Precision Agric
https://doi.org/10.1007/s11119-018-9566-5

How to define the optimal grid size to map high resolution spatial data?

B. Tisseyre[1] · C. Leroux[1,3] · L. Pichon[1] · V. Geraudie[2] · T. Sari[1]

Abstract The development and the release of sensors capable of providing data with high spatial resolution (> 4 000 points ha^{-1}) in agriculture raises new questions as to how to represent this spatial information. The objective of this study was to propose a methodology to help define the optimal grid size to map high resolution data in agriculture. The geostatistical method finds the grid size which maximizes the sum of two components: (i) the proportion of nugget variance that is removed, and (ii) the proportion of sill variance that remains in the data. The optimum grid size was found to be dependent on the resolution of the available information and the spatial structure of the raw data. Experiments on simulated datasets with varying data resolution (from 500 to 2 000 pts.ha^{-1}) and spatial structure (range of variogram between 10 and 45 m) showed that the proposed methodology was able to define varying optimal grid sizes (from 5 to 12 m). The proposed geostatistical approach was then applied on a real dataset of total soluble solids/sugar content of table grape so that the optimal mapping grid size could be found. Once it was defined, two interpolation methods: simple averaging over blocks and block kriging, were applied to mapping the data. Results show that both methods help depict the within-field variability in the data. While the averaging procedure is easier to automate, the block kriging approach provides users with a level of uncertainty in the aggregated data. Both mapping approaches significantly impacted the within-field spatial structure: (i) the small-scale variations were ten times lower than in the raw data, and (ii) the signal-to-noise ratio of the aggregated data with the optimal grid was twice as high as that of the raw data. As the proposed geostatistical methodology is a first attempt to define the optimal grid size to map high resolution spatial data, areas for future development applications are also proposed.

Keywords Filtering · High spatial resolution · Mapping

✉ B. Tisseyre
bruno.tisseyre@supagro.fr

[1] ITAP, Montpellier SupAgro, Irstea, Univ. Montpellier, Montpellier, France

[2] Pellenc S.A, route de Cavaillon, 84000 Pertuis, France

[3] SMAG, Montpellier, France

Published online: 10 March 2018

Introduction

The development of sensors such as Physiocap®, lidar, Multiplex®, Greenseeker®, etc. (Debuisson et al. 2010; Baluja et al. 2012) capable of acquiring data with a high spatial resolution (> 4 000 points ha^{-1}) raises new questions as to how to represent spatial information. Those data usually contain multiple sources of noise that can be related to (i) the sensor itself, (ii) the acquisition conditions and (iii) the short-range variations in the measured property (Taylor et al. 2010). To lessen the influence of this noise when mapping spatial information, users often aggregate data into a spatial unit (SU) with a larger spatial resolution than that of the raw information. Many terms can be used to refer to this spatial unit such as a block, a quadrat or a support. Note that there is a clear difference between the spatial resolution and the SU of the data. The spatial resolution is fixed and characterizes the number of observations within a standard area. In contrast, the SU can vary and represents an area over which data are aggregated. This data aggregation procedure generally involves a two-step process. First, there is a need to define an optimal SU or grid size to represent the spatial information. Then, data have to be resampled to match this new spatial resolution. Common approaches involve simple averaging data over those SU or more complex ones such as inverse distance weighting or kriging. With the development and release of embedded sensors, this issue of data representation becomes critical. The analyst needs to know whether it is better to (i) aggregate observations over a larger SU to suppress noise and see the spatial information in the data, or (ii) determine an optimal SU that retrieves most of the spatially structured variance in the data while keeping the noise relatively low.

The choice of an optimal SU size for data aggregation has been discussed little within the precision agriculture community. It is acknowledged that larger grids, by aggregating more data, can result in a loss of details and useful information (Blackmore et al. 2003). However, interpolating data on larger blocks lowers the error of prediction and makes it more likely to find significant differences between management zones (Bramley et al. 2011). Cressie (1996) also showed that the relationships between two or more spatial variables could be influenced by the level of aggregation of the data. In contrast, small SU will not counter the influence of noisy observations. Most studies report the grid size that is useful for aggregation, mostly cells of 1 × 1, 5 × 5 or 10 × 10 m, without clearly detailing the reason for this choice. However, some authors will provide an explanation with respect to grid size, especially due to practical or operational constraints, e.g. the resolution of auxiliary information, an appropriate size for site-specific management, or the spatial support of the within-field observations. For instance, the grid size can be related to that of another available information layer or to that of the less spatially resolute information (Acevedo-Opazo et al. 2008; Li et al. 2008; Tagarakis et al. 2013). Lauzon et al. (2005) worked with squared grids of 3, 6 and 9 m resolution for the temporal analysis of yield monitor data because those SU represent the minimum management size possible. Serrano et al. (2010) selected a grid size based on the spatial support of the within-field measurements. This support can be sensor-dependent because it integrates measurements over a given area or it can be the result of an average made on a pre-defined SU. For instance, it is common to make measurements on several plants at the same sampling site to limit random variations of the measured property for perennial crops such as vineyards and orchards. In this case, observations are often averaged and allocated to a point corresponding to the centroid of the SU (Acevedo-Opazo et al. 2008; Baluja et al. 2012).

Choice of block or grid size to aggregate spatial data, also known as the change of support problem, has been well investigated by the mining community. Journel and Huijbregts (1978) proposed one of the most commonly-reported approaches in geostatistics to deal with this issue: block kriging. Contrary to point kriging for which predictions are made at specific locations over the study area, block kriging predicts averages of the variable of interest over larger blocks (Bivand et al. 2008). Using this technique, prediction errors are known to be much lower than those arising from the point kriging approach. One useful way of using the block kriging methodology is to look for the block size which minimizes the prediction errors. By using a procedural approach and calculating the estimates errors across a large set of block sizes, one is able to find a block size that best matches the requirements of the study in terms of prediction variance. The block kriging approach is interesting in that the final objective of any prediction is to be as precise as possible. However, by solely minimizing prediction error, one does not consider how the spatial data structure is affected as the block size increases. This aspect is important to precision agriculture given that the spatial structure is what drives many field management decisions.

The aim of this study was to propose a methodology to help define the optimal grid size to map high resolution spatial data. This approach uses the geostatistical theory of change of support and aims at lowering the noise in the data while maintaining the spatially structured variance as high as possible. The methodology is first validated on simulated data with known statistical distribution, variance and spatial autocorrelation. The approach is then tested on real high resolution total soluble solids (TSS) data in a particular vineyard.

Materials and methods

Theoretical background

General classical considerations

For additive variables, it is possible to predict the change in variance related to different sizes of SU through classical statistical theory. In fact, considering a population and a group of n independent SUs with a punctual spatial support drawn from that population of mean μ and variance σ^2, the mean value of this sample ($\bar{X}$) is itself a random variable of mean μ and variance σ^2/n. When there is no spatial structure in the data and following the classical statistical theory, the variance within a quadrat is expected to decrease linearly with the number of sampling units inside the quadrat (Bellehumeur et al. 1997) as expressed in Eq. (1):

$$\mathrm{var}(Z_v/A) = \mathrm{var}(Z_p/A)/N \tag{1}$$

where $\mathrm{var}(Z_v/A)$ is the variance of the composite sample Z_v in an area A, $\mathrm{var}(Z_p/A)$ is the original variance of the punctual sampling unit Z_p in the same area A (Fig. 1), N is the number of punctual sampling units in a composite sample. A composite sample is formed by combining several adjacent SUs.

The previous equation can be considered valid solely if the SUs are independent of each other, i.e. Z_p does not exhibit any spatial structure. In case the variable Z_p is spatially autocorrelated, Eq. (1) is no longer valid. The geostatistics community has tackled this issue of change of spatial support when the spatial structure of Z_p cannot be left behind, mainly for ore reserve estimation (Journel and Huijbregts 1978) and ecological (Bellehumeur et al. 1997) studies.

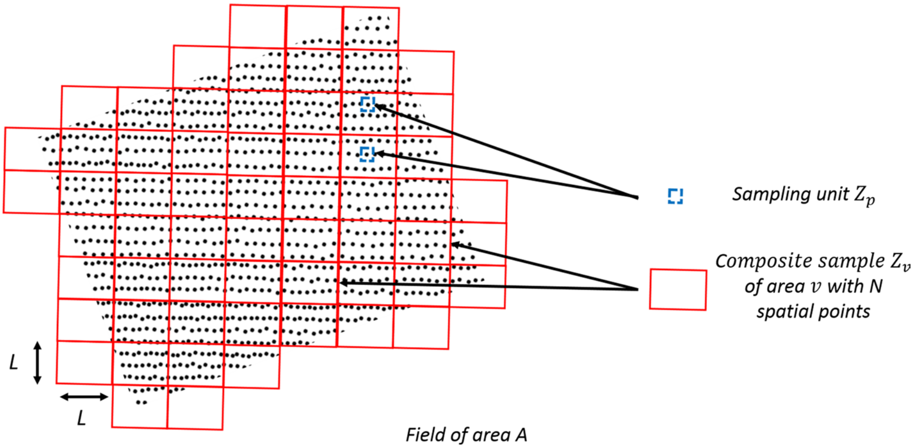

Fig. 1 Characterization of the sampling unit Z_p and the composite sample Z_v

Based on previous work, Bellehumeur et al. (1997) proposed and developed relationships to express the variogram of aggregated contiguous quadrats as a function of the variogram of a punctual support. The main relationships are summarized hereafter.

In nested designs (Fig. 2), the additivity property of variances implies the following:

$$\mathrm{var}\left(Z_p/A\right) = \mathrm{var}\left(Z_p/v\right) + \mathrm{var}(v/A) \tag{2}$$

where var(Z_p/A) is the variance of the punctual variable Z_p in an area A, var(Z_p/v) is the variance of Z_p in quadrats of size v and var(v/A) is the variance of a SU v in the area A. In the following, v will refer to a quadrat corresponding to a given SU and Z_v will refer to the variable resulting from a composite sample corresponding to a SU v.

From Eq. (2) and assuming stationarity, Bellehumeur et al. (1997) defined the variogram parameters of aggregated quadrats of size v as follows (Eqs. 3, 4 and 5).

$$C_{1_v} = C_{1_p} - \bar{\gamma}(v, v) \tag{3}$$

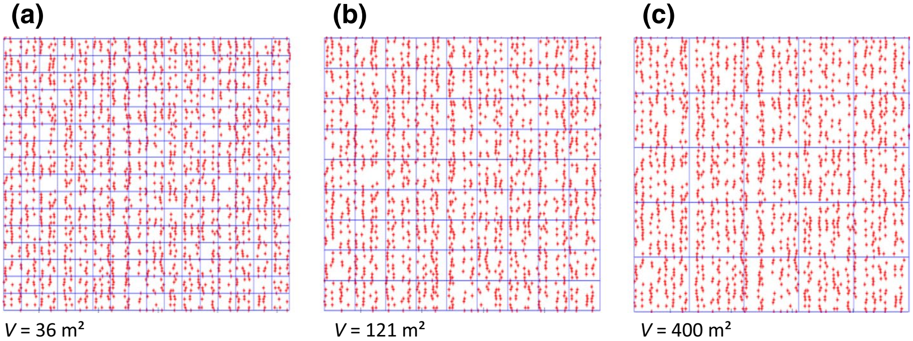

Fig. 2 Three examples of different size of sampling unit (v) over a hypothetical field of 1 ha (Data2), **a** v = 36 m^2, **b** v = 121 m^2, **c** v = 400 m^2

$$a_v = a_p + L \tag{4}$$

where C_{1v} and C_{1p} are the proportions of variance corresponding to the spatially correlated component for Z_v and Z_p respectively, $\bar{\gamma}(v, v)$ is the average point variogram value calculated over all possible distance vectors contained in v, a_p is the range of the variogram for Z_p, a_v is the range of the variogram for Z_v and v is the quadrat of side length L so that $L \times L = v$.

Calculation of the nugget effect C_{0v} when increasing the grid size

Following Eq. (1), Eq. (5) describes how the nugget effect associated with aggregated quadrats of size v is derived (Bellehumeur et al. 1997).

$$C_{0v} = \frac{C_{0p}}{n_v} \tag{5}$$

where C_{0v} is the nugget effect corresponding to Z_v, C_{0p} is the nugget effect corresponding to Z_p and n_v the number of observations available within v.

Assuming Z_p presents a random or regular spatial distribution over the field under study, changes in C_{0v} can be therefore written as a function of the spatial distribution of the variable Z_p (Eq. 6):

$$C_{0v} = \frac{C_{0p}}{r.L^2} \tag{6}$$

where C_{0v} is the nugget effect corresponding to Z_v, C_{0p} is the nugget effect corresponding to Z_p and r stands for the average spatial resolution Z_p over the study area.

Calculation of the spatially structured variance C_{1v} when increasing the grid size

Now that the nugget effect associated with quadrats of size v has been written with simple terms, there is a need to also simplify the expression of C_{1v}, i.e. that is to say that of $\bar{\gamma}(v, v)$ (Eq. 3). Here, as $\bar{\gamma}(v, v)$ is defined as the average point variogram value calculated over all possible distance vectors contained in v, and because v is a squared area of side L, it was decided to approximate $\bar{\gamma}(v, v)$ by the semi-variogram value at half the side distance of v (Eq. 7):

$$\bar{\gamma}(v, v) \approx \gamma(L/2) \tag{7}$$

It is acknowledged that this assumption strictly holds only when the relationship of the variogram with the distance is linear. However, the approximation of $\bar{\gamma}(v, v)$ that is proposed might be recognized to be viable when small grid sizes are considered, i.e. small L. In fact, at small lags, exponential and spherical variograms can be quite well fitted with a linear model. Note that this approximation will be investigated later on. For large grids, this assumption would effectively not hold as there might be strong deviations from linear models. Anyway, there would be no interest in using large grid sizes as the spatial structure would be totally lost. It is also acknowledged that the approximation in Eq. (7) would not hold for specific forms of variograms, such as gaussian variograms, because the deviation would be too strong with regard to a linear model even at small lags. However, this issue is not of primary importance given the fact that gaussian variograms are not widely present in the case of precision agriculture data.

From now on, it is considered that the spatial structure of Z_p is modelled by an exponential semi-variogram model. This model was chosen because it is commonly used for precision agriculture data. Note however that Table 1 reports the results of these calculations for the most commonly used semi-variogram models. Following Eqs. (3) and (7), Eq. (8) describes how C_{1_v} can be expressed as a function of the original spatially structured variance C_{1_p} when the variable Z_p is fitted by an exponential variogram model.

$$C_{1_v} = C_{1_p}.\exp\left(-\frac{L}{2.a}\right) \tag{8}$$

where C_{1_v} and C_{1p} are the proportions of variance corresponding to the spatially correlated component for Z_v and Z_p respectively, a is the parameter of the model (with 3*a corresponding to the practical range) and L stands for the distance between observations so that $L \times L = v$.

Equation (6) shows that C_{0_v} logically decreases with increasing SUs, which is what the analyst or the practitioner expects when mapping spatial data. Note however that increasing the SU also leads to a decrease in C_{1_v} which is not desirable when working with spatial data (Eq. 8).

Defining the optimal grid size to map spatial data

The objective of this work can be considered from a signal processing point of view. In fact, C_{0_v} may be considered as a random noise (N) that the practitioner wants to minimize (or even eliminate). In contrast, C_{1_v} may be considered as the signal (S) corresponding to the relevant proportion of variability to map that the practitioner wants to keep at the highest possible level. Given that an increase in the SU leads to a decrease in N and S at the same time, the aforementioned goal might be difficult to reach. To tackle this issue, two indices are proposed (Eqs. 9 and 10).

$$P_{NR} = \frac{\left(C_{0p} - C_{0v}\right)}{C_{0p}} \tag{9}$$

$$P_S = \frac{C_{1v}}{C_{1p}} \tag{10}$$

where P_{NR} estimates the proportion of the initial noise (C_{0p}) that is removed when a given SU (v) is considered, and P_S estimates the proportion of remaining spatially structured variance within a given v.

Changes in P_{NR} and P_s can simultaneously be taken into account through a function f_v that considers the sum of both proportions (Eq. 11).

$$f_V = P_{NR} + P_S \tag{11}$$

Note that a more complex form of f_v could be proposed, i.e. by giving weights to the different terms of the function. Such a refinement would require choosing appropriate weights for each of these terms. For simplification purposes, as a first attempt, f_v will be solely defined by the sum of both P_{NR} and P_s without considering different weights.

Considering Eqs. (6) and (8), f_v can be defined as a function of (i) L and (ii) the characteristics of Z_p, i.e. r and a (Eqs. 12 and 13):

$$f_V = \frac{\left(C_{0p} - \frac{C_{0p}}{r.L^2}\right)}{C_{0p}} + \frac{C_{1p}.\exp(-\frac{L}{2.a})}{C_{1p}} \tag{12}$$

By simplifying Eq. 12, f_v can be expressed as follows:

$$f_V = \exp\left(-\frac{L}{2.a}\right) + 1 - \frac{1}{r.L^2} \tag{13}$$

The optimal grid size length, L_{OPT}, corresponding to the SU that maximizes f_v is defined as the size L for which the derivative of f_v is null. Table 1 reports the L_{OPT} value for the most commonly-used variogram models. Some of the equations in Table 1 express L_{OPT} as a function of the Lambert function, referred to as W. This function is the inverse of the following function:

$$f(W) = We^W \tag{14}$$

where e^W is the exponential function, that is to say y = W(x) if and only if x = ye^y.

The use of the Lambert function is necessary so that a simple expression of L_{OPT} can be defined. Interested readers can refer to Weisstein (2002) for further information.

Datasets used

Two types of data were used to perform the tests:

- *Hypothetical fields* of known spatial structure were obtained from a simulated annealing procedure (Goovaerts 1997). Variogram parameters of these hypothetical fields were based on data from grape fields harvested with real-time yield monitors (Taylor et al. 2005). For these fields, the theoretical variogram is an exponential model for which the nugget effect is approximately one-third of the sill. It was decided to apply a nugget effect of 3 and a sill of 9 (arbitrary unit). As such, the different fields differ only in the ranges of their variograms. These theoretical fields were presented in a previous paper (Tisseyre and McBratney 2008). Three fields were used with practical ranges for the variogram model of 10, 20 and 45 m. From now on, these three hypothetical fields will be referred to as Data2, Data3 and Data4 (ranges of 10, 20 and 45 m, respectively). All the fields have an area of 1 ha (100 × 100 m) and the spatial resolution was set to 2 000 points ha^{-1} to match the resolution of classical monitored data (i.e. yield data).

Table 1 Variogram models and corresponding optimal grid size

Variogram model	Function to maximize	Maximum
Exponential	$fv = e^{-\left(\frac{l}{2a}\right)} + 1 - \frac{1}{rl^2}$	$L_{OPT} = -6a * W\left[\frac{-1}{3\left(2ra^2\right)^{\frac{1}{3}}}\right]$
Spherical	$fv = C_{1P}\left[1 - \frac{3l}{4a} - \frac{l^3}{16a^3}\right] + 1 - \frac{1}{rl^2}$	$L_{OPT} = 2\sqrt{3rC_{1P}\left(\frac{L}{2}\right)^5 + 3a^2rC_{1P}\left(\frac{L}{2}\right)^3 - a^3}$

W Lambert function

Data were not regularly distributed over the fields (Fig. 2). To evaluate the influence of the raw data spatial resolution on the definition of the optimal grid size, a random sampling was performed on Data4 to obtain datasets with similar spatial structures but with spatial resolutions of 1 000 and 500 points per hectare. Those will be referred to as Data4_res1000 and Data4_res500 respectively.

- *Real data* corresponding to total soluble solids (TSS) concentration in degree Brix (°Brix). TSS was measured at harvest over a vine field with a hand-held spectrometer (Spectron®, Pellenc, Pertuis, France). The 1.2 ha field was planted with cv. Syrah and harvested in 2012 in Provence (43.70°, 5.47° WGS 84). The average sampling rate was around 1200 measurements per hectare. The spatial support of the data was very small (3–5 berries) and there was considerable variation from one grape to another. Usually, wine industry practitioners averaged data over several grapes and several vines to smooth local variations and remove short-range variations.

Validation Tests

Experiments aimed to evaluate whether the values of C_{1v} computed numerically with the help of Eqs. 3 and 7, and the values of C_{0v}, computed numerically using Eq. 6 were similar to those obtained after fitting a model to the regularized variogram. From now on, the values obtained with the aforementioned equations are referred to as the 'estimated' C_{1v} and C_{0v} values, while those obtained with the fit of the regularized variogram are referred to as the 'observed' C_{1v} and C_{0v} values.

Validations were performed by considering a division of the fields into a variable number of sampling units v. Figure 2 shows three examples of the same theoretical field divided into sampling units of size $v = 36$, 100 and 400 m^2, respectively for Fig. 2a, b and c. Points represent the locations of measurements (Z_p). Experiments were conducted with 9 different grid sizes (cells of size $v = 4,16,36,64,100,144,196,324,400$ m^2). For each size v, the following operations were performed:

(i) Within each cell, calculation of the variable Z_v, where Z_v is the mean of Z_p values observed

(ii) Numerical computation of the estimated C_{1v} and C_{0v} values using Eqs. 3, 6 and 7

(iii) Calculation of a regularized variogram and determination of the observed C_{1v} and C_{0v} values

(iv) Comparison of the observed and estimated C_{1v} and C_{0v} values using classical statistics such as the coefficient of determination (R^2).

Data analysis and mapping

Once the optimal grid length L_{OPT} was determined, two methods were considered to aggregate the high resolution spatial data. The first, more accurate, method was to fit a variogram model to the data and then perform block kriging over a grid of size L_{OPT}. Because this first approach requires some skills in variogram modelling, a second method was proposed. This latter approach simply aims at averaging the observations inside each cell of the grid of size L_{OPT}. It is acknowledged that this second methodology is not optimal because the average estimate and the overall variance within the block might be biased. However, the averaging approach is easier to implement with common GIS software.

Mapping of the data was done with QGIS (Quantum GIS Development, V1.8, Open Source Geospatial Foundation Project) by importing Eastings and Northings and values for each of the fields. Specific calculations like z_v determination over the different sampling units were performed using our own functions developed under the R statistical environment (R Core Team, 2017) along with dedicated packages of spatial analyses such as *sp, gstat. maptools* and *rgeos*. The TSS data were mapped in 33% quantiles resulting in three classes for each map: low (0–33% quantile), medium (34–67% quantile) and high (68–100% quantile).

Results and discussion

Validation of the methodology on hypothetical fields

Strong correlations ($R^2 > 0.99$) between the observed and estimated C_{0v} and C_{1v} parameters are shown in Fig. 3. Strong correlations ($R^2 = 0.97$) were also found for the range parameter, i.e. a_v (results not shown). These findings demonstrate the relevance of the theoretical approach and its ability to estimate the parameters C_{1v}, C_{0v}, and a_v of a variable Z_v, derived from Z_p for the different sizes of sampling units v considered in this study. The estimated parameters are used in the rest of the paper.

Evolution of C0 and C1 as a function of the size of the sampling unit (v)

Increasing the size of v resulted in a decrease in the nugget effect (C_{0v}) (Fig. 4, left). This result is explained by a higher number n of measurements available on v (Eq. 5). Regarding C_1 (Fig. 4, right), it also decreased with an increasing size of v. The correlated component,

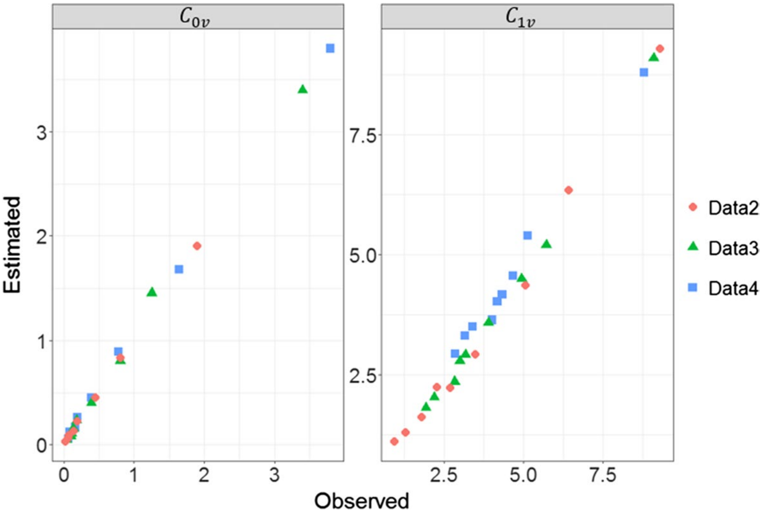

Fig. 3 Estimated VS observed C_{0v} (left) and C_{1v} (right) for the three hypothetical fields and 9 different sizes of sampling units (v)

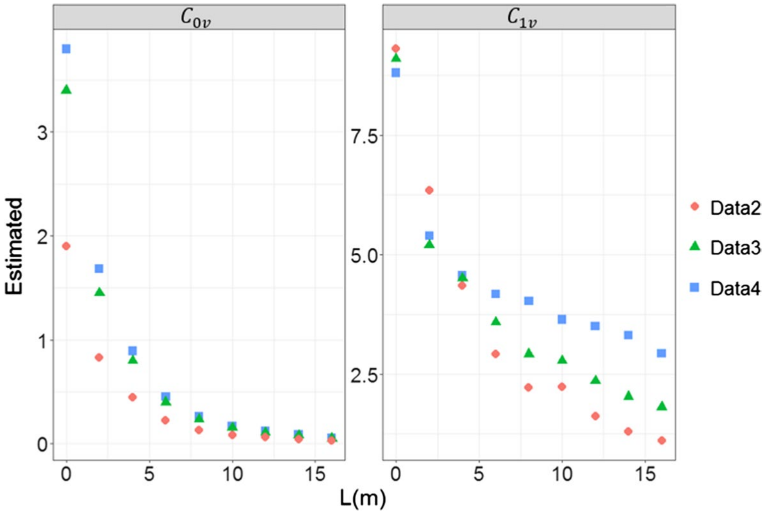

Fig. 4 Change in C_{0v} (left) and in C_{1v} (right) as a function of the size of the sampling unit v ($v = L\times L$) for the three hypothetical fields

C_{1v}, decreased faster when the range of the variogram of Z_p was smaller. Equation (3) explains this result since the variance corresponding to a spatial structure (not random) can only decrease with an increasing size of v. This decrease was even faster when $\bar{\gamma}(v, v)$ (the mean variogram value of Z_p within v) was large. The correlated component, C_{1v}, tended to zero when the size of v became larger than the range of the Z_p variogram.

Finding the optimal grid size to aggregate spatial information

The function $f_v = P_{NR} + P_s$ exhibited a clear maximum value for all the datasets under study (Fig. 5, left). Not surprisingly, f_v values were higher for Data4 than for Data2. Indeed, the most spatially correlated datasets come along with stronger signal information which means that the amount of remaining signal with respect to C_{1p}, i.e. the proportion of remaining signal, will be higher. The optimal grid length, L_{OPT}, is the size for which the function '$P_S + P_{NR}$' reaches a maximum value. It appears that the optimal grid length also increased from Data2 to Data4. As the spatial structure increased from Data2 to Data4, it should be possible to aggregate data over a larger grid without losing an important part of the signal. By using the function '$P_S + P_{NR}$', the same weight has been given to the proportions of remaining signal and noise removed. This choice should be further investigated because there is no reason for these weights to be similar. Indeed, it would not be interesting to attach the same importance to smoothing out the nugget variation in the situations where (i) the nugget is a tiny fraction of the partial sill and (ii) the partial sill is a tiny fraction of the nugget.

Decreasing spatial resolution resulted in a decrease of the f_v values and an increase in the optimal grid size (Fig. 5, left). It can be seen that the f_v function cannot be defined for small grid sizes when the spatial resolution is low. It must be clear that the observed differences are only due to a change in the raw data spatial resolution as the spatial structures of these data are similar. Given Eqs. (6) and (8), the spatial resolution only affects the

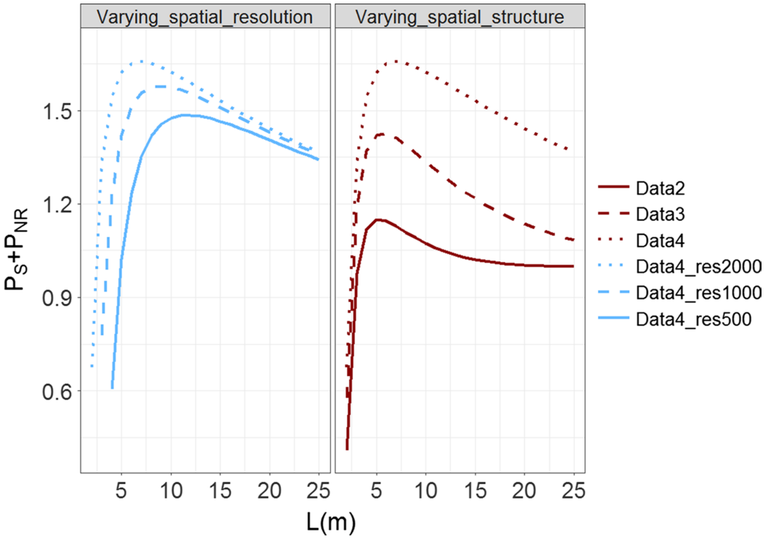

Fig. 5 Change in 'P_S + P_{NR}' as a function of v ($v = L \times L$) for varying spatial resolution [left] and spatial structure [right]

Table 2 Summary of signal and noise information after data aggregation over a grid of length L_{OPT}

Dataset	Resolution (pts ha^{-1})	L_{OPT} (m)	$P_S + P_{NR}$	*S/N*
Data2	2 000	5	1.15	1.69
Data3	2 000	6	1.43	1.71
Data 4/Data4_res2000	2 000	7	1.66	2.69
Data4_res1000	1 000	9	1.58	2.06
Data4_res500	500	12	1.48	1.62

calculation of the C_{0v} parameter, i.e. the nugget effect associated with Z_v. For a given grid size, as the raw data spatial resolution decreased, C_{0v} increased which makes the proportion of noise removed, P_{NR}, lower. As a consequence, the f_v values also diminished.

So far, it has been shown that the function 'P_S + P_{NR}' was derivable for most classical functions and that a clear maximum value was identifiable for a specific grid length, L_{OPT} (Table 2). Given this optimal length, it was then possible to quantify the amount of remaining spatially structured component, i.e. the signal, with respect to the remaining noise inside the field. However, as the 'P_S + P $_{NR}$' function is defined with proportions, it does not account for the initial absolute amount of noise in the dataset. Indeed, there would be no interest in aggregating data if the noise is initially very low. To overcome this issue, an interesting procedure could be to make use of the Cambardella index, i.e. $\frac{C_{0v}}{C_{0v}+C_{1v}}$ for which spatial structures rules have been defined (Cambardella et al. 1994). These authors have stated that datasets exhibiting a Cambardella index lower to 25% could be characterized as very well spatially structured datasets. By computing the inverse of this index, datasets would be said of high spatial structure with a value over 4, i.e. $\frac{100\%}{25\%}$. Here, as we consider the signal to be solely the spatially structured variance, i.e. C_{1v}, contrary to $C_{0v} + C_{1v}$ for Cambardella et al. (1994), data aggregation might be stopped when the *S/N* ratio exceeds a value of 3, $\frac{75\%}{25\%}$. Be aware that, for some datasets, the *S/N* ratio might never reach this defined threshold, especially if the initial amount of noise in the dataset is too strong (Table 2). In these cases, choosing L_{OPT} may remain the best option to define the grid length. Users should also note that the optimal grid length, L_{OPT}, should be much lower than the range of the point variogram because the spatial structure would be completely lost otherwise.

It must be clear that there needs to be a combination of the two functions '$P_S + P_{NR}$' and '*S/N*' because they are both important in the characterization of the signal and noise in the remaining dataset. The '*S/N*' ratio cannot be used alone because this index can be extremely high if the signal and noise are both very low. Similarly, if one uses solely the '$P_S + P_{NR}$' function, there is no indication on whether there remains a strong signal with respect to the noise. Please note that originally, the authors' first objective function (Eq. 11) was not the sum '$P_S + P_{NR}$ but rather the ratio of the random noise N to the signal S in the data because it was a more intuitive objective function. However, using this objective function led to identification of grid sizes that were larger than the range in these data. These findings were mainly due to the form of the functions that needed to be derived. As such, a more appropriate objective function was used, i.e. '$P_S + P_{NR}$' was considered.

Study case on real data

The TSS measurements observed over the vine field are noisy (Fig. 6a). The variogram model of the data exhibits a large nugget effect ($C_0 = 3.2$) compared to the sill ($\text{sill} = C_0 + C_1 = 3.7$). This characteristic is related to the method of measurement on a punctual support (Z_p) chosen at random. This results in a large nugget variance that arises from: (i) the measurement approach (sensor + operator), (ii) the within-cluster variability

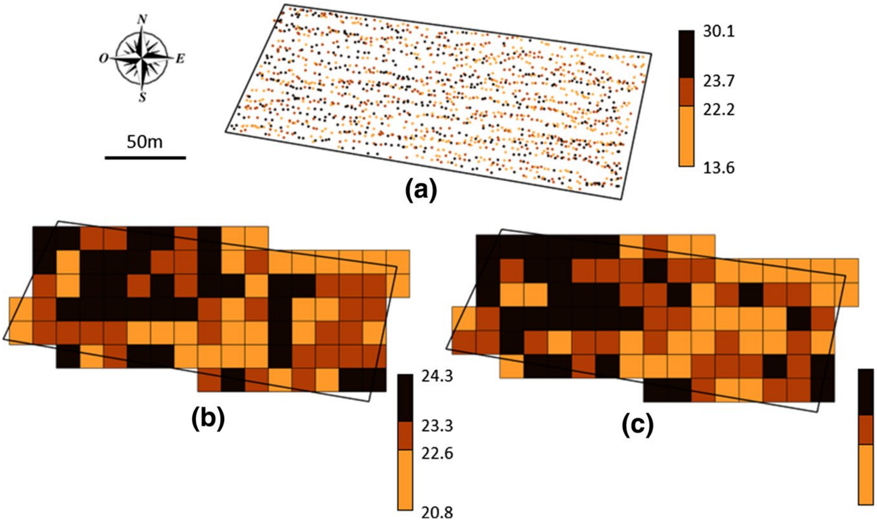

Fig. 6 Total soluble solid maps (°Brix). **a** Raw data; **b**. Raw data interpolated by block kriging over the optimal grid size: 10 m; **c** Raw data averaged over the optimal grid size: 10 m. Three classes of TSS were considered for each map: low (0–33% quantile), medium (34–67% quantile) and high (68–100% quantile)

of the berries, (iii) the within-plant clusters variability, and (iv) the variability from one plant to another. This problem is usually solved at sampling by averaging TSS values over several plants and assigning this value to a central measurement site. In this case, Z_v is defined with the sampling process but the size of v is defined arbitrarily by the practitioner without any prior studies.

Given the high spatial resolution of the data (~ 1200 points.ha^{-1}), an analysis of the evolution of 'P_S + P_{NR}' was done. It was possible to identify an optimal grid length, L_{OPT} = 10 m, that maximized the previously defined function f_v. For $v > 100$ m^2, a decrease in 'P_S + P_{NR}' was observed indicating that a loss in signal is expected. This optimal grid size can be used to aggregate the TSS measurements, either by block kriging (Fig. 6b) or by simple averaging of values inside each cell of the grid (Fig. 6c). It must be clear that if the geostatistical model that was fitted to the data is valid, then the predictions originating from the block kriging approach will be more accurate than those obtained from the averaging procedure and the prediction error variances will also be lower. However, averaging the data inside each cell of the grid does not require modelling skills and can be automated quite easily. Here, the average block kriging prediction error variance on the optimal grid size is relatively low over the field, i.e. 0.36. Note also that the averaging procedure significantly impacted the within-field spatial structure in the data and especially the small-scale variations ($C_0 = 0.4$; sill $= C_0 + C_1 = 0.55$). Be aware that the signal-to-noise ratio of the aggregated data with the optimal SU is twice as high as that of the raw data. It can be seen that both aggregation methods exhibit a similar spatial structure with the presence of a zone with larger TSS in the north-western part of the field and a zone with smaller TSS in the north-eastern part of the field. Note that this spatial structure was hardly visible by solely plotting the initial observations (Fig. 6a). Be aware that here, the block kriging approach seems to have smoothed the map to a greater extent than the averaging procedure. This can be explained by the relatively low spatial structure exhibited by the TSS observations.

It must be noted that these maps are only informative. From a practical point of view, the patterns identified need additional observations to understand their origin or to consider site-specific management. Decision-maker must remain cautious about interpreting these zones since the prediction inside each support is also associated to a prediction error variance. Any decision must take into account this within v variability. Note, however, that in many studies, this variability is seldom taken into account (if ever). In general, and more especially for perennial crops, initial observations already originate from some aggregation procedures, i.e. the initial support Z_p is in fact already a support Z_v. This procedure is very common and even if it is almost never discussed, it is aimed at reducing the variance at small distances. One could argue that, in this case, the proposed method could help identify the optimal measurement support and help practitioners define the best sampling so that the measurement noise is minimized. One strong advantage of this procedure would be to preserve the punctual observations as long as possible in the analysis and mapping process. In this work, the size of the grid has not been put in relation with the structure of the crop being grown. For instance, wheat is grown continuously over the field while vines are much more row-structured crops. In the case of vineyards, there would be no interest in averaging data in small cells that fall into these inter-rows. This might be of concern for the definition of optimal grid sizes.

Be aware that the proposed methodology, as emphasized by Bellehumeur et al. (1997), only applies to additive variables. In the present case, this condition remains questionable for TSS since its estimation over a grid cell will necessarily depend on the yield of the vine. In this work, TSS was considered additive locally, since previous studies have shown a strong autocorrelation of yield in viticulture (Taylor et al. 2005). As a result, in this study, the yield was considered relatively homogeneous within the same grid cell because those cells were small enough. Nonetheless, practitioners must remain cautious with this condition of additivity when applying the proposed methodology.

Another advantage of the methodology would be to help define general signal filtering methods for a large set of fields. For instance, the idea could be first to acquire reference information in several vineyards, e.g. using Physiocap® or Greenseeker® sensors (Debuisson et al. 2010; Baluja et al. 2012). Then, the proposed methodology could be run to finally propose a specific aggregation of the information that could be adapted to the majority of the fields under study. This approach would also enable to lower the amount of exchanged data and so the processing of such data. This might be very interesting considering that, in general, cloud services are used to process the data.

Conclusion

This study has focused on the question of the optimal grid size for data aggregation procedures in precision agriculture. In this work, it was shown to be possible to find an optimal grid size for which the proportions of remaining signal and of noise removed are maximized. By using a signal to noise ratio when the optimal grid size is defined, the practitioner is able to quantify the amount of remaining spatially structured component with respect to the noise in its datasets. The use of these indicators may provide the analyst with a decision support to choose the best grid size to map high resolution information. Further investigations will focus on tests and validations on a larger data base to assess their relevance.

References

Acevedo-Opazo, C., Tisseyre, B., Guillaume, S., & Ojeda, H. (2008). The potential of high spatial resolution information to define within-vineyard zones related to vine water status. *Precision Agriculture, 9*(5), 285–302.

Baluja, J., Diago, M., Goovaerts, P., & Tardaguila, J. (2012). Assessment of the spatial variability of anthocyanins in grapes using a fluorescence sensor: relationships with vine vigour and yield. *Precision Agriculture, 13,* 457–472.

Bellehumeur, C., Legendre, P., & Marcotte, D. (1997). Variance and spatial scales in a tropical rain forest: Changing the size of sampling units. *Plant Ecology, 130,* 89–98.

Bivand, R. S., Pebesma, E. J., & Gomez-Rubio, V. (2008). *Applied spatial data analysis with R*. New York, USA: Springer.

Blackmore, S., Godwin, R. J., & Fountas, S. (2003). The analysis of spatial and temporal trends in yield map data over six years. *Biosystems Engineering, 84*(4), 455–466.

Bramley, R. G. V., Trought, M. C. T., & Praat, J. P. (2011). Vineyard variability in Marlborough, New Zealand: Characterizing variation. *Australian Journal of Grape and Wine Research, 17,* 72–78.

Cambardella, C. A., Moorman, T. B., Novak, J. M., Parkin, T. B., Karlen, D. L., Turco, R. F., et al. (1994). Field-scale variability of soil properties in central Iowa soils. *Soil Science Society of America Journal, 58,* 1501–1511.

Cressie, N. (1996). Change of support and the modifiable areal unit proble. *Geographical Systems, 3,* 159–180.

Debuisson, S., Germain, C., Garcia, O., Panigai, L., Moncomble, D., Le Moigne, et al. (2010). Using Multiplex® and Greenseeker™ to manage spatial variation of vine vigor in Champagne. In *10th International Conference on Precision Agriculture*. https://www.ispag.org/proceedings/?action=year_abstracts. Accessed 9 March, 2018.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation. Applied geostatistics*. New York, USA: Oxford University Press.

Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. London, England: Academic Press.

Lauzon, J. D., Fallow, D. J., O'Halloran, I. P., Gregory, S. D. L., & von Bertoldi, A. P. (2005). Assessing the temporal stability of spatial patterns in crop yields using combine yield monitor data. Canadian Journal of Soil Science, *85*(3), 439–451.

Li, Y., Shi, Z., Wu, C.-F., Li, H.-Y., & Li, F. (2008). Determination of potential management zones from soil electrical conductivity, yield and crop data. *Journal of Zheijang University, 9,* 68–76.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Serrano, J. M., Peça, J. O., Marques da Silva, J. R., & Shahidian, S. (2010). Mapping soil and pasture variability with an electromagnetic induction sensor. *Computers and Electronic in Agriculture, 73*(1), 7–16.

Tagarakis, A., Liakos, V., Fountas, S., Koundouras, S., & Gemtos, T. A. (2013). Management zones delineation using fuzzy clustering techniques in grapevines. *Precision Agriculture, 14,* 18–39.

Taylor, J., Acevedo-Opazo, C., Ojeda, H., & Tisseyre, B. (2010). Identification and significance of sources of spatial variation in grapevine water status. *Australian Journal of Vine and Wine Research, 16,* 218–226.

Taylor, J., Tisseyre, B., Bramley, R., & Reid, A. (2005). A comparison of the spatial variability of vineyard yield in European and Australian production systems. In: Stafford, J. V. (Ed.), Proceedings of the 4th European Conference on Precision Agriculture. The Netherlands: Wageningen Academic Publishers, pp 907–914.

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

Weisstein, E. W. (2002) "Lambert W-function," MathWorld, A Wolfram Web Resource., http://mathworld.wolfram.com/LambertW-Function.html. Accessed 10 August, 2017