Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

A multifunctional matching algorithm for sample design in agricultural plots

N. Ohana-Levi [a,*,1], A. Derumigny [b,1], A. Peeters [c], A. Ben-Gal [d], I. Bahat [e,f], L. Katz [d,e,f,g], Y. Netzer [i,j], A. Naor [h], Y. Cohen [e]

[a] Independent Researcher, Variability, Ashalim 85512, Israel
[b] Department of Applied Mathematics, Delft University of Technology, Mourik Broekmanweg 6, 2628 XE Delft, the Netherlands
[c] TerraVision Lab, Midreshet Ben-Gurion 8499000, Israel
[d] Institute of Soil, Water and Environmental Sciences, Agricultural Research Organization, Gilat Research Center, Mobile post Negev 2, 85280, Israel
[e] Institute of Agricultural Engineering, Agricultural Research Organization, Volcani Center, P.O. Box 15159, Rishon LeZion 7505101, Israel
[f] The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, The Hebrew University of Jerusalem, The Robert H. Smith Faculty of Agriculture, Food & Environment, Rehovot 76100, Israel
[g] Department of Soil and Water Sciences, The Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, P.O. Box 12, Rehovot 7610001, Israel
[h] Department of Precision Agriculture, MIGAL Galilee Research Institute, Kiryat Shmona 11016, Israel
[i] Department of Agriculture and Oenology, Eastern R&D Center, Israel
[j] Department of Chemical Engineering, Ariel University, Ariel 40700, Israel

ARTICLE INFO

Keywords:
Partially-observed data
Representative sampling given covariates
Two-phase study
Agricultural sampling
Spatial autocorrelation

ABSTRACT

Collection of accurate and representative data from agricultural fields is required for efficient crop management. Since growers have limited available resources, there is a need for advanced methods to select representative points within a field in order to best satisfy sampling or sensing objectives. The main purpose of this work was to develop a data-driven method for selecting locations across an agricultural field given observations of some covariates at every point in the field. These chosen locations should be representative of the distribution of the covariates in the entire population and represent the spatial variability in the field. They can then be used to sample an unknown target feature whose sampling is expensive and cannot be realistically done at the population scale.

An algorithm for determining these optimal sampling locations, namely the multifunctional matching (MFM) criterion, was based on matching of moments (functionals) between sample and population. The selected functionals in this study were standard deviation, mean, and Kendall's tau. An additional algorithm defined the minimal number of observations that could represent the population according to a desired level of accuracy. The MFM was applied to datasets from two agricultural plots: a vineyard and a peach orchard. The data from the plots included measured values of slope, topographic wetness index, normalized difference vegetation index, and apparent soil electrical conductivity. The MFM algorithm selected the number of sampling points according to a representation accuracy of 90% and determined the optimal location of these points. The algorithm was validated against values of vine or tree water status measured as crop water stress index (CWSI). Algorithm performance was then compared to two other sampling methods: the conditioned Latin hypercube sampling (cLHS) model and a uniform random sample with spatial constraints. Comparison among sampling methods was based on measures of similarity between the target variable population distribution and the distribution of the selected sample.

MFM represented CWSI distribution better than the cLHS and the uniform random sampling, and the selected locations showed smaller deviations from the mean and standard deviation of the entire population. The MFM functioned better in the vineyard, where spatial variability was larger than in the orchard. In both plots, the spatial pattern of the selected samples captured the spatial variability of CWSI. MFM can be adjusted and applied

[*] Corresponding authors.
E-mail address: noao@post.bgu.ac.il (N. Ohana-Levi).
[1] Equal contribution.

https://doi.org/10.1016/j.compag.2021.106262
Received 27 December 2020; Received in revised form 7 June 2021; Accepted 8 June 2021
Available online 15 July 2021

using other moments/functionals and may be adopted by other disciplines, particularly in cases where small sample sizes are desired.

1. Introduction

Accurate information is necessary for decision-making and management in precision agriculture. Proliferation and improvement in technological applications and data availability enable extension of precision agriculture and “smart farming” (Kamilaris et al., 2017). Gaining more product with less input in agricultural plots can be achieved using multiple applications, including field management according to the spatial variability of certain characteristics, namely management zones (MZs), for irrigation, fertilization and pest control (Monaghan et al., 2013; Nawar et al., 2017; Park et al., 2007). Tools for spatial management include sampling soil and specific plants, directly or indirectly, to collect relevant data for decision-making. Such data can include soil properties, crop size and growth, crop physiological status, and yield characteristics (De Lannoy et al., 2006; Gandah et al., 2000; Jonckheere et al., 2004; Menzel & Simpson, 1986) collected manually or via remote or proximal sensing. In-field sensors, continuously collecting environmental, soil or crop-related data add a temporal dimension to potential decision-making processes.

Although sampling technologies have become more affordable and, in some cases, automated, managers remain limited regarding the number of sampling points that may be feasibly collected within a given field. Therefore, there is a need to develop methods to select points or individual plants to sample that represent the entire field/population. Sampling techniques within agricultural fields vary according to the type of representation required. Spatial sampling is commonly needed to address strategies for precision agriculture (Oliver, 2010). Spatial representations are designed to account for within-field heterogeneity and spatial autocorrelation (Haining, 2015b) and represent the spatial distribution. In many cases, the samples are the basis for spatial interpolation of the collected variables, which account for differences or similarities between measurements (Stein & Ettema, 2003). Sampling plans for interpolation purposes are frequently performed using random, stratified random, systematic, and adaptive spatial sampling throughout the agricultural field (Haining, 2015b; Stein & Ettema, 2003). In these cases, the size or density of the data collected is of major importance and is a function of the degree of spatial autocorrelation (Kerry et al., 2010). An additional related topic of interest has, in recent years, focused on the deployment of sensors across the field (Ben-Gal et al., *In Press*). For distribution of wireless sensors, a main concern for their location is achieving maximum coverage of the field without obstruction (Aqeel-Ur-Rehman et al., 2014), but for all sensors, as for sampling in general, the distance between the sensors is often based on field attributes (Visalini et al., 2019). Commonly in agricultural practices, MZs are used as a general frame for sampling since it is assumed that each MZ is homogeneous in space (Gavioli et al., 2019). Therefore, following a delineation into MZs, sample-points are selected arbitrarily within each zone (Johnson et al., 2003) to represent field properties. This approach may be problematic, since the spatial distributions of MZs are subjected to the selected clustering algorithm used to generate them (Chang et al., 2003). Additionally, clustering algorithms rarely generate perfect separations of groups, and tend to introduce dissimilarities among clusters (i.e., misclassified individuals) (Raskutti & Leckie, 1999), thus introducing bias to the sample. Few studies have dealt with sampling with the goal of representing the distribution of an unknown target variable across the field while using a set of known variables, such as the method proposed by Bazzi et al. (2019). While sampling for interpolation purposes requires estimated values for as many points in the field as possible, sampling for representation of the population distribution commonly aims to reduce the number of locations/observations and to represent the field using the lowest possible number of points. This study proposes an approach that aims to optimize data collection efficiency by collecting a small number of observations representing the statistical distribution of target field properties while avoiding information redundancy due to spatial autocorrelation and accounting for data values that are not independent (Haining, 2015a). This means that the main purpose is to sample the population distribution and not the spatial distribution, although the two frequently coincide (López-Vicente et al., 2020; Roudier et al., 2008).

In the field of digital soil mapping, conditioned Latin hypercube sampling (cLHS) was introduced by Minasny & McBratney (2006) as a sampling strategy of an area where ancillary data are available. The cLHS algorithm is a stratified random procedure that provides sample design based on multiple variables and their distribution, assuming that these variables represent the variability of the target variable to be sampled. The technique optimizes the N sites to be included in the sample, such that the multivariate distribution of the ancillary data is maximally stratified (Minasny & McBratney, 2006), providing full coverage of each explanatory variable. The cLHS method has been widely used in the field of digital soil mapping to assess variability in soil properties at regional scales, using ancillary data from a wide range of sources such as governmental data and remote sensing retrievals (Mulder et al., 2013). The method is also commonly used in modeling for selection of calibration and validation sets of the data (Richter et al., 2016). Over time, several extensions to the cLHS were introduced. The variance quadtree algorithm was designed to account for non-stationarity in the spatial sampling (Minasny et al., 2007). The flexible Latin hypercube sampling was suggested for selecting alternative sites in cases where sampling would only be possible in specific regions. For example, in proximity to tracks and roads or avoiding privately owned land (Clifford et al., 2014). cLHS has rarely been used for agricultural purposes (as shown by Adamchuk et al., 2011 & Israeli et al., 2019), and was shown to be less favorable in cases where a low number of locations is required for sampling (Werbylo & Niemann, 2014). However, the notion of using ancillary data to optimize the sampling design was shown to be highly effective (Zhang et al., 2017).

Using known field properties can assist in determining the pattern and optimal sampling intervals to avoid over- or under-sampling (Kerry et al., 2010). The representation of a target variable using a minimum number of samples relies on a set of predefined attributes with maximum coverage in the field. These previously measured variables are assumed to be related to the target variable, thus building confidence that their characteristics and statistical moments (i.e., statistical parameters that describe the location, scale or shape of a distribution) can serve as proxies for the unknown target variable. The use of multiple variables that are efficiently and easily acquired in the field in terms of labor, time, and cost, should therefore be beneficial.

This work proposes an algorithm to quantify a multifunctional matching (MFM) criterion to statistically define the “best” set of locations for sampling within a field. The algorithm was illustrated based on two datasets, one from a peach orchard and one from a wine grape vineyard. The resulting selection of observations to sample (e.g. trees/vines) were chosen to be representative of the joint distribution of several known variables, which are supposed to be related in some way to the distribution of the unknown target variable. Therefore, observations of the target variables at those points can be representative of the entire (unknown) population distribution of the target variable. The algorithm considers the spatial variability in the field as a cause of potential information redundancy and is designed to avoid it by applying a spatial constraint. However, as MFM is not a spatial algorithm, it was not designed to calculate and account for spatial autocorrelation or spatial stratified heterogeneity (SSH) (Wang et al., 2016) inherently. A second algorithm was further suggested to determine the most suitable number of observations within the field according to a predefined level of accuracy reduction from the maximum number of observations necessary, given the spatial constraint. The benefits of the MFM algorithm may support agronomic applications for both commercial and scientific purposes, such as berry sampling prior to harvest in vineyards (i.e, °Brix, pH, total acidity) (Kasimatis & Vilas, 1985; Sinton et al., 1978) or fruit mass/size and red skin coloration in orchards (Gasic et al., 2015; Shinya et al., 2013).

The main algorithm is based on the comparison between multiple moments of the population and of the sample. It finds the best sample whose moments are simultaneously most similar to those observed in the whole population. Comparison of moments has long been common in statistics, dating from Pearson (1894) who first proposed to estimate a parameter by matching empirical to theoretical moments. Since Pearson, the so-called "method of moments" has become popular due to its generality and wide applications, especially in econometrics (Hall, 2005; Hansen, 1982), but also in other fields where complex models are used, such as electromagnetism (Gibson, 2014) and hydrology (Kitanidis, 1988) among many others. The main idea behind the method of moments is very intuitive: given a parametric statistical model and an observed sample, the unknown parameter is estimated as the value implicitly defined by theoretical moments matched to empirical ones. In this study, the framework is slightly atypical, even if the same kind of intuition can be used. We do not assume any parametric model for the dataset but wish to find a subset of the dataset whose distribution is close to the distribution of the whole population. The definition of closeness will be based on the supremum of the distance between several functionals of the population distribution and of the sample distribution. Various commonly used distances between distributions, such as the Kolmogorov distance, the total variation distance and the Wasserstein distance, can be defined as the suprema of functionals (Gibbs & Su, 2002).

To the best of our knowledge, there has not been any previous scientific focus on developing a method for using moment (or functional) matching to represent a (multivariate) population of field attributes using a limited number of observations for agricultural practices. However, much research has dealt with quantifying associations, both in space and time, among different environmental, soil, plant, and terrain factors (Goovaerts & Kerry, 2010; Heuvelink & van Egmond, 2010) that enable efficient determination of the appropriate set of variables used to define the sample design needed to represent an unknown target variable. Fortunately, some of these factors may be acquired at the field scale with low effort and costs.

The main objective of this study was to develop a method for optimizing locations to sample an unknown target feature across an agricultural field by accurately representing a multivariate distribution of a population of observations. Specific secondary objectives were: (1) to develop an algorithm that selects an approximated best set of observations; (2) to define the number of observations according to a specified accuracy level of population representation; and (3) to validate the algorithm performance using actual agricultural datasets and compare the proposed algorithm against uniform random sampling and cLHS.

2. Methodology

2.1. Study sites and experimental design

The MFM was tested on two datasets, based on measurements conducted in two agricultural experimental sites, a wine grapes vineyard and a peach orchard (Fig. 1), both subjected to experiments investigating the potential for variable rate irrigation management. The vineyard (Fig. 1b) is located in the central mountains in Israel and covers an area of 2.5 ha, 2.3 of which was used for the study. The area is characterized by Mediterranean climate with hot, dry summers and rainy winters and is located at an elevation ranging between 675 and 900 m above sea level. The commercial vineyard of *Vitis vinifera* L. 'Cabernet Sauvignon' was planted in 2011 with vine spacing of 1.5 m within rows and 3 m between rows, resulting in a density of 2222 vines per hectare. The number of vines within the study site was 3893. Border vines were removed from the analysis since sampling grapevines in border-rows may introduce a bias to the data due to margin effects (Murolo et al., 2014). After this step, the population consisted of 3523 vines. The vineyard is surrounded by natural and planted vegetation (figs and indigenes vine), which include low shrubs, a pine forest and Mediterranean trees (Ohana-Levi et al., 2019).

The peach orchard (Fig. 1c) is located in Israel's Upper Galilee region and covers an area of 4 ha. It is also characterized by Mediterranean

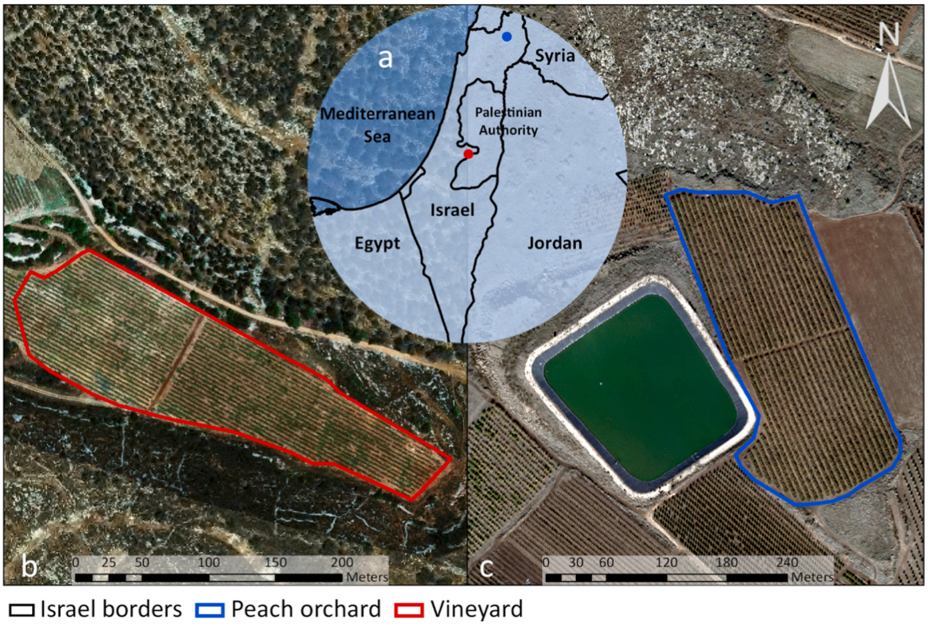

Fig. 1. The two study site locations within Israel (a): the vineyard near Mevo Beitar (red) (b) and the peach orchard near Mishmar HaYarden (blue) (c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

climate and the elevation of this plot ranges between 168 and 187 m above sea level. A commercial peach orchard of *Prunus persica cv. 1881* was planted in 2007 with spacing of 2.6 m between trees and 5 m between rows, resulting in a density of 770 trees/ha. The number of trees within the study site was 2402. Border trees were removed from the analysis, resulting in a population of 2151 trees. The orchard is located east of a water reservoir and adjacent to other orchards.

2.2. Data collection

Data used for testing the algorithm were acquired during the growing season of 2017 for both plots. The following four variables were used to test the algorithm, based on the premise that these attributes may be easily acquired and have a strong effect on, or may serve as indicators to, the plant water conditions within the field:

Slope: Extracted from a digital elevation model (DEM) based on an elevation survey conducted across the vineyard with a spatial resolution of 1 m. In the peach orchard, the elevation was based on a DEM with a spatial resolution of 4 m, provided by Survey of Israel. The slope indicates the maximum rate of change in elevation between one cell and its eight immediate neighbors. The specific cell is given a value in percent, signifying the steepest downhill descent, meaning that lower values correspond to flatter terrain. Slope was calculated using the Slope tool in the Surface toolset, ArcGIS PRO (ESRI Inc., 2017). Slope has been established as one of the main factors affecting soil moisture and consequently plant physiology (Ohana-Levi et al., 2020) and plant drought stress (Becker et al., 1988).

Topographic wetness index (TWI): This index simulates static soil moisture conditions and quantifies the distribution of accumulated area across downslope pixels. TWI was calculated using Eq. (1) and applied to the DEM data with the “dynatopmodel” package in R (Metcalfe et al., 2018):

$$TWI = \ln(A/tan\beta) \tag{1}$$

where A is the upstream contributing area (m^2) for each pixel and β is the local slope in the steepest downhill direction in degrees. TWI acquires the spatial resolution of the original DEM, in this case 1 m in the vineyard and 4 m in the peach orchard. TWI is used to assess soil moisture and its effect on plant conditions, mostly in ecological applications (Petroselli et al., 2013) and is known to influence vine vigor and water status (Bahat et al., 2021).

Apparent soil electrical conductivity (ECa): A survey was conducted in each of the plots to acquire soil ECa measurements. In the vineyard, the survey was performed on April 10, 2018 by a four-probe soil resistance sensor (Soil EC 3100) (Veris Technologies) at a field capacity, representing 0–30 cm depth. In the peach orchard, the survey was conducted on September 25, 2017 using an EM38-MK2 (Geonics Ltd., Ontario, CA), representing 0–150 cm depth. Continuous ECa maps were created using 4037 and 35,640 points recorded for the vineyard and peach orchard respectively, and interpolated using kernel interpolation, by way of a moving window calculating the shortest distance between points (Mühlenstädt & Kuhnt, 2011).

Normalized difference vegetation index (NDVI): Calculated from multispectral imagery that was acquired by an unmanned aerial vehicle (UAV) equipped with a multispectral MicaSense RedEdge camera (MicaSense* Inc, Seattle, USA). NDVI uses the red and near infrared bands of the image according to Eq. (2):

$$NDVI = \frac{\rho(\mathrm{NIR}) - \rho(\mathrm{red})}{\rho(\mathrm{NIR}) + \rho(\mathrm{red})} \tag{2}$$

where ρ is the reflectance value for each pixel, ranging from −1 to 1. The images were acquired on September 5 and August 29, 2017, for the vineyard and orchard, respectively. The value given to each vine was the average NDVI value for a rectangle with a length of 1.4 m and width of 80 cm covering the central area of the canopy. For the peach trees, the NDVI values were averaged for a 1 m radius circle buffer around the tree center. NDVI is a well-established measure of relative green vegetation and crop status. NDVI has been used to estimate a large number of vegetation properties, including leaf area (Johnson, 2003), photosynthetic response, and water stress (Aguilar et al., 2012; Kim & Glenn, 2015).

Validation of the MFM algorithm for each plot was performed using measurements of crop water stress index (CWSI), to assess how well the selected observations represent the entire population distribution of a target variable.

Crop water stress index: CWSI was computed using thermal UAV imagery acquired with a FLIR SC2000 thermal camera (FLIR* Systems Inc., Billerica, MA, USA), and is defined in Eq (3):

$$\mathrm{CWSI} = \frac{\mathrm{Tcanopy} - \mathrm{Twet}}{\mathrm{Tdry} - \mathrm{Twet}} \tag{3}$$

where Tcanopy is the temperature of a specific pixel, Twet is the lower baseline calculated from the entire field and Tdry is the upper baseline. The calculation was conducted on pixels consisting of pure vegetation, after masking out the soil surface and mixed pixels. Local air temperature was measured using meteorological stations located in the agricultural plots. Tdry was defined as air temperature +7 °C in the vineyard and air temperature +5 °C in the peach orchard (Rud et al., 2013). Twet was calculated by averaging the lowest 5% of temperature values in the thermal image (Cohen et al., 2017; Rud et al., 2014). The value provided for each observation in the vineyard was calculated as the minimum CWSI value at a radius of 70 cm around each vine in the vineyard (Ohana-Levi et al., 2019), and the mean value of the lowest 33% of CWSI pixels present within a radius of 150 cm around each tree in the peach orchard (Katz et al., In Review; Meron et al., 2010). CWSI is widely used to quantify water status in plants, specifically in agricultural crops (Khanal et al., 2017; Meron et al., 2010), including orchards and vineyards (Bellvert et al., 2016). It is known to be associated with terrain, soil and vegetation factors (Colaizzi et al., 2003; O’Shaughnessy et al., 2011; Ohana-Levi et al., 2019) and was therefore selected as a target variable indicating water condition of the crops.

The descriptive statistics of the five variables (four inputs and validation) are summarized in Table 1 for the vineyard and the peach orchard, including mean, standard deviation (SD), coefficient of variation (CV) and the range of values. Spatial representation of these five variables is illustrated in Figs. S1 and S2 (Supplementary material A). The relationships among the input variables for each of the plots are provided in a Kendall’s tau rank correlation matrices (Fig. 2).

2.3. Spatial autocorrelation

Positive spatial autocorrelation is present where adjacent observa-

Table 1
Descriptive statistics of the variables used as input and validation for the multivariate moment-matching criterion algorithm. TWI, ECa, NDVI, and CWSI stand for topographic wetness index, apparent electrical conductivity, normalized difference vegetation index, and crop water stress index, respectively. SD and CV stand for standard deviation and coefficient of variation, respectively.

Variable	Mean	SD	CV (%)	Range
	Vineyard			
Slope (°)	4.33	1.72	39.67	1.72–10.28
TWI	7.23	1.49	20.67	4–11.51
ECa (mS/m)	32.41	10.76	33.19	12.32–64.21
NDVI	0.5	0.08	15.81	0.16–0.74
CWSI	0.29	0.14	49.65	0–0.89
	Peach orchard			
Slope (°)	5.74	2.11	36.82	0.37–17.37
TWI	7.31	1.17	16.04	3.41–12.23
ECa (mS/m)	57.25	27.89	48.71	17.58–187.56
NDVI	0.67	0.02	3.72	0.56–0.75
CWSI	0.19	0.11	57.89	0–0.68

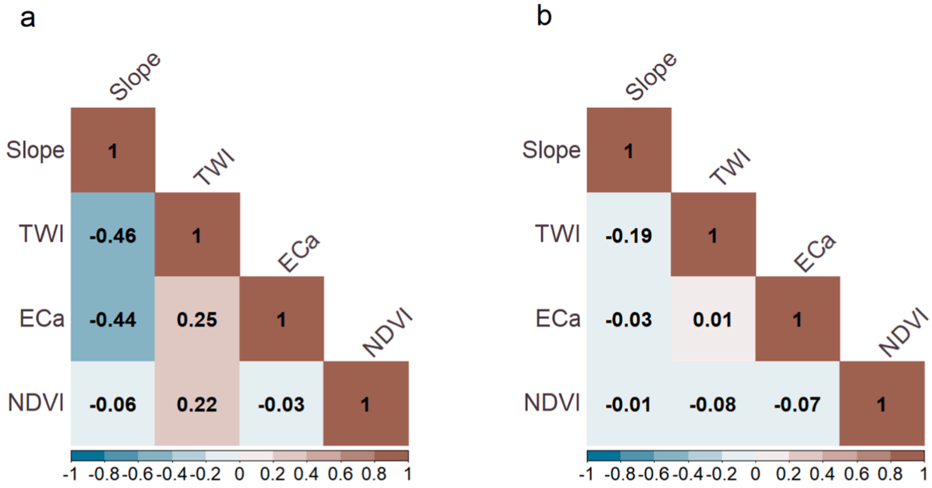

Fig. 2. Kendall's tau rank correlation matrices for the input variables of the vineyard (a) and peach orchard (b). TWI, ECa, and NDVI stand for topographic wetness index, apparent electrical conductivity, and normalized difference vegetation index, respectively.

tions have similar data values. Indication of spatial autocorrelation implies information redundancy, which, in the case of representative sampling while using a small number of observations, should be avoided. The stronger the spatial autocorrelation, the more important it is to minimize duplication of information within the sample size that is selected (Haining, 2015a). A representation of the distribution of an unknown variable based on a set of known variables should reduce the duplication of information to a minimum. Applying the MFM algorithm, therefore, requires definition of a minimum distance between potential sample observations within a specific subset of the population. We first quantified spatial autocorrelation using Moran's I statistic ($\alpha = 0.05$) to confirm that all variables display spatial dependence (Moran, 1948). Then, we computed a multivariate empirical semivariogram, showing how spatial structure varies with distance (Isaacs & Srivastava, 1989) to determine the minimum distance between two observations within a sample, when the data lacks stationarity (Oliver & Webster, 1986). An empirical semivariogram plots half the squared difference between two observations (i.e. semivariance) against their actual distance in space, averaged for predefined distance classes. The semivariogram consists of parameters defining its structure, including the sill (average half squared difference of two observations), range (the maximum distance at which pairs of observations display spatial dependence), and nugget (the variance within the sampling unit). The model semivariogram assumes that there is no spatial association for distances larger than the range (Wagner, 2003). The multivariate semivariogram computation, equivalent to summing univariate variograms, included the set of four input variables. It was calculated using the “adespatial” package in R (Stéphane Dray et al., 2020), with a minimum distance value of 8 m, a maximum distance that is equal to half the maximum length of each plot, and 80 classes of lag distances. A variogram model was then fitted

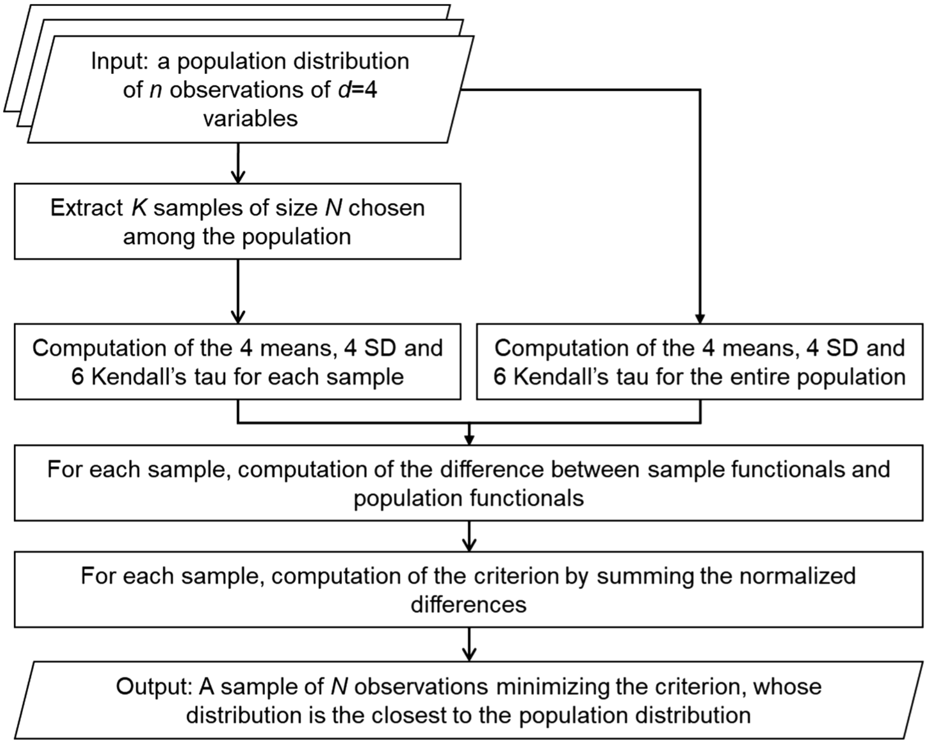
Input: a population distribution of *n* observations of *d*=4 variables
Extract *K* samples of size *N* chosen among the population
Computation of the 4 means, 4 SD and 6 Kendall's tau for each sample
Computation of the 4 means, 4 SD and 6 Kendall's tau for the entire population
For each sample, computation of the difference between sample functionals and population functionals
For each sample, computation of the criterion by summing the normalized differences
Output: A sample of *N* observations minimizing the criterion, whose distribution is the closest to the population distribution

Fig. 3. Flowchart of the multivariate moment-matching criterion for the specific case studies on the vineyard and peach orchard (Fig. 1). SD stands for standard deviation.

using the “gstat” package in R (Pebesma, 2004), with the Bessel and Matern (with an optimally fitted kappa) variogram models used for the vineyard and peach orchard, respectively. For each semivariogram, the range values were used as the minimum distance between observations within a sample. It should be noted that other methods of computing the minimum distance required to avoid spatial dependence exist (Griffith, 2004; Koenig, 1999), and the user may select any method they see fit to determine this parameter.

2.4. Multivariate functional-matching criterion

In this study, we considered the problem of finding the optimal location of observations for the variable CWSI given observations from the variables TWI, NDVI, Slope and ECa. This means that the observations space is $\mathscr{X} = \mathbb{R}^d$ with $d = 4$ variables and for $j = 1, \cdots, 4$, $X_{(j)}$ denotes the jth component of the random vector X. Algorithm 1, represented in Fig. 3, proposes to match 14 functionals of the population distribution to the sample distribution and find the sample that has the closest moments. These functionals are:

- the 4 different means $\varphi_1(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}\left[X_{(1)}\right], \varphi_2(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}\left[X_{(2)}\right], \varphi_3(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}\left[X_{(3)}\right], \varphi_4(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[X_{(4)}]$,
- the 4 different standard deviations $\varphi_5(\mathbb{P}) = sd_{\mathbb{P}}\left[X_{(1)}\right]$, $\varphi_6(\mathbb{P}) = sd_{\mathbb{P}}\left[X_{(2)}\right], \varphi_7(\mathbb{P}) = sd_{\mathbb{P}}\left[X_{(3)}\right], \varphi_8(\mathbb{P}) = sd_{\mathbb{P}}[X_{(4)}]$,
- the 6 different Kendall’s tau $\varphi_9(\mathbb{P}) = \tau_{\mathbb{P}}\left[X_{(1)}, X_{(2)}\right]$, $\varphi_{10}(\mathbb{P}) = \tau_{\mathbb{P}}\left[X_{(1)}, X_{(3)}\right]$, $\varphi_{11}(\mathbb{P}) = \tau_{\mathbb{P}}\left[X_{(1)}, X_{(4)}\right]$, $\varphi_{12}(\mathbb{P}) = \tau_{\mathbb{P}}\left[X_{(2)}, X_{(3)}\right]$, $\varphi_{13}(\mathbb{P}) = \tau_{\mathbb{P}}\left[X_{(2)}, X_{(4)}\right], \varphi_{14}(\mathbb{P}) = \tau_{\mathbb{P}}\left[X_{(3)}, X_{(4)}\right]$, where $\tau_{\mathbb{P}}[X, Y]$ denotes Kendall’s tau for the distribution $\mathbb{P}$ between variables X and Y.

Therefore, the criterion is defined as $\varphi(\mathbb{P}) := \sum_{j=1}^{14} \varphi_j(\mathbb{P})$. In practice, there is no access to the true distribution $\mathbb{P}^*$ of the variables, but only to a population of size n, whose empirical distribution will be denoted by $\mathbb{P}_n$. One could wonder about the difference between the true criteria $\varphi(\mathbb{P})$ and its estimated population version $\varphi(\mathbb{P}_n)$. As $\varphi_1, \cdots, \varphi_{13}$ are regular functionals, we can apply the central limit theorem to them (under suitable regularity conditions), to get the asymptotic approximation $\varphi(\mathbb{P}) - \varphi(\mathbb{P}_n) \approx \sqrt{n}G$ for some Gaussian variable G. If S a random sample of size N (among the n population points) is considered, the quantity $\Delta_{n,S} := \left(\varphi(\mathbb{P}_n) - \varphi(\mathbb{P}_{\mathbb{S}})\right)^2$ is a random variable bounded below by 0. Let us call its (essential) minimum $\Delta_{n,\min}$ conditionally to $\mathbb{P}_n$. By classical results, the best sample S^* of K random samples should satisfy $\Delta_{n,S_{k^*}} \approx \Delta_{n,\min} + O(K^{-1})$. A precise computation of the distribution of the best error $\Delta_{n,\min}$ achievable by a sample of size N among n observations is out of the scope of this article and left for future research.

Algorithm 1. (*Multivariate functional-matching algorithm for selection of a representative sample*)

```
Input 1: A population of $n$ observations of the $d$ variables and their positions in a
  coordinate system
Input 2: Tuning parameters: minimum distance $r$, size of the sample $N$, number of
  samples $K$.
1. Select a subset $\mathbb{S}$ of the population to be represented
2. For each $k = 1$ to $K$ do:
   a. Choose at random one sample $S_k \subset \mathbb{S}$ of size $N$ satisfying the minimum distance
      condition.
   b. Compute the $d$ means using the sample $S_k$.
   c. Compute the $d$ standard deviations using the sample $S_k$.
   d. Compute the Kendall’s tau among all pairs of $d$ variables using the sample $S_k$.
3. Compute the means, standard deviations, and Kendall’s tau on $\mathbb{S}$.
4. Compute and normalize the difference between computed on the subset $\mathbb{S}$ and on
   sample $S_k$.
5. For each $k = 1$ to $K$ do:
   a. Compute the criterion of $Crit(S_k)$ by summing the normalized differences.
6. Compute $k^*$ as the minimizer of the criterion $Crit(S_k)$
Return the sample $S_{k^*}$ and the associated criterion $Crit(S_{k^*})$.
```

The output of Algorithm 1 is not an estimator, but rather a sample. It consists of a list of locations (e.g., trees, grapevines) that are recommended to be used for sampling.

Note that several constraints should apply to the set of inputs. First, it is better to use variables that are related to the target variable. If this condition is not satisfied, then Algorithm 1 is only as good as simply selecting a sample at random among the population. Second, the minimum distance r and the size of the sample N should not be too large in both cases so that the set of samples satisfying the conditions has enough elements. Finally, the number of samples K to be compared should not be too small to have a good approximation of the best criterion, but it should not be too large to ensure a reasonable computation time. Should the user wish to express different importance levels for specific variables, weights may be considered. This algorithm is available in the R package “MFunctMatching” (Derumigny & Ohana-Levi, 2020).

2.5. Criterion reduction for defining sample size

In some cases, the user may determine a predefined sample size S (according to existing resources such as a specific number of sensors, limited manpower to conduct the sampling, and so on). This will require determining a set of observations that represent the population distribution as accurately as possible, given this limitation. In that case, they may directly apply Algorithm 1 to the data. In other cases, the choice of the number of observations to include in the sample is less obvious. One possibility is to determine how many observations are necessary to represent the population distribution within a given accuracy. We propose a method to choose a minimum sample size, for cases where the number of observations is not predefined. The sample size is a function of the error that the user is willing to tolerate in representing of the population. The algorithm for defining the sample size to select relies on different executions of Algorithm 1 for N in an interval $[N_{min}, N_{max}]$, until reaching the maximum sample size, or until reaching convergence. In order to have a higher precision, for each value of N, Algorithm 1 was run $R = 1000$ times. The user may select the smallest sample size that satisfies a certain reduction of the maximum number achieved, i.e. the smallest sample size N^* such that the reduction of the criteria from N_{min} to N^* is a given fraction α, such as 90%, of the maximal reduction of the criteria from N_{min} to N_{max}. This procedure is displayed below as Algorithm 2 and is available in the R package “MFunctMatching” (Derumigny & Ohana-Levi, 2020).

Algorithm 2. (*Choice of the optimal sample size satisfying a reduction of the criteria*)

```
Inputs 1–2: As in Algorithm 1.
  Input 3: A target percentage $\alpha$ denoting the chosen reduction of the mean criteria.
  Input 4: A range of relevant sample sizes $[N_{min}, N_{max}]$.
  Input 5: A number $R$ of replications for the computation.
  Input 6 (optional): A tolerance $tol$ for the convergence of the mean criteria
1. For each $N = N_{min}$ to $N_{max}$ do
   a. Run $R$ times Algorithm 1 using the sample size $N$. This gives a mean
      criterion $C_{mean,N}$.
   b. If $|C_{mean,N} - C_{mean,N-1}| < tol$ then affect $N_{max} := N$ and go to step 2.
2. Compute $N^* := \min\{N : C_{mean,N_{min}} - C_{mean,N} > \alpha(C_{mean,N_{min}} - C_{mean,N_{max}})\}$
Return the chosen sample size $N^*$
```

The minimum sample size N_{min} should be chosen large enough to enable reliable calculation of moments. Note that it is possible to choose $N_{max} = n$ and wait for the convergence of the criteria to 0. The chosen sample size N^* increases with the criterion reduction fraction α, as a bigger reduction corresponds to a larger sample size. In a complementary way, a visualization of the mapping $N \mapsto C_{mean,N}$ may provide information on the degree of accuracy that is granted by a larger sample size, or that is lost by a smaller sample size. In this study, the criteria reduction percentage was chosen as $\alpha = 90\%$.

An empirical analysis of the sensitivity of Algorithm 2 to the minimum distance r was performed, to evaluate the effect of spatial autocorrelation on the sample size for each of the plots. Algorithm 2 was tested using$K = 500$, and$10 \leq r \leq 16$. In this study, *tol* was chosen as $\min\limits_{N \in [N_{min}, N_{max}]} C_{mean,N}$, however tolerance may be defined according to the user's choice.

2.6. Validation against uniform random sampling and cLHS

In this study, we know the measured value of CWSI at each of the trees/vines, and therefore this information can be used to compare the population distribution of CWSI with a sample distribution of CWSI. This allows us to measure the success rate of the proposed algorithm and the accuracy by which the selected sample represents the target variable CWSI values for the entire population of vines/peach trees (specified in subsection 2.2). Fig. 4 illustrates a density estimate of CWSI for the two plots, using kernel smoothing. The CWSI values in the vineyard were, on average, higher than those in the peach orchard, due to the strategy of deficit irrigation and subsequent greater water stress. The variance of CWSI in the vineyard was also higher due to a stronger spatial heterogeneity of terrain and soil attributes.

An analysis was conducted to assess the quality of the outcome sample and compare the distribution of the population CWSI with the CWSI of the sample generated by Algorithm 1 after defining the optimal number of locations for the sample using Algorithm 2. The comparison between the sample and the entire population was performed using the Kolmogorov-Smirnov Goodness-of-Fit (KS) test statistic (Marsaglia et al., 2003), SD, mean, CV, range, quantiles of the population represented by the sample and value of the criterion of the chosen sample.

Finally, the MFM algorithm was compared to two reference methods, namely a uniform random sample with identical spatial constraints (e.g. the same minimum distance as specified in Algorithm 1, Input 2), as well as to the cLHS provided in the R package "clhs" (Roudier, 2011). A qualitative comparison between MFM and cLHS is also detailed in Table 2. The cLHS algorithm was applied using the default parameterization values specified by the "clhs" package. These included the number of iterations for the simulated annealing process, initial temperature (the control parameter of the annealing process) at which the simulated annealing begins (between 0 and 1), and a decreasing factor between 0 and 1, signifying the rate at which temperature decreases at each iteration. We ran each of the three methods (MFM, uniform random sampling, both with spatial constraints and cLHS) 3000 times. Each time, the MFM chose the best sample (i.e., corresponding to the lowest criterion) amongst 2000 random samples drawn uniformly. For each of these best samples, we computed statistics comparing the population and sample distributions of the target variable CWSI: differences in mean, differences in standard deviation, and KS. We compared the

Table 2
A qualitative comparison between the conditioned latent hypercube sampling (cLHS) and the multifunctional matching (MFM) algorithms. N is population size, while n stands for sample size.

	cLHS	MFM
Assumption regarding the sample size	$n \ll N$	$n \ll N$
Tuning parameter for the model	Simulated annealing with temperature T, the rate of temperature decrease d, number of iterations or stopping criteria	Straightforward: number of tested subsamples
Difficulty calibrating the tuning parameter	Depends on the convergence properties of the Markov chain	Straightforward: in general, 10,000 subsamples yield acceptable performance
Type of matching	Iterative process	Brute-force comparison
Moments to match	Fixed: $n \times k$ quantiles + $k(k+1)/2$ correlations, where k is the number of numerical variables	Flexible: any combination of moments; by default, k means + k standard deviations + $k(k-1)/2$ Kendall's tau
Bias/variance compromise for large sample size	Small bias: the quantiles ensure that the whole distribution of each variable in the subsample is close to their population distribution	Less effective for large sample sizes than the cLHS concerning the accuracy of other moments than the ones used for the fit
Bias/variance compromise for small sample size	Less effective for small sample sizes (Werbylo & Niemann, 2014)	Small variance: the (by default) smaller number of moments to match makes the estimation more stable for small sample sizes
Type of method	Nonparametric (the number of constraints increase with n)	Parametric (fixed number of constraints) but could be nonparametric if the set of moments is increased

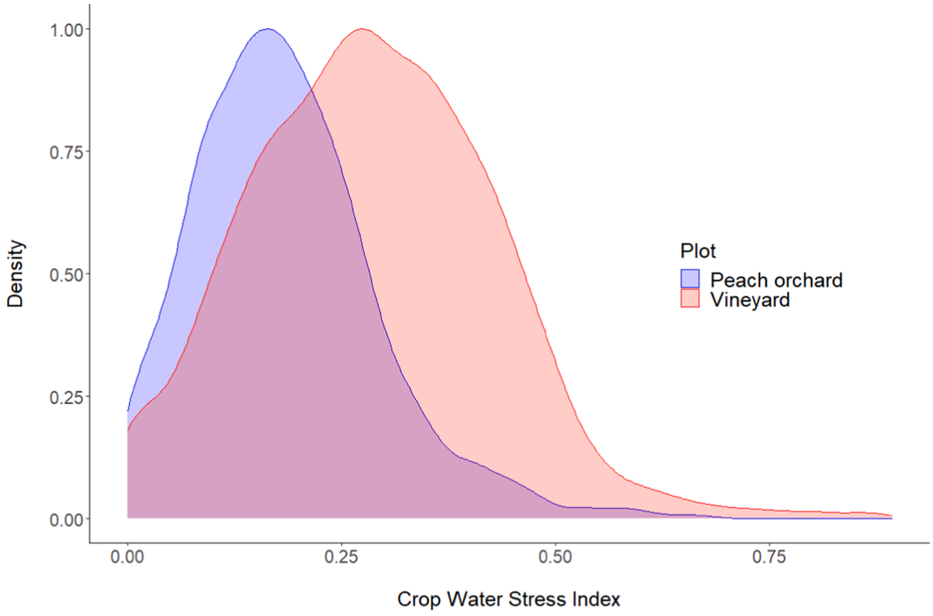

Fig. 4. Crop water stress index density plots quantified by drone images acquired for the vineyard (September 5, 2017) and peach orchard (August 29, 2017).

distribution of these 3 goodness-of-fit measures for MFM against the two other methods using Welch's *t*-test for statistical difference between two means, and we compared them graphically using quantile–quantile plots (QQ-plots). The Q-Q plots enable visualization of the distribution of differences between the sampled moments and the population moments for the MFM against the two reference methods. This procedure was implemented in R (R Core Team, 2020).

2.7. Assessment of management zones

In recent years, variable-rate in-field spatial management of various agricultural practices (irrigation, fertilization, and plant protection) has been recognized by the scientific community as valuable due to the potential for increased profitability and benefits for sustainable crop production (Ben-Gal et al., In Press; Zhang et al., 2002). One of the most efficient techniques for precision agriculture is using site-specific MZs. MZs are homogeneous sub-areas within a field based on spatially quantitative measures (Gavioli et al., 2019). An additional aspect of testing the accuracy of population representation through the sample derived by the MFM algorithm is determining the number of sample observations that fall within each management zone. In an accurate representation of the partitioning of the plots, the population share of observations in each MZ should be equal to the sample share of observations in each zone, where the sample has been chosen by the MFM algorithm. The MZs for each plot were estimated using a weighted multivariate clustering method (Ohana-Levi et al., 2020). First, the variables were given weights according to their respective contribution to the explained variance by applying principal component analysis (PCA) (Dunteman, 1989). This step was followed by iteratively applying fuzzy c-means algorithm for a different number of clusters each time. The best clustering result was determined by calculating an average silhouette width (ASW), comparing cluster tightness and separation (Rousseeuw, 1987). For this, the iteration with the highest ASW was chosen to define the number of MZs for each plot. The MZ delineation process was performed using the R packages "ppclust" (Cebeci, 2019) applying the fuzzy c-means algorithm and "cluster" (Maechler et al., 2019) for the ASW analysis.

3. Results

3.1. Defining minimum distance by estimating spatial autocorrelation

The four input variables, as well as the target variable, showed significant spatial autocorrelation according to Moran's I test at the level of$\alpha = 0.0001$. The multivariate semivariogram applied to the four variables provided the structural parameters. For this study, we considered the estimated range as the minimum distance between each pair of sample observations. Minimum distance values were 13.89 and 15.26 m for the vineyard and peach orchard, respectively.

3.2. Chosen sample size and location of the sample observations

The results for the sensitivity analysis are presented in Table 3, illustrating the relationship between spatial autocorrelation and sample size. In the vineyard, with a smaller areal coverage then the peach orchard, the spatial constraint prevented a convergence of the mean criterion for$r < 10$. In the peach orchard, which consists of a larger area, r was not found to constrain the spatial distribution of the observations for$r < 16$. In both cases, higher r values were associated with smaller sample size.

The inputs to the MFM algorithm are specified in Table 4. The results in Fig. 5 represent the outcome of Algorithm 2 using the respective minimum distance$r$ that were defined by the ranges of the semivariogram for each of the two fields. Convergence in the vineyard occurred at 69 observations and for the peach orchard, the spatial constraints allowed for 89 observations. For both plots, the sample size for 90% criterion reduction was 22. The final location of the sample observations for both plots are illustrated in Fig. 6.

3.3. Validation

After applying the MFM algorithm (Algorithm 1), resulting in an approximate best sample of observations for each plot, analysis was conducted to compare the sample to the true population values of CWSI. In Table 5 the sample (22 observations) and the entire population are compared in terms of KS test statistic, SD, mean, CV, range, quantiles of the population represented by the sample and value of the criterion of the chosen sample. For both plots, there was no significant difference between the distributions of the entire population of CWSI values and the sample ($p = 0.71$ and $p = 0.1$ for the vineyard and orchard, respectively). For both plots, SD differences between CWSI population and the selected sample were small (0.04 and 0.01 for vineyard and peach orchard, respectively), and the sample set was able to represent 99.6 and 93.3% of the range of values of the CWSI population.

The results of the validation of the MFM algorithm against uniform random sampling with spatial constraints and cLHS are detailed in Table 6. Generally, the MFM method performed better than the uniform

Table 4
Input variables and parameters for the multiple functional-matching criterion algorithm. TWI, NDVI and ECa stand for topographic wetness index, normalized difference vegetation index and apparent electrical conductivity, respectively.

	Vineyard	Peach orchard
Variables	Slope, TWI, ECa, NDVI,	Slope, TWI, ECa, NDVI
Population size$n$	3523	2151
Sample size$N$	22	22
Minimum distance$r$ (m)	13.9	15.3
Number K of samples considered	20,000	20,000

Table 3
Sample size estimation, maximum number of observations and sample size at convergence of the mean criterion,$C_{mean,N}$, according to different minimum distance values, with$\alpha = 90\%$ criterion reduction, $K = 500$ samples for each minimum distance$r$.

Minimum distance (m)	Vineyard			Peach orchard		
	Sample size	Maximum number of observations	Sample size at convergence of$C_{mean,N}$	Sample size	Maximum number of observations	Sample size at convergence of$C_{mean,N}$
10	24	127	124	24	200	194
11	23	105	–	24	171	170
12	23	94	–	23	148	147
13	22	76	–	23	112	111
14	22	68	–	22	101	98
15	21	61	–	22	92	91
16	21	52	–	22	82	–

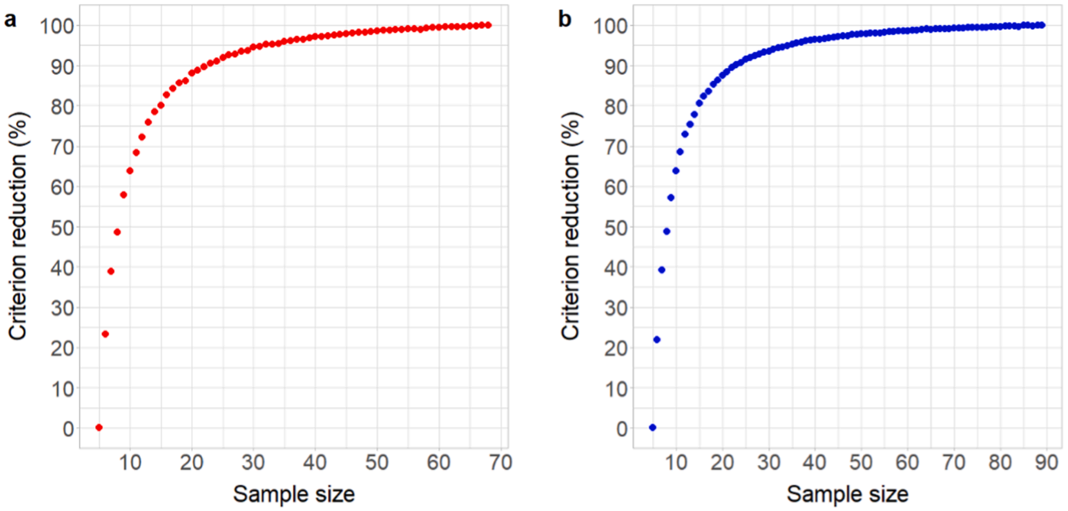

Fig. 5. Illustration of simulated percent criterion reduction against sample size for the vineyard (a) and peach orchard (b).

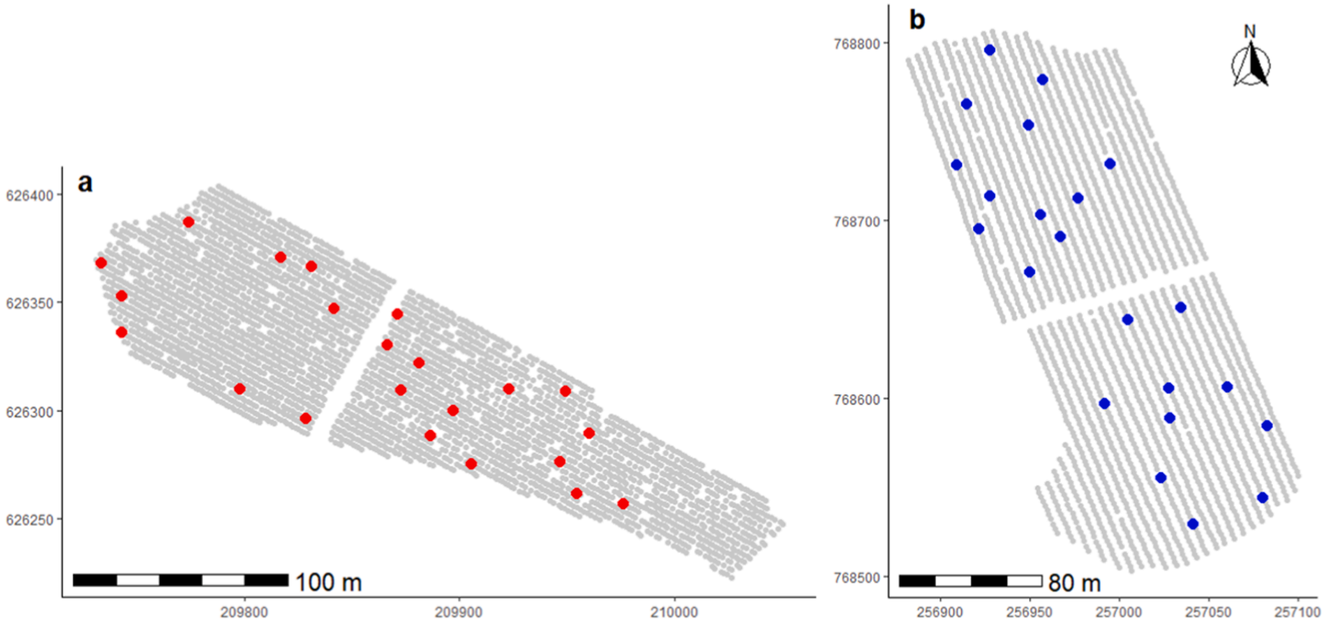

Fig. 6. Spatial representation of the optimized sample location according to the multivariate moment-matching criterion in the vineyard (a) and the peach orchard (b).

random sampling and cLHS for both fields. This means that the moments of the samples selected by the MFM algorithm were closer to the population moments compared to the moments of the uniform random samples or those generated by cLHS. More precisely, the difference between each MFM sample moment and the corresponding population moments were on average smaller than the differences with the uniform random sample moment and those of cLHS. The only exception was the SD for the peach orchard which had a slightly larger difference with the MFM sample than both reference methods, and the SD difference range was smaller for cLHS. The SD differences for the peach orchard data of MFM were not significantly different from the other two reference methods (*p-value* = 0.61 & 0.704 for the uniform random sample and cLHS, respectively). The highest dissimilarities between the MFM sample results and both the uniform random sampling and cLHS, for both plots, were for differences between population and sampled range values. On average, the MFM sampled observations showed lower ranges of differences for all moments compared to the ranges of differences resulting from the uniform random sampling or cLHS. In the vineyard, the differences between the MFM and the two reference methods were significant for mean, SD and KS statistic, while in the peach orchard there were significant differences only for the average differences in mean values, and between MFM and cLHS for the KS statistic.

Differences between the methods were also visualized using QQ-plots (Figs. S3, S4). In general, at lower values, there were almost no differences between MFM and both reference sampling methods for all moments (mean, SD, and KS), with an exception for MFM against cLHS in the vineyard. However, larger differences between the sampling techniques were observed in the vineyard dataset (Figs. S3, S4 a, c, e). The QQ-plot (dotted line) spanned further away from the diagonal (continuous line) towards the Y-axis, representing the uniform random sampling method/cLHS, illustrating larger differences between samples and population moments. These results support the *t*-test analysis provided in Table 6, showing that the MFM method produced sample observations that were significantly closer to the population CWSI in terms of mean, SD, and distribution (KS), compared to sample observations

Table 5
Comparison between population and chosen sample statistics of crop water stress index for the vineyard and peach orchard plots, where the sample was chosen among $K = 20,000$ uniformly drawn samples.

	Population	Sample
	Vineyard	
Kolmogorov-Smirnov goodness-of-fit		$p = 0.711$
Standard deviation	0.14	0.18
Mean	0.29	0.30
Coefficient of variation	0.5	0.6
Range	0–0.89	0–0.79
Quantiles of population represented by sample (%)		0–99.6
Criterion for the chosen sample set		2.5
	Peach orchard	
Kolmogorov-Smirnov goodness-of-fit		$p = 0.098$
Standard deviation	0.11	0.1
Mean	0.19	0.15
Coefficient of variation	0.58	0.65
Range	0–0.68	0.006–0.4
Quantiles of population represented by sample (%)		2.2–95.5
Criterion for the chosen sample set		3.02

derived from the uniform random sampling or cLHS. For the peach orchard, the differences between the methods were significant only for mean differences (uniform random sampling) and for mean differences and KS (cLHS). This result is also supported by Figs. S3b & S4b, f, showing a larger distance between the QQ-plot (dotted line) and the diagonal than for the other moments of the peach orchard (Figs. S3 d, f & S4 d).

3.4. Management zone representation

In Fig. 6, the location of the chosen set of observations for both plots is shown. After applying a weighted, multivariate fuzzy C-means cluster analysis to the plots' datasets, they were each separated into two clusters (Fig. 7). The population share of vines/trees in cluster 1 and cluster 2 was, correspondingly, 0.41 and 0.59 for the vineyard and 0.34 and 0.66 for the peach orchard. The selected observations showed a similar distribution, where, in the vineyard, eight observations (36% of selected observations) were included in cluster 1 and 14 observations (64% of selected observations) in cluster 2. For the peach orchard, there were six observations (27% of selected observations) located within cluster 1 and 16 (73%) in cluster 2.

4. Discussion

The main objective of this study was to construct an algorithm for defining locations to sample a target feature across an agricultural field, using known explanatory variables observed over the whole population. The proposed algorithm (Algorithm 1) requires several inputs, the first of which (Input 1) is a set of multivariate spatially defined observations. The variables that serve as inputs to the model should be linked to the target variable. Such an association may be established using expert knowledge, based on past experiments, and literature review. An input set of variables independent from the target variable to be sampled will not yield an accurate sample set of observations. In this study, some of the variables selected (e.g. TWI, NDVI, ECa and slope) are known to be associated with soil moisture levels (Costa et al., 2014; Mohanty & Skaggs, 2001; Nanda et al., 2019), which in turn determine water availability to the plants (Prueger et al., 2019). NDVI is an indicator of plant vigor, that has been found to be associated with drought stress in crops (Lee & Park, 2019). The choice of the input variables was also based on the degree of availability and efficiency of measurements in the field. TWI and slope are both products that were generated from a DEM, which is commonly available, NDVI is easily derived from UAV, airborne, or satellite platforms and is widely used in agricultural research and practices (Kamilaris et al., 2017). Measurements of ECa, representing soil properties that affect crop productivity, including soil texture, salinity, organic matter, drainage conditions, and subsoil features, are reliable, quick and easy to acquire (Corwin & Lesch, 2013). It should be noted that the algorithm may also be modified and applied using other types of data, such as categorical variables, circular variables, etc.

Acknowledging the presence of spatial autocorrelation in geographically distributed data is imperative for sample size determination (Griffith, 2013), thus the final set of input variables was used to define the minimum distance between each pair of sample observations (Input 2). The main purpose of this step is to avoid having a number of observations that provide dependent information and could be biased

Table 6
Validation results for the target variable "crop water stress index", based on 3000 replications. A comparison between averaged differences between the functionals of the samples and the population was based on multiple runs of the multivariate functional-matching algorithm, the conditioned Latin hypercube sampling (cLHS) and uniform random sampling with a spatial constraint for the vineyard and peach orchard, using a sample size of 22 (corresponding to 90% criterion reduction) for both plots. The percentages in parentheses denote the relative difference between MFM and the uniform random sample and cLHS.

Average differences between normalized sample and population moments of the target variable	MFM (reference)	cLHS	Uniform random sample
	Vineyard		
Average difference of means	0.017	0.023 (−25.3%)	0.021 (−18.87%)
Range of mean difference	0.087	0.11 (−18.23%)	0.11 (−17.46%)
Average difference of SD	0.017	0.019 (−13.18%)	0.018 (−7.67%)
Range of SD difference	0.067	0.09 (−25.3%)	0.103 (−34.25%)
Average D statistic	0.160	0.175 (−8.38%)	0.167 (−3.8%)
Range of D statistic	0.296	0.342 (−13.28%)	0.321 (−7.87%)
Welch's *t*-test			
Differences between means		$t = -12.08, p < 0.001$	$t = -9.55, p < 0.001$
Differences between SD		$t = -6.05, p < 0.001$	$t = -3.71, p < 0.001$
Differences between KS statistics		$t = -9.66, p < 0.001$	$t = -4.77, p < 0.001$
	Peach orchard		
Average difference of means	0.016	0.17 (−4.99%)	0.017 (−4.44%)
Range of mean difference	0.077	0.087 (−11.62%)	0.094 (−18.32%)
Average difference of SD	0.017	0.016 (+0.87%)	0.016 (+1.06%)
Range of SD difference	0.066	0.064 (+3.95%)	0.077 (−13.92%)
Average D statistic	0.172	0.176 (−2.14%)	0.173 (−0.76%)
Range of D statistic	0.316	0.373 (−15.12%)	0.345 (−8.37%)
Welch's *t*-test			
Differences between means		$t = -2.106, p = 0.035$	$t = -2.08, p = 0.038$
Differences between SD		$t = 0.38, p = 0.704$	$t = -0.51, p = 0.61$
Differences between KS statistics		$t = -2.268, p = 0.023$	$t = -0.88, p = 0.377$

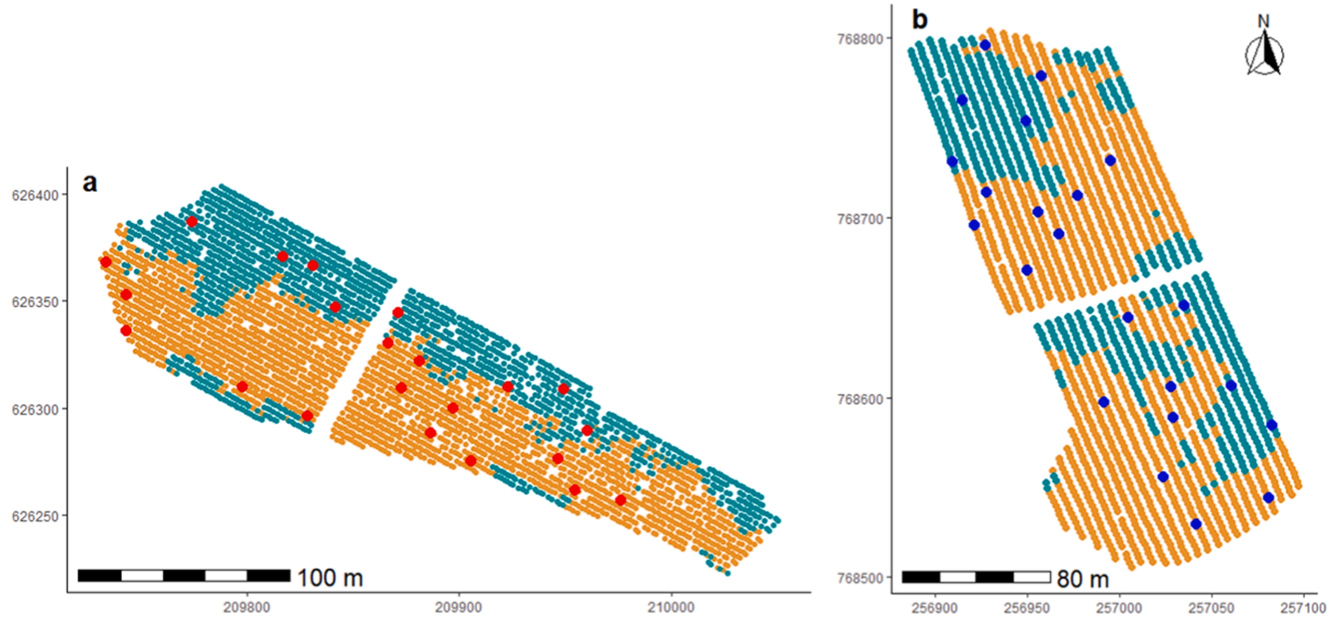

Fig. 7. A spatial representation of the clusters in the vineyard (a) and the peach orchard (b), along with the sampled points selected by the multifunctional matching algorithm in the vineyard (red) and peach orchard (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

compared to the entire population. In general, the more heterogeneous the population, the larger the sample size needed to obtain a specific level of precision (Israel, 1992). In addition, a stronger spatial autocorrelation (and weaker spatial variability), with a larger model-semivariogram range value, requires a smaller sample size (Input 2) to represent the multivariate distribution, since more observations are similar to others (Haining, 2015b). In the case study presented in this work, the characteristics of the vineyard displayed both higher variability (Table 1) and a higher spatial variability (Table 4) than the attributes of the peach orchard. The number of observations suggested by Algorithm 2 was slightly sensitive to the minimum distance$r$, which corresponds to the spatial variability. More precisely, the estimated number of samples for a givenα may shift with varying values of r. In some cases, a small r will not enable enough observations to be included for the mean criterion to converge, thus decreasing the approximated sample size K when r increases. While applying Algorithm 2, other parameters can also have an influence on the outcome, as well as the structure of the data itself. The outcome of Algorithm 1 depends on the chosen number of replications (Input 2). A higher number of replications will result in a higher probability of reducing the criterion and finding a more accurate fit to the population moments, thereby introducing a better sample to represent the distribution of interest.

In this present study, the vineyard was smaller in area than the orchard (with a ratio of 57%). However, the chosen number of observations for both plots, after applying a criterion reduction of 90%, was 22. With a larger area and lower tree-density, the peach orchard was found to have a higher range of spatial autocorrelation (Table 4) and lower variability (Table 1) than the vineyard. Although the vineyard consisted of a more observations than the orchard (3523 vines vs 2151 trees), when increasing the percentage of criterion reductionα above 90% (and thereby increasing the accuracy of the representation of the population), the number of observations required to represent the population was larger for the orchard than for the vineyard (for example, 95% criterion reduction resulted in 34 vs 31 observations for the orchard and the vineyard, respectively). These differences in the estimated required number of observations for higher accuracy levels can be explained by the difference between the two distributions. Indeed, the underlying data-generating processes that correspond to each field are not identical; and even if we use some normalization (step 3 of Algorithm 1), higher-order effects would be expected to have increasing influence when the accuracy level is sufficiently high. When convergence does not occur (for example because of a small population size or equivalently, because of a high autocorrelation), the size of the plot may also have an effect on sample size. In other words, in the cases where the spatial constraint limits the sample size, larger plots will require more observations to maintain a certain level of accuracy. Similarly, another factor that affects the sample size is spatial autocorrelation, since a larger spatial variability will require a larger sample size.

Validation against measured CWSI was conducted to provide an assessment of the performance of the suggested algorithm. In Table 5, the sample moments are shown in comparison to the population moments, to assess the effectiveness of the algorithm in finding a representative sample. The results show small differences between sample and population mean and SD of CWSI. The range of values was not expected to overlap between the sample values and the population, since outliers should not be incorporated in the selected sample. Indeed, a useful sample should not account for extreme values. Therefore, we chose to use only means, SDs and Kendall's tau as functionals of interest (Algorithm 1), instead of the minimum and the maximum which would tend to select outliers. Table 5 shows that the quantiles of the population represented by the range of the sample were quite high, 99.6% for the vineyard and 93.3% for the peach orchard, indicating that the selected set of observations for both plots was able to represent most of the population distribution, excluding extreme values. In cases where the user wishes to represent a narrower part of the distribution (only the inter-quantile range, for example), the algorithm should be applied on a subset of the multivariate population data that lies within the specified frequency levels. In this study, we chose to exclude border vines/trees, so the application of the algorithm was based on a subset of the entire population. The MFM algorithm is randomized, therefore different applications may yield different outputs even with the same input. This is not problematic, as it is entirely possible that different sets of samples could each satisfactorily represent the population.

Validating the performance of the algorithm also included a comparison to a uniform random sampling with the semivariogram-derived minimum distance and to cLHS for each plot. Table 6 shows that overall, the MFM algorithm performed better than the uniform random sampling and cLHS techniques in the vineyard, where the dataset had greater variability and spatial heterogeneity. In the peach orchard, improvements shown by the MFM model were more subdued. The *t*-test analyses showed that the sample set generated by the MFM algorithm accounted only for significant improvement in representing the population mean of CWSI against both reference methods, and improvement in representing the population distribution compared to cLHS. However, in both plots, the reduction in the ranges of difference between sample set functionals and population functionals was high and significant, with an exception of SD for the peach orchard, where cLHS and MFM algorithms were not found to be significantly different. This means that, on average, the MFM performed better than uniform random sampling for the peach orchard in representing the (unknown) average value of the target but similarly in terms of representing the SD. In other words, on 3000 replications of the process, the MFM model showed higher levels of accuracy compared to the uniform random sampling or cLHS (smaller differences on average, Table 6). cLHS has been widely used in digital soil mapping, usually applied to cases where a high number of observations should be sampled within large areas of research (Brungard & Boettinger, 2010). Defining sample designs within agricultural fields may benefit from the MFM, which has fewer tuning parameters and a straightforward calibration process. From this study, it also seems to be more suitable for smaller sample sizes for plots with spatial autocorrelation (Table 2). Further study should be conducted to test this assumption. Additionally, the MFM model was not designed to specifically account for SSH (Wang et al., 2016), and it is currently assumed that the accuracy and efficiency of the algorithm is unaffected by either the presence or the absence of SSH in the data. This assumption should be tested in the future followed by possible specific modifications to the algorithm.

Finally, the MFM algorithm generated sampling sets that corresponded to the spatial patterns of the plots (Fig. 7). Using multiple variables, two clusters were estimated for each plot, with a different number of vines/trees in each cluster. The population shares of observations in each cluster were quite similar to the selected sample shares of observations in each cluster. This means that the suggested algorithm could account for the spatial variability and succeed in representing the population's repartition into the clusters. However, the MFM algorithm is a non-spatial platform for representing the distribution of a certain population, with a recommended choice of incorporating a minimum distance condition to the calculation. MZs may be used as a validation tool to assess the certainty of the resulting sample set, and indeed, as the sample is intended to be representative of the population, it is coherent that the repartition of the observations of this chosen sample into both clusters should be close to the repartition of the whole population.

The suggested framework was designed to select the best locations to sample in order to represent an unknown population distribution of a certain field attribute. Our study case demonstrated a scenario where a relatively low number of samples represented a large number of population observations (0.6% and 1% of the vines/trees, respectively). These approximately best locations are determined using information from other proxy variables. Using such a method may allow for a relatively small number of samples that would sufficiently represent the entire population of objects in the field, such as in the cases of sampling or sensing for soil or plant water or nutrient status, plant size, yield or plant physiological parameters, provided that the input variables are related closely enough to the target variable for sampling. The MFM may be applied to non-spatial data as well, when dealing with problems of representing population distributions, such as sampling honeybee colonies for infections (Fries et al., 1984) or analysis of genotypic variation for specific crops (He et al., 2017; O'Toole & Moya, 1978). Similarly, it may be used with time-series data to analyze specific intervals/timesteps that may be sampled to represent the entire dataset.

5. Summary and conclusions

The new MFM algorithm was designed to select the best set of sample locations based on functional matching of a random vector. Specified functionals of each set of observations are compared to the population functionals and result in a criterion that is determined by their differences. This work shows how to apply mean, SD and Kendall's Tau as the functionals of matching; however, the method can be adjusted and applied using other moments/functionals that could help to represent the distribution. The algorithm can also be applied using only one input variable, as long as it is associated to the target variable. Validation of the model results against a specified target variable showed a higher accuracy, better precision, and an overall stronger performance of the MFM model compared to a uniform random sampling with spatial constraint and to cLHS. Model performance was better for the vineyard that had higher variability in the data than for the peach orchard.

We applied the MFM based on a static dataset; however, it may be used dynamically throughout the season. Some variables in agricultural applications are known to change during the growing season (e.g. NDVI), and a sampling design may change along with the known variables for each sampling campaign.

The number of sample sites defined using Algorithm 2 was based on an arithmetic reduction of the criterion. Future work may include a reduction of the criterion based on minimizing cost function from an economical point of view, or the penalty of losing information that may lead to inferior estimations in terms of expenses to the farmer, as opposed to the cost of sampling. This type of modeling approach will have to rely on quantifying the cost of error or bias to the farmer. Additional data collection and studies in this direction are needed in order to establish such a modeling framework.

Finally, the application of the proposed algorithm may be adopted by other fields of research where small sample sizes are required and is not limited to agricultural practices. Sampling designs of spatially and non-spatially varying datasets for representation of their distribution may yield high accuracy levels of information by utilizing the modeling framework and assisting in decision-making processes.

CRediT authorship contribution statement

N. Ohana-Levi: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **A. Derumigny:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. **A. Peeters:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing – review & editing. **A. Ben-Gal:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **I. Bahat:** Data curation, Investigation, Methodology, Formal analysis. **L. Katz:** Data curation, Investigation, Methodology, Formal analysis. **Y. Netzer:** Data curation, Investigation, Methodology, Resources, Project administration. **A. Naor:** Data curation, Resources. **Y. Cohen:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is a part of the "Eugene Kendel" Project for Development of Precision Drip Irrigation funded via the Ministry of Agriculture and Rural Development in Israel (Grant No. 20-12-0030). The project has also received funding from the European Union's Horizon 2020 research and innovation programme under Project SHui, grant agreement No 773903. The authors wish to thank NaanDan Jain for the elevation survey in Mevo Beitar vineyard, to Carmel Winery for the vineyard contribution, and the grower, Avi Yehuda. The authors would also like to thank the grower, Shlomo Cohen, for collaborating and allowing the research to be conducted in his peach orchard and to the Northern R&D technical support team.

Appendix. Supplementary materials A and B

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2021.106262.

References

Adamchuk, V.I., Viscarra Rossel, R.A., Marx, D.B., Samal, A.K., 2011. Using targeted sampling to process multivariate soil sensing data. Geoderma 163 (1–2), 63–73. https://doi.org/10.1016/j.geoderma.2011.04.004.

Aqeel-Ur-Rehman, Abbasi, A.Z., Islam, N., Shaikh, Z.A., 2014. A review of wireless sensors and networks' applications in agriculture. Comput. Stand. Interfaces 36 (2), 263–270. https://doi.org/10.1016/j.csi.2011.03.004.

Bahat, I., Netzer, Y., Grünzweig, J.M., Alchanatis, V., Peeters, A., Goldshtein, E., Ohana-Levi, N., Ben-Gal, A., Cohen, Y., 2021. In-season interactions between vine vigor, water status and wine quality in terrain-based management-zones in a 'cabernet sauvignon' vineyard. Remote Sens. 13 (9), 1636. https://doi.org/10.3390/rs13091636.

Bazzi, C.L., Schenatto, K., Upadhyaya, S., Rojo, F., Kizer, E., Ko-Madden, C., 2019. Optimal placement of proximal sensors for precision irrigation in tree crops. Precis. Agric. 20 (4), 663–674. https://doi.org/10.1007/s11119-018-9604-3.

Becker, P., Rabenold, P.E., Idol, J.R., Smith, A.P., 1988. Water potential gradients for gaps and slopes in a Panamanian tropical moist forest's dry season. In: Journal of Tropical Ecology, Vol. 4. Cambridge University Press, pp. 173–184. https://doi.org/10.2307/2559656.

Bellvert, J., Zarco-Tejada, P.J., Marsal, J., Girona, J., González-Dugo, V., Fereres, E., 2016. Vineyard irrigation scheduling based on airborne thermal imagery and water potential thresholds. Aust. J. Grape Wine Res. 22 (2), 307–315. https://doi.org/10.1111/ajgw.12173.

Ben-Gal, A., Cohen, Y., Peeters, A., Naor, A., Netzer, Y., Ohana-Levi, N., Bahat, I., Katz, L., Shaked, B., Linker, R., Yulzary, S., & Alchanatis, V., n.d. Precision drip irrigation for horticulture. Acta Horticulturae, In Press.

Brungard, C.W., Boettinger, J.L., 2010. Conditioned Latin hypercube sampling: optimal sample size for digital soil mapping of arid rangelands in Utah, USA. In: Digital Soil Mapping. Springer, Netherlands, pp. 67–75. https://doi.org/10.1007/978-90-481-8863-5_6.

Cebeci, Z., 2019. Comparison of internal validity indices for fuzzy clustering. J. Agric. Inf. 10 (2). https://doi.org/10.17700/jai.2019.10.2.537.

Chang, J., Clay, D.E., Carlson, C.G., Clay, S.A., Malo, D.D., Berg, R., Kleinjan, J., Wiebold, W., 2003. Different techniques to identify management zones impact nitrogen and phosphorus sampling variability. Agron. J. 95 (6), 1550–1559. https://doi.org/10.2134/agronj2003.1550.

Clifford, D., Payne, J.E., Pringle, M.J., Searle, R., Butler, N., 2014. Pragmatic soil survey design using flexible Latin hypercube sampling. Comput. Geosci. 67, 62–68. https://doi.org/10.1016/j.cageo.2014.03.005.

Cohen, Y., Alchanatis, V., Saranga, Y., Rosenberg, O., Sela, E., Bosak, A., 2017. Mapping water status based on aerial thermal imagery: comparison of methodologies for upscaling from a single leaf to commercial fields. Precis. Agric. 18 (5), 801–822. https://doi.org/10.1007/s11119-016-9484-3.

Colaizzi, P.D., Barnes, E.M., Clarke, T.R., Choi, C.Y., Waller, P.M., 2003. Estimating soil moisture under low frequency surface irrigation using crop water stress index. J. Irrig. Drain. Eng. 129 (1), 27–35. https://doi.org/10.1061/(ASCE)0733-9437(2003)129:1(27).

Corwin, D.L., Lesch, S.M., 2013. Protocols and guidelines for field-scale measurement of soil salinity distribution with ECa-directed soil sampling. J. Environ. Eng. Geophys. 18 (1), 1–25. https://doi.org/10.2113/jeeg18.1.1.

Costa, M.M., de Queiroz, D.M., de Carvalho Pinto, F. de A., dos Reis, E.F., Santos, N.T., 2014. Moisture content effect in the relationship between apparent electrical conductivity and soil attributes. Acta Scientiarum - Agronomy, 36(4), 395–401. https://doi.org/10.4025/actasciagron.v36i4.18342.

De Lannoy, G.J.M., Verhoest, N.E.C., Houser, P.R., Gish, T.J., Van Meirvenne, M., 2006. Spatial and temporal characteristics of soil moisture in an intensively monitored agricultural field (OPE3). J. Hydrol. 331 (3–4), 719–730. https://doi.org/10.1016/j.jhydrol.2006.06.016.

Derumigny, A., Ohana-Levi, N., 2020. R package: MFunctMatching.

Dunteman, G.H., 1989. Principal Component Analysis. SAGE Publications Ltd.

ESRI Inc, 2017. ArcGIS Pro (Version 2.1.2). Environmental Systems Research Institute, Redlands, CA.

Fries, I., Ekbohm, G., Villumstad, E., 1984. Nosema apis, sampling techniques and honey yield. J. Apic. Res. 23 (2), 102–105. https://doi.org/10.1080/00218839.1984.11100617.

Gandah, M., Stein, A., Brouwer, J., Bouma, J., 2000. Dynamics of spatial variability of millet growth and yields at three sites in Niger, West Africa and implications for precision agriculture research. Agric. Syst. 63 (2), 123–140. https://doi.org/10.1016/S0308-521X(99)00076-1.

Gasic, K., Reighard, G.L., Windham, J., Ognjanov, M., 2015. Relationship between fruit maturity at harvest and fruit quality in peach. Acta Hortic. 1084, 643–648. https://doi.org/10.17660/ActaHortic.2015.1084.86.

Gavioli, A., de Souza, E.G., Bazzi, C.L., Schenatto, K., Betzek, N.M., 2019. Identification of management zones in precision agriculture: an evaluation of alternative cluster analysis methods. Biosyst. Eng. 181, 86–102. https://doi.org/10.1016/j.biosystemseng.2019.02.019.

Gibbs, A.L., Su, F.E., 2002. On choosing and bounding probability metrics. Int. Stat. Rev. 70 (3), 419–435. https://doi.org/10.1111/j.1751-5823.2002.tb00178.x.

Gibson, W.C., 2014. The Method of Moments in Electromagnetics, Second Ed. Chapman and Hall/CRC.

Goovaerts, P., Kerry, R., 2010. Using ancillary data to improve prediction of soil and crop attributes in precision agriculture. In: Oliver, M.A. (Ed.), Geostatistical Applications for Precision Agriculture. Springer, Netherlands, pp. 167–194. https://doi.org/10.1007/978-90-481-9133-8_7.

Griffith, D.A., 2004. Spatial autocorrelation. In: Encyclopedia of Social Measurement. Elsevier Inc., pp. 581–590. https://doi.org/10.1016/B0-12-369398-5/00334-0

Griffith, D.A., 2013. Establishing qualitative geographic sample size in the presence of spatial autocorrelation. Ann. Assoc. Am. Geogr. 103 (5), 1107–1122. https://doi.org/10.1080/00045608.2013.776884.

Haining, R., 2015a. Spatial Autocorrelation. In: International Encyclopedia of the Social & Behavioral Sciences, Second Edition. Elsevier Inc., pp. 105–110. https://doi.org/10.1016/B978-0-08-097086-8.72056-3

Haining, R., 2015b. Spatial sampling. In: International Encyclopedia of the Social & Behavioral Sciences, Second Edition. Elsevier Inc., pp. 185–190. https://doi.org/10.1016/B978-0-08-097086-8.72065-4

Hall, A.R., 2005. Generalized Method of Moments. Oxford University Press.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50 (4), 1029. https://doi.org/10.2307/1912775.

He, J., Jin, Y., Du, Y.-L., Wang, T., Turner, N.C., Yang, R.-P., Siddique, K.H.M., Li, F.-M., 2017. Genotypic variation in yield, yield components, root morphology and architecture, in soybean in relation to water and phosphorus supply. Front. Plant Sci. 8, 1499. https://doi.org/10.3389/fpls.2017.01499.

Heuvelink, G.B.M., van Egmond, F.M., 2010. Space-time geostatistics for precision agriculture: a case study of NDVI mapping for a Dutch Potato Field. In: Geostatistical Applications for Precision Agriculture. Springer, Netherlands, pp. 117–137. https://doi.org/10.1007/978-90-481-9133-8_5.

Isaacs, E.H., Srivastava, M., 1989. An Introduction to Applied Geostatistics. Oxford University Press.

Israel, G.D., 1992. Determining Sample Size.

Israeli, A., Litaor, M., Emmerich, M., Shir, O.M., 2019. Statistical learning in soil sampling design aided by Pareto optimization. In: GECCO 2019 - Proceedings of the 2019 Genetic and Evolutionary Computation Conference, pp. 1198–1205. https://doi.org/10.1145/3321707.3321809.

Johnson, C.K., Mortensen, D.A., Wienhold, B.J., Shanahan, J.F., Doran, J.W., 2003. Site-specific management zones based on soil electrical conductivity in a semiarid cropping system. Agron. J. 95 (2), 303–315. https://doi.org/10.2134/agronj2003.3030.

Johnson, L.F., 2003. Temporal stability of an NDVI-LAI relationship in a Napa Valley vineyard. Aust. J. Grape Wine Res. 9 (2), 96–101. https://doi.org/10.1111/j.1755-0238.2003.tb00258.x.

Jonckheere, I., Fleck, S., Nackaerts, K., Muys, B., Coppin, P., Weiss, M., Baret, F., 2004. Review of methods for in situ leaf area index determination: Part I. Theories, sensors and hemispherical photography. Agric. For. Meteorol. 121 (1–2), 19–35. https://doi.org/10.1016/J.AGRFORMET.2003.08.027.

Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. In: Computers and Electronics in Agriculture, Vol. 143. Elsevier B.V, pp. 23–37. https://doi.org/10.1016/j.compag.2017.09.037.

Kasimatis, A.N., Vilas, E.P., 1985. Sampling for degrees brix in vineyard plots. Am. J. Enol. Viticult. 36 (3).

Katz, L., Ben-Gal, A., Litaor, M., Naor, A., Peres, M., Bahat, I., Netzer, Y., Peeters, A., Alchanatis, V., Cohen, Y., n.d. A methodology to evaluate the impact of variable rate application: cases from a drip-irrigated peach orchard and a wine grape vineyard. Precis. Agric.

Kerry, R., Oliver, M.A., Frogbrook, Z.L., 2010. Sampling in precision agriculture. In: Geostatistical Applications for Precision Agriculture. Springer, Netherlands, pp. 35–63. https://doi.org/10.1007/978-90-481-9133-8_2.

Khanal, S., Fulton, J., Shearer, S., 2017. An overview of current and potential applications of thermal remote sensing in precision agriculture. Comput. Electron. Agric. 139, 22–32. https://doi.org/10.1016/J.COMPAG.2017.05.001.

Kim, J.Y., Glenn, D.M., 2015. Measurement of photosynthetic response to plant water stress using a multi-modal sensing system. Trans. ASABE 58 (2), 233–240. https://doi.org/10.13031/trans.58.10873.

Kitanidis, P.K., 1988. Prediction by the method of moments of transport in a heterogeneous formation. J. Hydrol. 102 (1–4), 453–473. https://doi.org/10.1016/0022-1694(88)90111-4.

Koenig, W.D., 1999. Spatial autocorrelation of ecological phenomena. In: Trends in Ecology and Evolution, Vol. 14. Elsevier Ltd., pp. 22–26. https://doi.org/10.1016/S0169-5347(98)01533-X. Issue 1.

Lee, D.-H., Park, J.-H., 2019. Comparison between NDVI and CWSI for waxy corn growth monitoring in field soil conditions. In: Neale, C.M., Maltese, A. (Eds.), Remote Sensing for Agriculture, Ecosystems, and Hydrology XXI, Vol. 11149. SPIE, p. 58. https://doi.org/10.1117/12.2533558.

López-Vicente, M., Calvo-Seas, E., Álvarez, S., Cerdà, A., 2020. Effectiveness of cover crops to reduce loss of soil organic matter in a rainfed vineyard. Land 9 (7), 230. https://doi.org/10.3390/land9070230.

Maechler, M., Rousseeuw, P.J., Struyf, A., Hubert, M., Hornik, K., 2019. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0. (R package version 2.1.0; p. Available online; https://svn.r-project.org/R-pack).

Marsaglia, G., Tsang, W.W., Wang, J., 2003. Evaluating Kolmogorov's distribution. J. Stat. Softw. 8 (1), 1–4. https://doi.org/10.18637/jss.v008.i18.

Menzel, C.M., Simpson, D.R., 1986. Plant water relations in lychee: Effects of solar radiation interception on leaf conductance and leaf water potential. Agric. For. Meteorol. 37 (4), 259–266. https://doi.org/10.1016/0168-1923(86)90064-X.

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. Precis. Agric. 11 (2), 148–162. https://doi.org/10.1007/s11119-009-9153-x.

Metcalfe, P., Beven, K., Freer, J., 2018. dynatopmodel: Implementation of the Dynamic TOPMODEL Hydrological Model. R package version 1.2.1 (version 1.2.1 from CRAN; p. Available online: https://CRAN.R-project.org/packa). https://rdrr.io/cran/dynatopmodel/.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009.

Minasny, B., McBratney, A.B., Walvoort, D.J.J., 2007. The variance quadtree algorithm: use for spatial sampling design. Comput. Geosci. 33 (3), 383–392. https://doi.org/10.1016/j.cageo.2006.08.009.

Mohanty, B.P., Skaggs, T.H., 2001. Spatio-temporal evolution and time-stable characteristics of soil moisture within remote sensing footprints with varying soil, slope, and vegetation. Adv. Water Resour. 24 (9–10), 1051–1067. https://doi.org/10.1016/S0309-1708(01)00034-3.

Monaghan, J.M., Daccache, A., Vickers, L.H., Hess, T.M., Weatherhead, E.K., Grove, I.G., Knox, J.W., 2013. More 'crop per drop': constraints and opportunities for precision irrigation in European agriculture. J. Sci. Food Agric. 93 (5), 977–980. https://doi.org/10.1002/jsfa.6051.

Moran, P.A., 1948. The interpretation of statistical maps. J. Roy. Stat. Soc.: Ser. B (Methodol.) 10 (2), 243–251. https://doi.org/10.1111/j.2517-6161.1948.tb00012.x.

Mühlenstädt, T., Kuhnt, S., 2011. Kernel interpolation. Comput. Stat. Data Anal. 55 (11), 2962–2974. https://doi.org/10.1016/J.CSDA.2011.05.001.

Mulder, V.L., de Bruin, S., Schaepman, M.E., 2013. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. Int. J. Appl. Earth Obs. Geoinf. 21 (1), 301–310. https://doi.org/10.1016/j.jag.2012.07.004.

Murolo, S., Mancini, V., Romanazzi, G., 2014. Spatial and temporal stolbur population structure in a cv. Chardonnay vineyard according to *vmp1* gene characterization. Plant. Pathol. 63 (3), 700–707. https://doi.org/10.1111/ppa.12122.

Nanda, A., Sen, S., McNamara, J.P., 2019. How spatiotemporal variation of soil moisture can explain hydrological connectivity of infiltration-excess dominated hillslope: observations from lesser Himalayan landscape. J. Hydrol. 579, 124146 https://doi.org/10.1016/j.jhydrol.2019.124146.

Nawar, S., Corstanje, R., Halcro, G., Mulla, D., Mouazen, A.M., 2017. Delineation of soil management zones for variable-rate fertilization: a review. In: Advances in Agronomy, Vol. 143. Academic Press Inc., pp. 175–245. https://doi.org/10.1016/bs.agron.2017.01.003

O'Shaughnessy, S.A., Evett, S.R., Colaizzi, P.D., Howell, T.A., 2011. Using radiation thermography and thermometry to evaluate crop water stress in soybean and cotton. Agric. Water Manag. 98 (10), 1523–1535. https://doi.org/10.1016/j.agwat.2011.05.005.

O'Toole, J.C., Moya, T.B., 1978. Genotypic variation in maintenance of leaf water potential in rice[1]. Crop Sci. 18 (5), 873–876. https://doi.org/10.2135/cropsci1978.0011183X001800050050x.

Ohana-Levi, N., Ben-Gal, A., Peeters, A., Termin, D., Linker, R., Baram, S., Raveh, E., Paz-Kagan, T., 2020. A comparison between spatial clustering models for determining N-fertilization management zones in orchards. Precis. Agric. 1–25 https://doi.org/10.1007/s11119-020-09731-5.

Ohana-Levi, N., Paz-Kagan, T., Panov, N., Peeters, A., Tsoar, A., Karnieli, A., 2020. Time series analysis of vegetation-cover response to environmental factors and residential development in a dryland region. GIScience and Remote Sensing 56 (3). https://doi.org/10.1080/15481603.2018.1519093.

Ohana-Levi, Noa, Bahat, I., Peeters, A., Shtein, A., Netzer, Y., Cohen, Y., Ben-Gal, A., 2019. A weighted multivariate spatial clustering model to determine irrigation management zones. Comput. Electron. Agric. 162, 719–731. https://doi.org/10.1016/J.COMPAG.2019.05.012.

Oliver, M.A., 2010. Geostatistical applications for precision agriculture. In: Geostatistical Applications for Precision Agriculture. Springer. https://doi.org/10.1007/978-90-481-9133-8.

Oliver, M.A., Webster, R., 1986. Semi-variograms for modelling the spatial pattern of landform and soil properties. Earth Surf. Proc. Land. 11 (5), 491–504. https://doi.org/10.1002/esp.3290110504.

Park, Y.L., Krell, R.K., Carroll, M., 2007. Theory, technology, and practice of site-specific insect pest management. J. Asia-Pac. Entomol. 10 (2), 89–101. https://doi.org/10.1016/S1226-8615(08)60337-4.

Pearson, K., 1894. Contributions to the mathematical theory of evolution. Philos. Trans. Roy. Soc. London (A.) 185, 71–110. https://doi.org/10.1098/rsta.1894.0003.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691. https://doi.org/10.1016/j.cageo.2004.03.012.

Petroselli, A., Vessella, F., Cavagnuolo, L., Piovesan, G., Schirone, B., 2013. Ecological behavior of Quercus suber and Quercus ilex inferred by topographic wetness index (TWI). Trees – Struct. Funct. 27 (5), 1201–1215. https://doi.org/10.1007/s00468-013-0869-x.

Prueger, J.H., Parry, C.K., Kustas, W.P., Alfieri, J.G., Alsina, M.M., Nieto, H., Wilson, T. G., Hipps, L.E., Anderson, M.C., Hatfield, J.L., Gao, F., McKee, L.G., McElrone, A., Agam, N., Los, S.A., 2019. Crop Water Stress Index of an irrigated vineyard in the Central Valley of California. Irrig. Sci. 37 (3), 297–313. https://doi.org/10.1007/s00271-018-0598-4.

R Core Team, 2020. R: A language and environment for statistical computing (p. Available online: https://www.R-project.org/). R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raskutti, B., Leckie, C., 1999. An evaluation of criteria for measuring the quality of clusters. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'99), pp. 905–910.

Richter, R., Reu, B., Wirth, C., Doktor, D., Vohland, M., 2016. The use of airborne hyperspectral data for tree species classification in a species-rich Central European forest area. Int. J. Appl. Earth Obs. Geoinf. 52, 464–474. https://doi.org/10.1016/j.jag.2016.07.018.

Roudier, P., 2011. clhs: a R package for conditioned Latin hypercube sampling. https://github.com/pierreroudier/clhs/.

Roudier, P., Tisseyre, B., Poilvé, H., Roger, J.M., 2008. Management zone delineation using a modified watershed algorithm. Precis. Agric. 9 (5), 233–250. https://doi.org/10.1007/s11119-008-9067-z.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20 (C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Rud, R, Cohen, Y., Alchanatis, V., Dar, Z., Levi, A., Brikman, R., Shenderey, C., Heuer, B., Markovits, T., Mulla, D., Rosen, C., 2013. The potential of CWSI based on thermal imagery for in-season irrigation management in potato fields. 721–727. https://doi.org/10.3920/978-90-8686-778-3_89.

Rud, Ronit, Cohen, Y., Alchanatis, V., Levi, A., Brikman, R., Shenderey, C., Heuer, B., Markovitch, T., Dar, Z., Rosen, C., Mulla, D., Nigon, T., 2014. Crop water stress index derived from multi-year ground and aerial thermal images as an indicator of potato water status. Precis. Agric. 15 (3), 273–289. https://doi.org/10.1007/s11119-014-9351-z.

Shinya, P., Contador, L., Predieri, S., Rubio, P., Infante, R., 2013. Peach ripening: segregation at harvest and postharvest flesh softening. Postharvest Biol. Technol. 86, 472–478. https://doi.org/10.1016/j.postharvbio.2013.07.038.

Sinton, T.H., Ough, C.S., Kissler, J.J., Kasimatis, A.N., 1978. Grape juice indicators for prediction of potential wine quality. I. Relationship between crop level, juice and wine composition, and wine sensory ratings and scores. Am. J. Enol. Viticult. 29 (4).

Stein, A., Ettema, C., 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. Agric. Ecosyst. Environ. 94 (1), 31–47. https://doi.org/10.1016/S0167-8809(02)00013-0.

Stéphane Dray, A., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., Wagner, H.H., 2020. Package "adespatial": Multivariate Multiscale Spatial Analysis (R package version 0.3-8). https://doi.org/10.1890/11-1183.1.

Visalini, K., Subathra, B., Srinivasan, S., Palmieri, G., Bekiroglu, K., Thiyaku, S., 2019. Sensor placement algorithm with range constraints for precision agriculture. IEEE Aerosp. Electron. Syst. Mag. 34 (6), 4–15. https://doi.org/10.1109/MAES.2019.2921177.

Wagner, H.H., 2003. Spatial covariance in plant communities: Integrating ordination, geostatistics, and variance testing. Ecology 84 (4), 1045–1057. https://doi.org/10.1890/0012-9658(2003)084[1045:SCIPCI]2.0.CO;2.

Wang, J.F., Zhang, T.L., Fu, B.J., 2016. A measure of spatial stratified heterogeneity. Ecol. Ind. 67, 250–256. https://doi.org/10.1016/j.ecolind.2016.02.052.

Werbylo, K.L., Niemann, J.D., 2014. Evaluation of sampling techniques to characterize topographically-dependent variability for soil moisture downscaling. J. Hydrol. 516, 304–316. https://doi.org/10.1016/j.jhydrol.2014.01.030.

Zhang, G., Liu, F., Song, X. dong., 2017. Recent progress and future prospect of digital soil mapping: A review. In: Journal of Integrative Agriculture, Vol. 16 Chinese Academy of Agricultural Sciences, Issue 12, pp. 2871–2885. https://doi.org/10.1016/S2095-3119(17)61762-3.

Zhang, N., Wang, M., Wang, N., 2002. Precision agriculture—a worldwide overview. Comput. Electron. Agric. 36 (2–3), 113–132. https://doi.org/10.1016/S0168-1699(02)00096-0.