average intra-zone heterogeneity $\frac{M_{II}+M_{JJ}}{2}$, for any two neighbouring zones z_I, z_J:

$$C_{IJ} = \frac{2M_{IJ}}{M_{II} + M_{JJ}} \qquad (3)$$

Following the decomposition given in Eq. 2 for M_{IJ} and M_{II}, C_{IJ} can be rewritten as:

$$C_{IJ} = 1 + \frac{(\mu_I - \mu_J)^2}{\sigma_I^2 + \sigma_J^2} \qquad (4)$$

This decomposition highlights the local standardization of the distance between two neighbouring zones, appearing in the second term.

Criterion

Finally the criterion is defined using the contrast indicator, as follows:

$$Crit(\mathrm{Z}) = min_{z_I \in Z,\, z_J \in N(z_I)} C_{IJ} \qquad (5)$$

In Eq. 5, $N(z_I)$ *is* the neighbourhood of zone z_I, i.e. composed of all zones that are neighbours of z_I. The choice of the minimum value over all pairs of neighbouring zones avoids compensations between pairs, as would be the case if average were used instead of minimum. It is the most severe criterion, and it ensures high contrast among all zones. The higher the criterion value, the better the zoning quality.

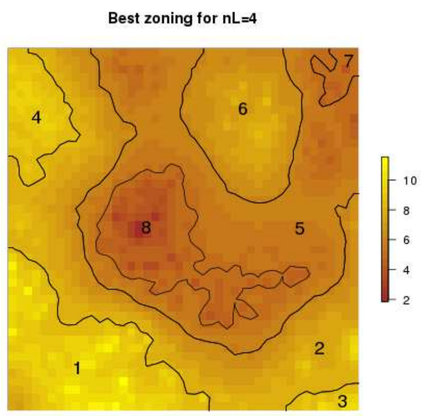
Properties

This criterion is based on statistical local indicators. To be meaningful, it requires a sufficient number of data points within zones. By definition, it is always greater than 1 and has no upper bound. In the following, the labels (or class labels) are defined, for each class, as the range of values within that class. Be aware that in this paper, these labels are not *a priori* defined. Their values result from the optimisation procedure itself

Illustration

Figure 1 and Figure 2 illustrate the criterion behaviour, on the same field (identical simulated data). Two different zonings are considered, the first one having 8 zones and the second one 7 zones. However the number of different labels assigned to the zones is the same (4) in both cases.

Each figure (high, resp. low criterion value) is divided into two parts, on the left a zoning and on the right the C_{IJ} matrix for all neighbour zones. Zone numbers are indicated on the map and are used as row and column names for the C_{IJ} matrix.

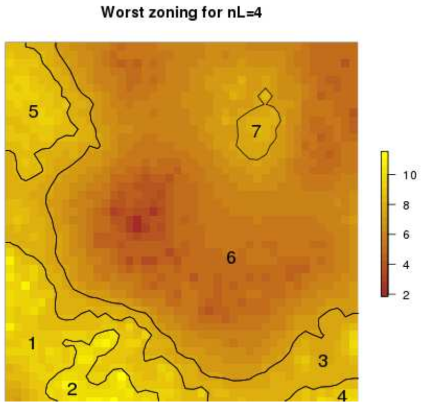


	1	2	3	4	5	6	7	8
1		4.48						
2			4.50	4.45	6.26			
3								
4								
5						4.99	3.46	3.81
6								
7								
8								

a) b)

Figure 1: a) Example of zoning with high criterion value and b) corresponding C_{IJ} matrix.

The criterion $Crit(Z)$ is the minimum value of the matrix, i.e. 3.46 in the first case, and 2.82 in the second one. Therefore the first zoning can be considered as much better than the second one.



a)

	1	2	3	4	5	6	7
1		2.82	3.34				
2							
3				3.78	4.05	6.20	
4							
5							
6							5.49
7							

b)

Figure 2: a) Example of zoning with low criterion value and b) corresponding C_{IJ} matrix.

Optimisation-based zoning

The objective of this part is to propose a zoning algorithm aiming at demonstrating the interest of the criterion. The proposed criterion is general enough to be easily integrated into different zoning methods. However, in order to better demonstrate its potential in comparing several zonings and in optimising zoning operation, it has been integrated into a simple zoning method. This zoning method takes into account an optimisation step of zone contours to show the interest of the criterion. Note that the same optimisation-based approach could be used with more complex zoning methods.
The approach takes as starting point, a zoning defined by a combination of contour lines. Contour lines delineate zones that are easy to label, all values inside the zones being either lower or higher than the cut-point. The zoning generation can then be reformulated as the question of finding the best combination of contour lines. Furthermore this guarantees a clear labelling for all zones. Labels are hereby determined by quantiles. However, emphasis must be put on the ability to structure the 'labels' to end-user requirements. if the farmer or advisor believes that a 1 t/ha change is worth managing, then contours based on 1 t changes in yield (but possibly started at different values e.g. 2.1 or 2.2 or 2.3 etc…) could be used as a starting point. Similarly, if another farmer believes he can manage to 0.5 t/ha differences, this level could be used. This may be a tool to introduce some agronomic knowledge and preference at the start of the process.

The algorithm consists of several embedded steps. Each step is thoroughly detailed below.

Use of probabilities as parameters of the procedure

There are several ways to build contour lines. A simple approach is to directly give the contour levels, either based on expert values or on automatic search for level values, based on data.
Another way, used in our approach, is to use quantiles. Quantiles are cut-points dividing the observations in a sample into contiguous intervals corresponding to given

probabilities. A quantile vector q_m univocally corresponds to a probability vector m. The choice of probability vectors simplifies the search procedure, and the corresponding cut-points automatically determine contour lines in the data map. These contour lines constitute a unique zoning for a given quantile vector.

The optimisation procedure consists in finding the probability vectors associated with the best zonings, by maximizing the criterion *Crit(Z)* given in Eq. 5. This is done by exploring a set of probabilities, each probability corresponding to a quantile value in the dataset.

Probabilities and associated quantiles are illustrated on Figure 3.

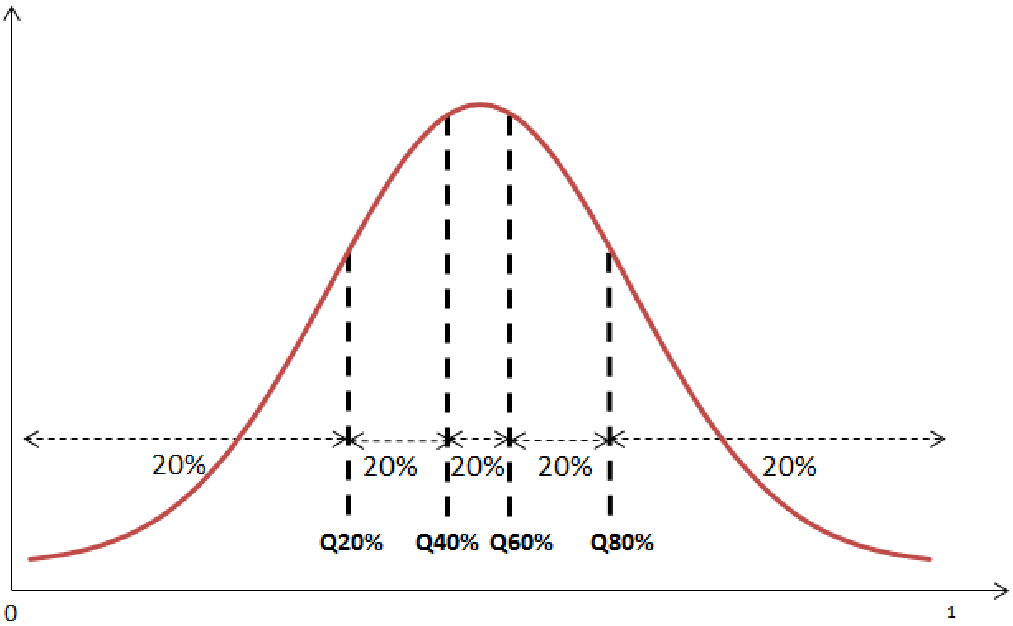


Figure 3: A Gaussian distribution. The probability vector m=(0.2,0.4,0.6,0.8) determines the quantile values at 20%, 40%, 60% and 80%.

For a given probability vector m of size $\boldsymbol{n_m}$, there are $\boldsymbol{n_L = n_m + 1}$ labels corresponding to the lower and upper limits of the quantiles. The probability vector size is therefore related to the precision required in the final zoning.

Zoning procedure for a given probability vector

The following procedure is called *Zoning(m)*, where m is a probability vector:

1. Create the initial zoning Z^m_{init} (contour lines) corresponding to probability vector m.
2. Explore the tree of possible corrections (see below for correction procedure).
3. At each step of the correction process, update the zone neighbourhood for all zones.

A set of zonings $Z^m_{(k)}$, $k = 1 \dots K_m$ is obtained, which is ranked by decreasing value of $Crit(Z^m_k)$, the best zonings (in the sense of the criterion) coming first. K_m depends on the number of corrections that were made on initial zoning Z^m_{init}.

The best zoning is denoted Z^*_m in the following.

Correction procedure: principle and illustration

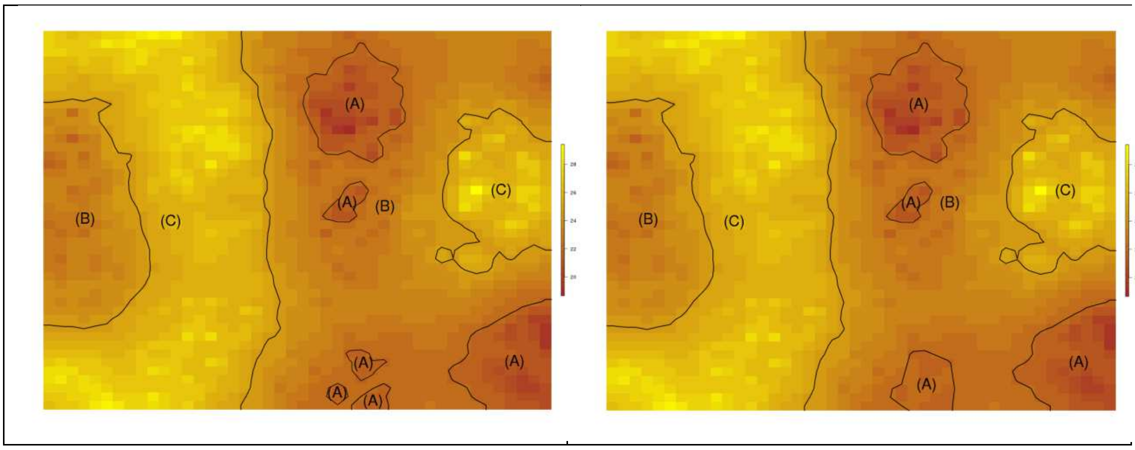
The correction procedure only considers the area (size) of the zones resulting from the contour line definition. Assuming that small zones may correspond to unmanageable regions, the correction procedure consists in modifying the initial zoning in order to respect these management constraints. The correction procedure may therefore lead to small zone removal and merging, or small zone growing. The zone size threshold is a parameter of the correction procedure. Given this threshold, two possibilities are examined and the criterion corresponding to each new zoning is computed:

1 – The *small* zone is integrated in to the closest surrounding zone to form a single zone, which is assigned the label of the surrounding zone. Therefore, the small zone label is locally lost, and this label may completely disappear from the zoning if there is no other zone carrying it.

2 - The *small* zone is expanded. If it is close enough to another zone carrying the same label, both are merged together. In all cases, the new zone keeps the same label.

At the end of the correction process, the number of zones can be drastically reduced as well as the number of labels n_L. Beyond the initial quantile vector size, the final number of labels should be considered for interpretability.

Figure 4a displays an initial three-label zoning corresponding to a probability vector $m = (0.1, 0.8)$. Quantile values associated to m are $(21.89, 25.66)$. Each zone is labeled A, B or C. The labels correspond respectively to attribute values $F \leq 21.89$, $21.89 < F \leq 25.66$, $F > 25.66$.



a) *b)*

Figure 4: A three-label zoning corresponding to a probability vector $m = (0.1, 0.8)$. *a) map of three label zoning and b) example of zoning corrections aiming at removing zones with small area.*

An example of small zone growing and merging is shown in Figure 4a and 4b. After two corrective merging steps, three small zones (at the bottom center in Figure 4a) are transformed into a single zone respecting the size constraint (Figure 4b).

Optimisation procedure for a given number of labels $\mathbf{n_L}$

The criterion does not allow choosing the optimal number of labels (n_L). It is designed to optimise the zoning when n_L is (expertly) given. Therefore, the optimisation procedure is run for a given number of labels. However, in some particular cases, during the zoning correction procedure, some class labels may disappear from the zoning. In this case, the corresponding zoning is assigned to the relevant closest number of labels.

The optimisation procedure called *Opt*(n_L) uses the *Zoning(m)* procedure in order to explore the various possibilities and to find relevant zonings. As mentioned above, n_L corresponds to a number of quantiles $n_m = n_L - 1$.
Let *P*(n_m) (with cardinal p_L) the discretized set of probability vectors of size n_m. For instance, $P(3) = \{(0.1,0.3,0.5),(0.1,0.3,0.6),\ldots,(0.5,0.7,0.9)\}$, for a step of 0.1 between successive probabilities and a minimum gap of 0.2 within an *m* vector.
This set univocally corresponds to a set of quantile vectors. Practically, a loop on probability values is done with a fixed step and gap between elements of vector m.
Opt(n_L) is detailed hereafter.

Opt*($\mathbf{n}_L$) *Procedure.

For *m* in *P*(n_m)
    Run *Zoning(m)*
    Keep the best zoning Z_m^*
End for
Store the subset of the *M* best zonings, excluding the ones that loose a label in the correction process: $Z_{\mathrm{n_L}}^{best} = (Z_{(1)}^*, \ldots, Z_{(k)}^*, \ldots, Z_{(M)}^*)$.

M is chosen so that $Crit(Z_{(1)}^*) - Crit\big(Z_{(M)}^*\big) \leq Thresh$, where $Thresh$ is a tunable threshold, chosen to eliminate the worst zonings.

Figure 5 shows the general flow of the zoning optimization algorithm: from a set of probabilities, a set of initial zonings Z_{init} is designed using contour lines corresponding to the quantile values associated to these probabilities; another set of best admissible zonings Z^* is obtained by a correction procedure; the best ones among them are selected.

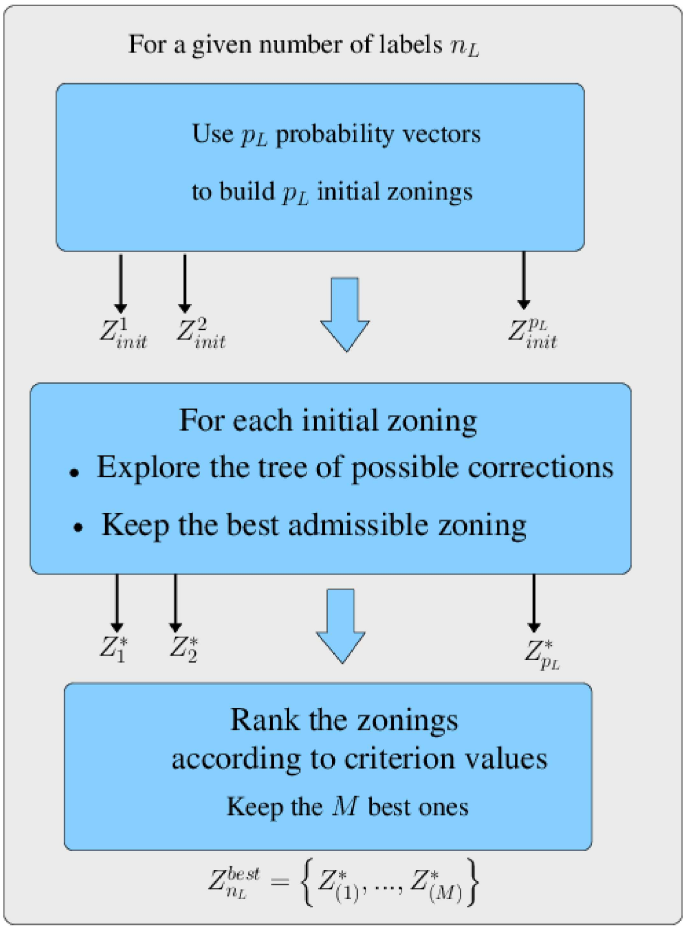


Figure 5: General flow of the zoning optimization algorithm.

Material and methods

The validation was conducted in two phases: first, by testing the method on simulated fields exhibiting Gaussian behaviour whose characteristics are known, and then by analysing real yield data.

Simulated data

50 hypothetical data sets were obtained as follows.
For each simulation, random co-ordinates were simulated at 600 locations in a square field. Then a Gaussian field (stationary isotropic covariance model, mean = 5 and standard deviation = 3) was simulated conditional on these data. The semi-variogram model had the following parameters: Nugget=1.2, partial Sill (sill minus the nugget)=5 and Range=0.2. The choice of these values was inspired by within-field wheat data from yield monitoring systems (Pringle et al., 2003).

Evaluation procedure

For each of the 50 hypothetical fields, the zoning procedure *Opt*(n_L) was applied for n_L=2 to 5. The maximum value (5) was chosen as a reasonable number of labels to summarize the information in an interpretable way. It is worth noting that n_L defines the number of labels and does not limit the number of zones.

Real data from precision agriculture

Real data were from a crop field located in Peterborough, England (52° 35′ N, 0° 15′ W) and harvested with a CLAAS combine with a 6 m cutting bar. It contained (approx. 1000 points/ha) yield values automatically geolocated by the dGPS receiver of the combine. 3966 sites were measured, the attribute $F(s_i)$ (yield) ranged from 0.70 to 7.13 Mg/ha, with a mean value of 4.73 Mg/ha.
The method was applied to kriged data (i.e. on data on a regular grid) to make easier the definition of contour lines. Note that it is general enough to apply to irregular spatial data if another contour generation function were used.

Data interpolation was performed using R. The interpolation method used in this study was based on punctual kriging with a grid size of 7*12 m. Semi-variogram analysis, model definition and data interpolation was performed with the gstat package in R (R 2008). The best semi-variogram model fitted on raw data presented the following characteristics (model type: Spherical, Nugget effects value: 0.07, partial Sill: 1.83, Range: 235 m).

Results and discussion

Results on simulated data

The distribution of best criteria values (Crit($Z_{\mathbf{n_L}}^{best}$)) over all simulations was computed and plotted as a boxplot for each n_L (Figure 6).

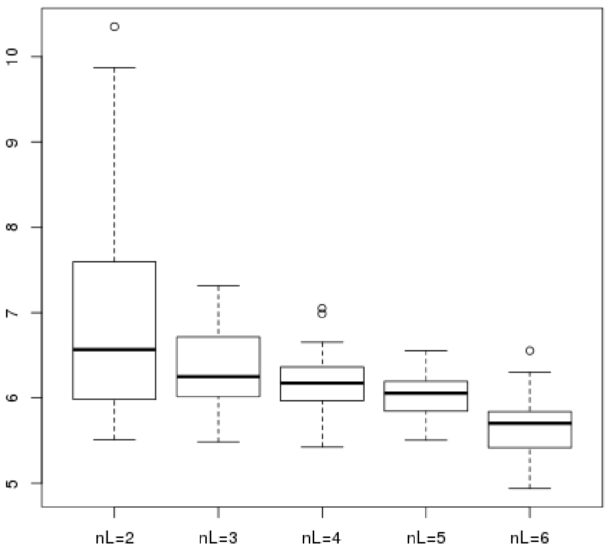


Figure 6: Distribution of best criterion values Crit($Z_{n_L}^{best}$) over all simulations.

For each n_L, thick black lines represent the median of the distribution of *Crit*, and the lines extending vertically from the boxes (whiskers) indicate variability outside the upper and lower quartiles.
The best criterion median value, as well as its quartiles, decreases with increasing n_L. Except for 2 labels ($n_L = 2$), the interquartile range is comparable for all n_L values. For $n_L = 2$, the criterion values are much more variable, with very low values sometimes, which means that the zonings are questionable in these cases.

Figure 7 shows criterion values for each of best zonings ($Z_{n_L}^{best}$) observed during a typical simulation and different n_L values. For clarity, the abscissae are centered on each n_L value. Figure 7 highlights the effect of the optimisation procedure which results in different *Crit(Z)* values for the same number of labels. Among all the corrections considered, *Crit* makes it possible to identify the zoning that simultaneously maximizes the heterogeneity between zones and the homogeneity within zones.

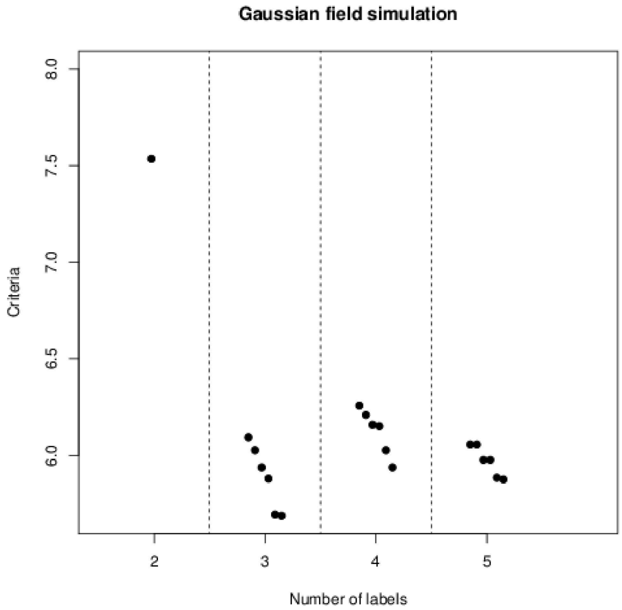


Figure 7: Criterion values (Crit) observed for each of the best zonings ($Z_{n_L}^{best}$, with thresh=0.8). For clarity, the abscissae are centered on n_L values and dash lines separate the different subsets.

For the same simulated field, Figure 8 shows zonings corresponding to the best *Crit* value for each n_L. As expected, the number of zones increases with the number of labels (n_L). Respectively 2, 6, 8, 13 zones are observed for $n_L = 2, 3, 4, 5$.

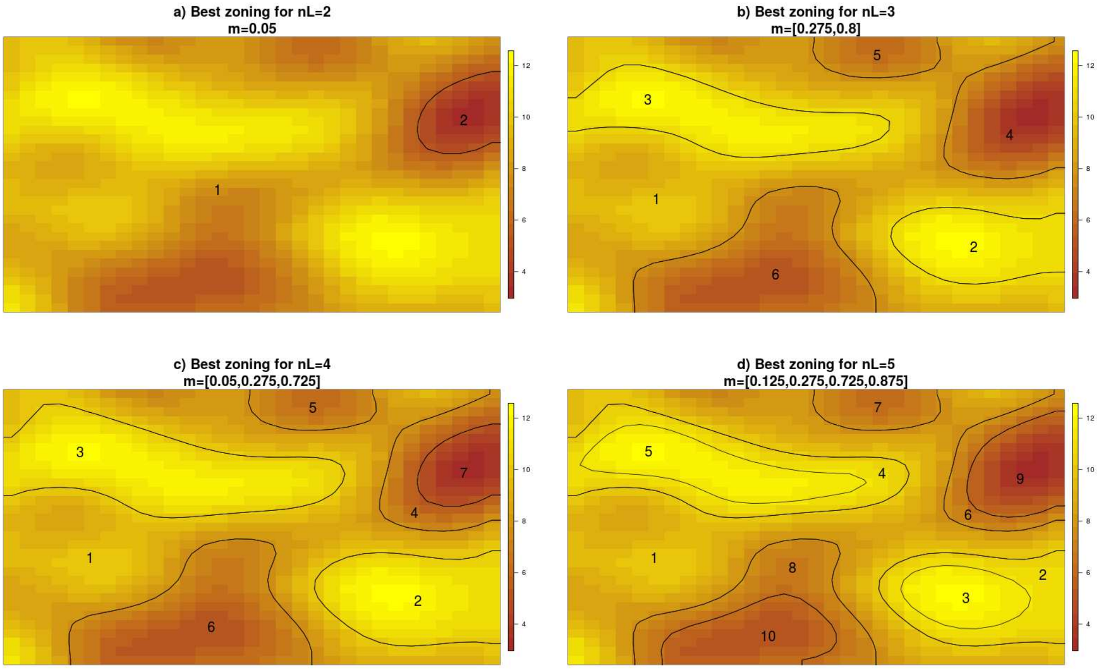


Figure 8: Best zonings for the same field and different optimisations with n_L= 2,3,4,5. The zone number is added to each zone.

The final choice of the best map depends on the granularity necessary for the decision to be applied. The ultimate selection can then be left to the end user. In Figure 8 d) some zones may be difficult to manage not because of their size, but because of their shape. This is especially the case of zones 4 and 8. These zones correspond to transition zones between core zones with clearly distinct labels.

In order to show the relevancy of the criterion to sort the possible zonings from the best to the worst, Figure 9 shows three different zonings for the same plot and the same number of labels (n_L=3), together with the distributions of within-zones attribute values. These distributions are plotted below the corresponding zonings. In the boxplots, the tick marks on the x-axes correspond to the zone numbers which are assigned by increasing mean value. The same zone numbers are indicated on the zoning.
The three zonings represented in Figure 9 correspond to the best one for n_L=3 (already shown in Figure 8), the second best one and the worst one.
Considering the best and the second best zoning, the first one corresponds to a probability vector (0.275, 0.8), while the second best one corresponds to (0.2, 0.725). They divide the field into 6 zones. Very close criterion values are observed for these two zonings (6.09 for the best and 6.03 for the second best), meaning that only small changes are expected. Regarding the worst zoning, it corresponds to the probability vector (0.65, 0.8). The low criterion value (3.18) is due to the large zone (#1) which has a high standard deviation, as highlighted by the corresponding boxplot.

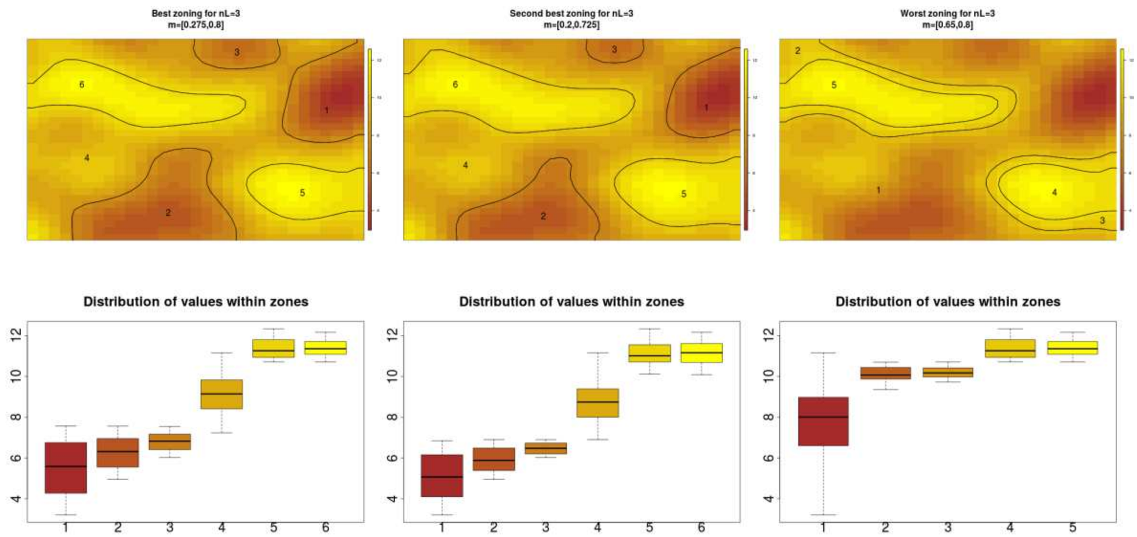


Figure 9- Top row) Three different zonings for the same plot and the same number of labels (n_L=3). Zone numbers are assigned by increasing zone mean value. They are plotted on each zone. Bottom row) Distributions of within-zones attribute values for the 3 zonings above. The tickmarks on the x-axes correspond to zone numbers.

Another point to be highlighted is that in Figure 8, the best zoning for n_L=4 roughly corresponds to the union of zonings for n_L=3 represented in the first two columns of Figure 9. This behaviour is often observed in the set of simulations that were performed.

Results on real data

The optimisation procedure $Opt(\mathrm{n_L})$ was applied for a number of labels n_L = *2, 3, 4, 5.* The best zonings obtained for yield data are shown in Figures 10 and 11.

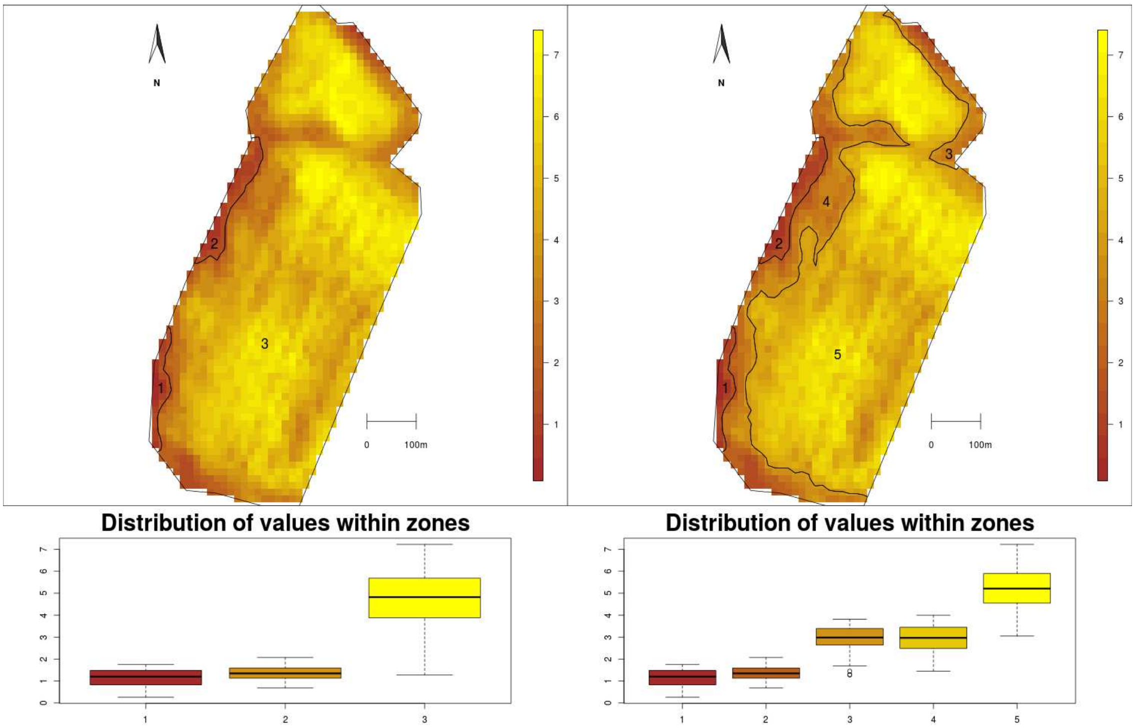


Figure 10: Best zonings for 2 labels (left – probability $\boldsymbol{m} = .05$) and 3 labels (right - probability vector $\boldsymbol{m} = (.05, .2)$). Zonings have respectively 3 and 5 zones and Crit values are 7.80 and 5.65.

As observed in the previous section with theoretical data, the criterion value generally decreases with the number of labels. Except for the zoning corresponding to 2 labels (Figure 10), the distribution of yield values within the zones shows a low variability. After a certain point, increasing $\boldsymbol{n_L}$ does not improve the number of significant zones but rather leads to a zone refinement. Results observed on yield data present the same behaviour as those observed on theoretical data showing that the proposed methodology can reasonably apply to a real data set.

Figure 11 exhibits a contrasted zoning for 4 and 5 labels as the *Crit* values shows (5.50 and 5.46). The four zonings (Figure 10 and 11) have embedded probability vectors: $(.05) \subset (.05,.20) \subset (.05,.20,.575) \subset (.05,.20,.575,.875)$. From 2 to 4 labels, the number of zones increases significantly (from 3 zones to 9 zones). After 4 labels, the increase is limited (from 9 to 11 zones). For practical use, the relevant information may be in 4 labels, as 2 and 3 labels are not enough detailed and the 5 labels zoning is just a tiny refinement of the previous zoning (4 labels).

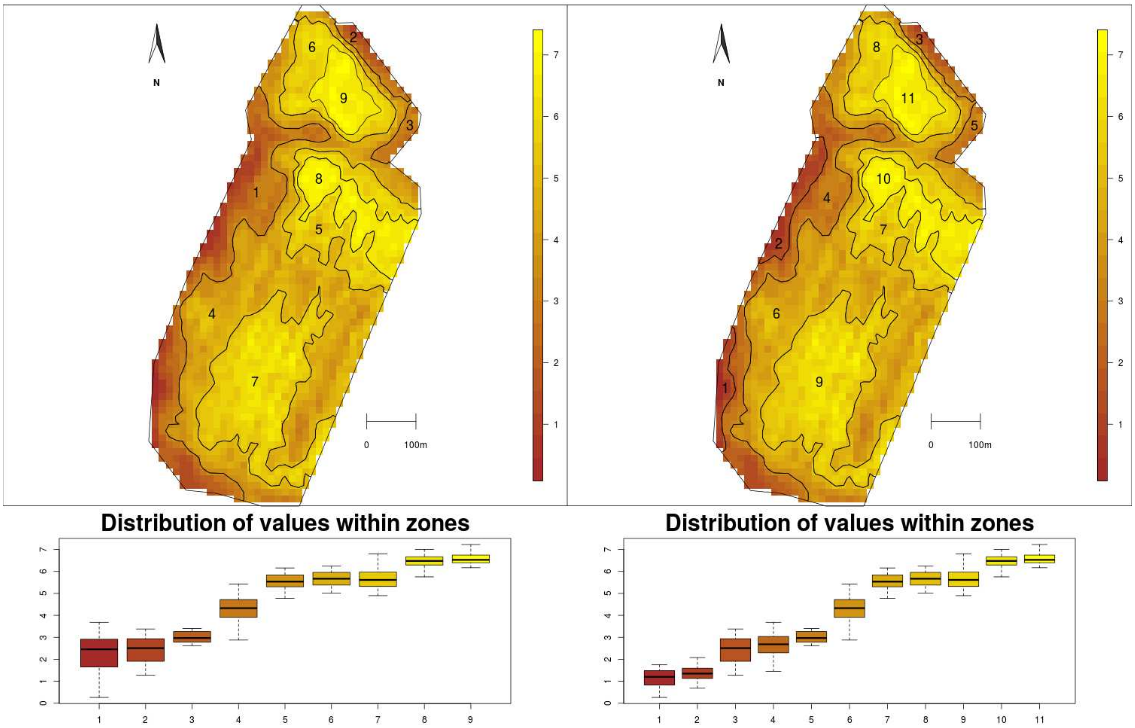


Figure 11: Best zonings for 4 labels (left - probability vector $m = (.05,.2,.575)$*) and 5 labels (right - probability vector* $m = (.05,.2,.575,.875)$*). Zonings have 9 and 11 zones and Crit values are equal to 5.50 and 5.46.*

In order to better explain the specificities of the behavior of the criterion, results were compared to those of a very common method used as a "zoning" approach in precision agriculture: k-means. Figure 12 exhibits the result of a k-means clustering using the correction procedure. Without a correction procedure, k-means results are strongly fragmented and the comparison would be unrealistic. The comparison is done on 4 and 5 labels. Without corrections, 18 and 25 zones with a *Crit* of 4.06 and 4.28, respectively, were obtained. *Crit* values are better for the method (5.50 and 5.46) than for the corrected k-means results (5.00 and 4.42), especially for 5 labels. The 4 and 5 label k-means zonings do not correspond to embedded probability vectors. This makes it difficult to compare and interpret maps.

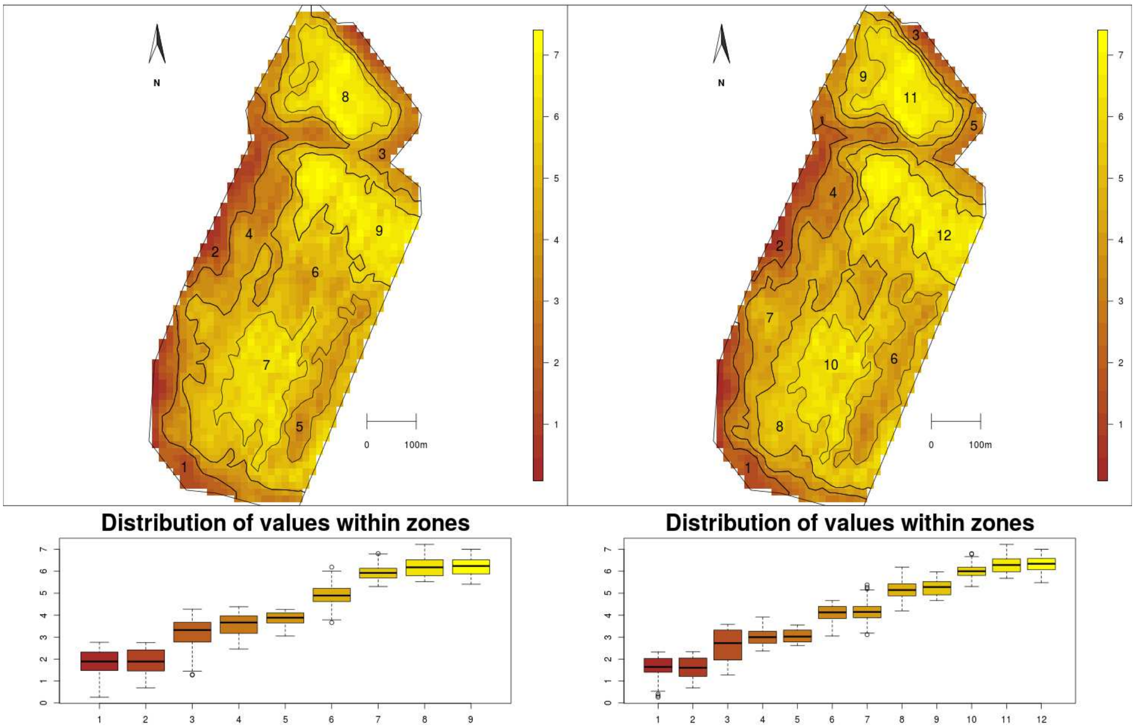


Figure 12: Best k-means zonings for 4 labels (left - probability vector $m =$ *(.12,.376,.715)) and 5 labels (right - probability vector* $m =$ *(.085,.224,.461,.759)). Zonings have 9 and 12 zones and Crit values are equal to 5.00 and 4.42.*

The zoning approach relies on a non-supervised algorithm. In the absence of a known reference for an ideal zoning, even for simulated data, the choice of the best number of labels is virtually impossible. Expert knowledge and the use case will together determine the best number, and therefore the resulting zonings. A key point is that the best labels are automatically selected by the procedure. The interest of the optimisation procedure is therefore to provide not a single solution for zoning but a ranked set of possible good zonings. Note that if expertise is available, a straightforward perspective of the method is to introduce fixed labels corresponding to classes which make sense according to application or to characteristics of the variable rate application machine.

Other information may be provided with the different maps: mean and variance of the zones, cost criteria (overall or per label) etc. All this information may constitute a support for the end-user to choose the best possible zoning in a specific context. Note that additional criteria could be added to a decision support tool: opportunity indices (Leroux et al., 2017) to assess the possibility to manage site specifically.

Practical implementation was done in R (R Development Core Team, 2008), based on standard tools for spatial data (packages sp, gstat, randomFields, and rgeos). A R package, called geozoning, is available on github.com and on the CRAN. On an Intel Core i7 3.10 GHz, run time was 188 s, for a typical *Opt*(n_L) procedure, applied to real yield data for $n_L = 5$.

Conclusion

This study proposed a criterion allowing taking into account simultaneously the homogeneity within zones and the difference between neighbouring zones on a two-dimensional map. To the authors' knowledge this is the first approach allowing the comparison between zonings either with different zones or different labels. A restriction currently applies concerning the number of labels, which must be the same to allow a fair comparison. The present paper shows the interest of such a criterion when used in association with a zoning algorithm to optimise the zones (providing that the zoning aims at maximizing the zone homogeneity and the difference between zones). In addition, this study also proposed an optimisation procedure that yields interpretable zonings, in which each zone is assigned a clear label.

This is of course a first approach. The criterion may be implemented in other zoning algorithms and, for instance, used to guide zone union choices in region merging algorithms. This opens up new research areas aiming at testing the relevancy of such a criterion when associated with other zoning algorithms. Similarly, the optimisation procedure may use other criteria.

Appendix

In Equation (1), M_{IJ} is decomposed as follows:

$$M_{IJ} = \frac{1}{A_I A_J}\sum_{s_i \in z_I}\sum_{s_j \in z_J}\left(F(s_j) - \mu_J + \mu_J - \mu_I + \mu_I - F(s_i)\right)^2 A(s_i)A(s_j)$$

$$= \frac{1}{A_I}\sum_{s_i \in z_I}\left(F(s_i) - \mu_I\right)^2 A(s_i) + \left(\mu_J - \mu_I\right)^2 + \frac{1}{A_J}\sum_{s_j \in z_J}\left(F(s_j) - \mu_J\right)^2 A(s_j)$$

$$+ \frac{2}{A_I}\sum_{s_i \in z_I}\left(F(s_i) - \mu_I\right) A(s_i)\frac{1}{A_J}\sum_{s_j \in z_J}\left(F(s_j) - \mu_J\right) A(s_j)$$

$$+ \frac{2}{A_I}\sum_{s_i \in z_I}\left(F(s_i) - \mu_I\right)A(s_i)\left(\mu_J - \mu_I\right) + \frac{2}{A_J}\sum_{s_j \in z_J}\left(F(s_j) - \mu_J\right)A(s_j)\left(\mu_J - \mu_I\right)$$

where:
$\sum_{s_i \in z_I}\left(F(s_i) - \mu_I\right)A(s_i) = \sum_{s_i \in z_I} F(s_i)\, A(s_i) - \mu_I \sum_{s_i \in z_I} A(s_i) = \mu_I \sum_{s_i \in z_I} A(s_i) - \mu_I \sum_{s_i \in z_I} A(s_i) = 0$. Hence the last three terms are equal to zero.

References

Barghout, L. & Lawrence, L. (2003). Perceptual information processing system. *U.S. Patent Application*, 10/618,543.

Bramley, R. G. V., Proffitt, A. P. B., Hinze, C. J., Pearse, B., & Hamilton, R. P. (2005). Generating benefits from Precision Viticulture through selective harvesting. *Precision Agriculture*, 5, 891-898.

Chassery, J.M. & Garbay, C. (1984). An iterative segmentation method based on a contextual color and shape criterion. *IEEE Transactions on PAMI*, 6(6),794-800.

Coquerez, J-P. & Philipp, S. (1995). *Analyse d'images : filtrage et segmentation (Image Analysis: filtering and segmentation).* Paris, France : Editions Masson, ISBN:2-225-84923-4, ISSN:0992-5538.

Cupitt, J. & Whelan, B.M. (2001). Determining potential within-field crop management zones. In G. Grenier & S. Blackmore (Eds) *ECPA 2001,* Proceedings of the 3rd European Conference on Precision Agriculture*,* Montpellier, France: agro-Montpellier, pp. 7-12.

Fridgen, J.L., Kitchen, N.R., Sudduth, K.A., Drummond, S.T., Wiebold, W.J. & Fraisse, C.W. (2004). Management zone analyst (MZA): Software for sub-field management zone delineation. *Agronomy Journal*, 96,100-108.

Hedley, C. B., & Yule, I. J. (2009). Soil water status mapping and two variable-rate irrigation scenarios. *Precision Agriculture*, 10(4), 342-355.

Lark, R.M. & Stafford, J.V. (1997). Classification as a first step in the interpretation of temporal and spatial variability of crop yield. *Annals of Applied Biology*, 130, 111-121.

Leroux C., Jones H., Clenet A. & Tisseyre B. (2017). A new approach for zoning irregularly-spaced, within-field data. *Computers and Electronics in Agriculture*, 141, 196-206

McBratney, A.B., Whelan, B.M., Taylor, J.A. & Pringle, M. (2000). A management opportunity index for Precision Agriculture. In P.C. Robert, R.H. Rust & W.E. Larson (Eds) *Precision Agriculture,* Proceedings of the 5th International Conference on Precision Agriculture, Madison, WI, USA: ASA/CSSA/SSSA, (CD publication).

Meron, M., Tsipris, J., Orlov, V., Alchanatis, V., & Cohen, Y. (2010). Crop water stress mapping for site-specific irrigation by thermal imagery and artificial reference surfaces. *Precision agriculture,* 11(2), 148-162.

Okabe A., Boots, B., Sugihara, K. & Chiu, S.N. (2000). *Spatial Tessellations – Concepts and Applications of Voronoi Diagrams.* 2nd edition. New York, USA: John Wiley, 671 pages, ISBN 0-471-98635-6

Pal, N. & Pal, S.K. (1993). A Review on Image Segmentation Techniques. Pattern Recognition, 26(9), 1277-1294.

Pedroso, M., Taylor, J., Tisseyre, B., Charnomordic, B. & Guillaume, S. (2010). A segmentation algorithm for the delineation of agricultural management zones. *Computers and Electronics in Agriculture*, 70 (1), 199-208.

Peralta, N. R., Costa, J. L., Balzarini, M., Franco, M. C., Córdoba, M., & Bullock, D. (2015). Delineation of management zones to improve nitrogen management of wheat. *Computers and Electronics in Agriculture*, 110, 103-113.

Pringle, M.J., McBratney, A.B., Whelan, B.M. & Taylor, J.A. (2003). A preliminary approach to assessing the opportunity for site-specific crop management in a field, using a yield monitor. *Agricultural Systems,* 76, 273-292.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria (http://www.R-project.org).

Roudier, P., Tisseyre, B., Poilvé, H. & Roger J-M. (2008). Management zone delineation using a modified watershed algorithm. *Precision Agriculture*, 9(5), 233-250.

Shaddad, S. M., Madrau, S., Castrignanò, A. & Mouazen, A. M. (2016). Data fusion techniques for delineation of site-specific management zones in a field in UK. *Precision Agriculture*, 17(2), 200-217.

Shapiro, L., & Stockman, G.C. (2001). *Computer vision*. Upper Saddle River, New Jersey: Prentice Hall.

Shatar, T.M. & McBratney A.B. (2001). Subdividing a field into contiguous management zones using a K-zones algorithm. In G. Grenier & S. Blackmore (Eds) ECPA 2001, Proceedings of the 3rd European Conference on Precision Agriculture*,* Montpellier, France: agro-Montpellier pp. 115-120.

Taylor, J.A., McBratney, A.B. & Whelan, B.M. (2007). Establishing management classes for broadacre grain production. *Agronomy Journal*, 99, 1366-1376.

Tisseyre, B., & McBratney, A.B. (2008). A technical opportunity index based on mathematical morphology for site-specific management: an application to viticulture. *Precision Agriculture*, 9(1-2):101-113.

Urretavizcaya, I., Santesteban, L. G., Tisseyre, B., Guillaume, S., Miranda, C., & Royo, J. B. (2014). Oenological significance of vineyard management zones delineated using early grape sampling. *Precision Agriculture*, 15(1), 111-129.

Zane, L., Tisseyre, B., Guillaume, S. & Charnomordic, B. (2013). Within-field zoning using a region growing algorithm guided by geostatistical analysis. In J V Stafford (Ed.) ECPA 2013*,* Proceedings of the 9th European Conference on Precision Agriculture*,* The Netherlands: Wageningen Academic Publishers, pp. 7-12.