Computers and Electronics in Agriculture 135 (2017) 163–174

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Assessing sampling designs for determining fertilizer practice from yield data

S.E. Muhammed [a], B.P. Marchant [b], R. Webster [a], A.P. Whitmore [a], G. Dailey [a], A.E. Milne [a,*]

[a] Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK
[b] British Geological Survey, Keyworth, Nottingham NG12 5GG, UK

ARTICLE INFO

Article history:
Received 18 December 2015
Received in revised form 2 February 2017
Accepted 5 February 2017
Available online 17 February 2017

Keywords:
Yield maps
Phosphorus
Sampling
Plant nutrients
Variograms

ABSTRACT

Many farmers sample their soil to measure the concentrations of plant nutrients, so as to decide how much fertilizer to apply. Now that fertilizer can be applied at variable rates farmers want to know whether maps of nutrient concentration made from grid samples or of field subdivisions (zones within their fields) are merited: do such maps lead to greater profit than would a single measurement on a bulked sample for each field when all costs are taken into account? We have examined the merits of grid-based and zone-based sampling strategies over single field-based averages using continuous spatial data on wheat yields at harvest in six fields in southern England and simulated concentrations of phosphorus (P) in the soil. We have taken into account current prices of wheat, P fertilizer and sampling and laboratory analysis. Variograms of yield provide guides for sampling. We show that where variograms have large variances and long effective ranges grid-sampling and mapping are feasible and have large probabilities of being cost-effective. Where effective ranges are short, sampling must be dense to reveal the spatial variation and be expensive, and variable-rate application of fertilizer is likely to be impracticable and almost certainly not cost-effective. We found zone-based sampling was less likely to be cost effective in a similar situation when the management zones were poorly correlated to P concentrations.

1. Introduction

We have known for more than 150 years that shortages of phosphorus and potassium in the soil limit crop growth. Thousands of experiments have been done to estimate the responses of crops to additions of these nutrients and to calculate the needs for fertilizers. Farmers now want to use these results to vary their applications within fields. Perhaps surprisingly, there are few reports linking variation in the concentrations of these elements in the soil to yields within individual fields on commercial farms. There are examples, however, where positive correlations were found for cereals (Frogbrook et al., 2002) and pastures (McCormick et al., 2009; Serrano et al., 2011). Many farmers in the United Kingdom sample their soil every four years to measure the nutrients, in particular phosphorus (P) and potassium (K), in the soil so as to decide how much fertilizer to apply to their crops. Sampling is often done at points on a 'W' shape across each field, and then individual samples are bulked before analysis in the laboratory (PDA, 2011). Even though this sampling configuration does not follow the principles of design-based statistics, it has been widely adopted by farmers as the results are generally no less accurate than those obtained from stratified random sampling (Marchant et al., 2012). By bulking the sample, however, all information on the variation of the nutrients across the field is lost, and so any local deficiency or excess is obscured. If a farmer wants to map the variation in nutrients across a field, so that fertilizer could be adjusted spatially, for example, then he or she should ideally sample the soil on a grid (with perhaps some additional points at closer spacings) and measure the nutrient content in each sample of soil separately (Mallarino and Wittry, 2004; Sawchik and Mallarino, 2007; Fu et al., 2013). Kriging, which makes best use of such data, could then be used to map the variation in nutrients (Kravchenko, 2003; Webster and Oliver, 2007). To krige, however, one needs an accurate estimate of the variogram or covariance function for the variable of interest, and for that at least 100 measurements are needed (Oliver and Webster, 2014). This creates a problem because measuring the concentration of P or K of each sample in the laboratory is costly.

Often the reason for the variation in yield across a field will be obvious to the farmer. For example, the farmer might know that a particular part of the field is prone to drought and that this is a major cause of the variation in yield. If the farmer suspects a local nutrient deficiency then it would make sense to divide the land

* Corresponding author.
E-mail address: alice.milne@rothamsted.ac.uk (A.E. Milne).

http://dx.doi.org/10.1016/j.compag.2017.02.002
0168-1699/

Nomenclature

$\mathbf{x} \equiv \{x_1, x_2\}$	spatial coordinates in two dimensions
y	yield of crop
$\tilde{y}$	standardized yield
y_r	realized yield
y_0	target yield
z	quantity of phosphorus, P
z^*	realization of z
z_{fert}	quantity of fertilizer P
z_{soil}	initial quantity of P in the soil
z_{total}	$z_{fert} + z_{soil}$
k	the number of classes in the k-means classification

variogram parameters

c_0	nugget variance
c_1	variance of spatially correlated structure
a	distance parameter

Costs

G_{wheat}	price of grain, assumed to be £150 t $^{-1}$
G_{fert}	price of P fertilizer, assumed to be £0.31 kg $^{-1}$
G_{sample}	cost of soil analysis, assumed to be £5 sample $^{-1}$

into management zones and estimate the nutrient status for each zone separately.

So, which of these three sampling approaches, commonly used by farmers and advisors, should be adopted in any particular situation to apply P and K fertilizer spatially? It is a question taxing agricultural advisers who want to advise farmers on best practice for within-field sampling for plant nutrients—see, for example, Oliver and Kerry (2013), Mylavarapu and Wonsok (2014) and Hawkins et al. (2016). The grid-based approach should result in the most accurate prediction of fertilizer requirement (Mallarino and Wittry, 2004; Sawchik and Mallarino, 2007). But the money saved by varying the application of the fertilizer locally within fields to match the requirement of the crop might be less than the cost of sampling and measurement (Fleming et al., 2000; Mallarino and Wittry, 2004). The balance of the two, and therefore the merit of the approach, depend on the magnitude and variation of the nutrient content of the soil (Sawchik and Mallarino, 2007). These two variables can be determined accurately only from measurements made on samples. The variation can, however, be assessed indirectly from crop yields. Many farmers are already monitoring yields as they harvest their crops, and variation in the data they record is in many instances a reflection of the variation in the availability of the nutrients in the soil (Stafford et al., 1999; Diker et al., 2004; Flowers et al., 2005). If the variation in yield is small then it is unlikely that the nutrients vary substantially. We know that factors other than nutrient supply can cause large variations in yield. Nevertheless, nutrient supply does dominate yield variation in many cases, and for present purpose we proceed on that assumption. In such circumstances Lark et al. (2003) proposed metrics based on the variogram of yield data (which captures the magnitude of variation of the yield) to assess the scope for variable rate management. Their approach was to use the metrics as factors in a decision tree designed to determine the potential for variable rate management.

In this context we aimed (i) to compare the merits of measuring plant nutrients by three sampling schemes (whole fields, zones within fields and grids) and (ii) to assess the extent to which yield maps might be used to determine the most cost-effective sampling. Which of the above sampling approaches is suitable for a given situation depends on a farmer's profit margin over the cost of fertilizer and soil sampling. It is not possible to do such a comparison in the field because the test requires perfect knowledge of how the nutrients vary across the field, we therefore resorted to simulation. In the approach presented here we simulated the variation in nutrients across fields from geostatistical models of the nutrients and used these to test the cost effectiveness of each sampling scheme. We modelled the associated yield variation for each realization and explored the use of the metrics of the yield variogram to decide which sampling strategy was likely to be most cost-effective.

2. Method

2.1. Data

We collated yield data, denoted y, from monitors on board combine harvesters and measurements of extractable P, denoted z, in the soil for six fields on a farm near Newbury, England (Table 1). Soils were medium to heavy textured with slight to moderate stoniness. The yields were of winter wheat from the seasons 2001 to 2011 recorded at approximately 20-m intervals by the monitor. The measurements of Olsen extractable P (sodium bicarbonate extract at pH 8.2) were made on a 100-m grid at 24–36 locations across each field. Fig. 1 displays yield maps for the six fields for a single year.

2.2. Determining management zones

Several methods have been devised for creating management zones (Oliver and Webster, 1989; Fleming et al., 2000; Diker et al., 2004; Flowers et al., 2005; Zhang et al., 2009). The multivariate technique of Dray et al. (2006) has most recently been applied successfully by Peralta et al. (2015) for wheat farming and by Córdoba et al. (2016) for grain cropping more generally.

We created management zones from the yield data using a spatially smoothed version of a fuzzy k-means classification devised by Lark (1998) (see also Milne et al., 2012). The data for the classification consisted of yields of wheat for p years at the n nodes of a square grid at intervals of 10 m. We denote the grid coordinates as $\mathbf{x} \equiv \{x_1, x_2\}$ and the yields in the p years as $y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_p(\mathbf{x})$.

From these data we created a classification by a 'hard' k-means algorithm. We standardized each of the $y_j, j = 1, 2, \ldots, y_p$ to zero mean and variance of 1, denoted $\tilde{y}_j$. For this method we choose k, the number of classes. We divided the whole set of the standardized data into that number of classes in such a way as to minimize the trace or determinant of the within-classes variance–covariance matrix. Each grid node then belonged to one and only one of the k classes, and in general it resembled other members of its class more than the members of the other classes.

For fuzzy k-means classification we first define a measure of dissimilarity, d, between an individual node i and a class q. A convenient measure is the Euclidean distance in the vector space:

$$d_{iq} = \sqrt{\sum_{j=1}^{p} \left(\tilde{y}_{ij} - \bar{\tilde{y}}_{jq}\right)^2}, \quad (1)$$

where $\tilde{y}_{ij}$ is the standardized yield at node i in the jth year, and $\bar{\tilde{y}}_{jq}$ is the mean of $\tilde{y}$ in class q in that year.