Field Crops Research 230 (2019) 31–45

Contents lists available at ScienceDirect

Field Crops Research

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

Establishing the precision and robustness of farmers' crop experiments

Ben Marchant[a,*], Sebastian Rudolph[a,b], Susie Roques[c], Daniel Kindred[c], Vincent Gillingham[d], Sue Welham[e], Colin Coleman[f], Roger Sylvester-Bradley[c]

[a] British Geological Survey, Keyworth, Nottinghamshire, NG12 5GG, UK
[b] Johann Heinrich von Thünen Institute, Institute of Rural Studies, Bundesallee 50, Braunschweig, Germany
[c] ADAS Boxworth, Cambridge, CB23 4NN, UK
[d] AgSpace, Dorcan Business Village, Swindon, SN3 5HY, UK
[e] VSN International Ltd., Software for Bioscientists, 2 Amberside, Wood Lane, Hemel Hempstead, Herts, HP2 4TP, UK
[f] Trials Equipment (UK) Ltd., Hudson's Hill, Hedingham Road, Wethersfield, Essex, Braintree, CM7 4EH, UK

ARTICLE INFO

Keywords:
Yield monitor
NDVI
Field-scale trials
Geostatistics

ABSTRACT

Precision farming technologies such as global positioning, input placement technologies and on-the-go yield monitoring now provide farmers with the means to conduct their own experiments at scales relevant to their decisions with minimal disruption. However, these experiments are generally incompatible with conventional statistical methods and alternative models of response variables (e.g. yield) must be estimated if the effect of the management decision is to be distinguished from other sources of variation. We explore the precision and robustness of such experiments using four sources of data and experimental designs of different degrees of complexity. We see that there is a trade-off between the precision of the experiment and its complexity and hence implementation cost. In yield experiments with small-grain cereals, standard errors of treatment effects in yield of less than 0.05 t/ha can potentially be achieved when the treatment is varied along the field traffic row and standard errors of less than 0.1 t/ha can potentially be achieved when single treatments are applied in each row but the experiment includes multiple disconnected repetitions of each treatment. Simpler split-field designs are less robust since it can be difficult to distinguish treatment effects from independent spatial trends and discontinuities in the response variable. In some instances, the potential precision is not realised because the data include noise or artefacts that are unrelated to crop performance. Further yield sensor developments are required to minimise these occurrences. The model-based statistical analyses of these experiments require assumptions regarding the variation of the response variable. We see that when these assumptions are inappropriate (e.g. if the correlation between response variable measurements is poorly modelled) then the inferences from the experiments can be unreliable. In particular, we see that the spatial correlation amongst yield measurements tends to be greater along the farm traffic row than perpendicular to it. Standard isotropic models of spatial correlation do not accommodate this feature and led to substantial under-estimation of the standard errors.

1. Introduction

Learning and progress in farming have, for centuries, occurred through sharing of experiences arising from 'trial and error' across multiple farms within a locality (Sylvester-Bradley, 1991). Through the last century, as farming became more sophisticated and farms became larger and less numerous, farmers have become increasingly reliant on more formal small-plot experiments and theories ('recommendations' e.g. AHDB, 2017) provided by their suppliers or advisors, or by the broader agricultural science community. However, the extensive extrapolations required in deriving broad-scale recommendations from small-plot experiments incur large uncertainties, particularly concerning the influence of soil variation, but also the effects of weather, farming system, and farming skill, which are all difficult to mimic and control at the small-plot scale. Thus, farmers are increasingly questioning whether these recommendations are relevant to their own farms or whether local factors might confound the experimental results (e.g. Pannell et al., 2006) and farmers are increasingly making field-scale comparisons of their own husbandry options.

In studies of within-field variation, Kindred et al. (2015) found that the optimal nitrogen fertiliser rate could vary by more than 100 kg/ha within a 4 ha wheat experimental area and Bramley et al. (2005) demonstrated how the choice of location for a viticulture plot trial could markedly impact the results. There is a strong case for farm-led

* Corresponding author.
E-mail address: benmarch@bgs.ac.uk (B. Marchant).

https://doi.org/10.1016/j.fcr.2018.10.006
Received 2 November 2017; Received in revised form 9 October 2018; Accepted 11 October 2018

experimentation at scales relevant to commercial decision-making so that bespoke guidelines specific to farm, field or management zone can be determined (Griffin et al., 2008). Indeed, MacMillan and Benton (2014) advocate a move away from the 'one-size-fits-all' approach of centralized research and a move towards research at a more local-scale, conducted by farmers. They describe case studies where networks of farmers share their experimental findings to determine more locally-specific management recommendations.

The implementation and analysis of agricultural experiments are not without challenges. Conventional small-plot experiments generally include a number of different treatments or management interventions that are allocated to different plots according to a randomized design (Fisher and Wishart, 1930; Little and Hills, 1978; The University of Reading, 2000). The same treatments are replicated on multiple plots so that classical statistical methods can be used to partition the variability of the response variable(s) and to draw reliable inferences. This approach requires few assumptions about the variation of the response. However, the work of applying treatments and collecting response data is laborious, requires specialist staff and equipment and interferes with the standard work practices of farms. Collecting yield data can be particularly disruptive when farmers are under pressure to complete the harvest in limited windows of appropriate weather conditions (Griffin et al., 2014).

Farmers' crop experiments tend to follow simpler, more systematic designs (Hicks at al., 1997). These designs are incompatible with the classical statistical methods so more elaborate statistical models of the variation of the response variable are required (Brus and De Gruijter, 1997), which entail specific assumptions about the variation of the response (Webster and Oliver, 2007). Treatments are applied using the existing farm equipment and yield measurements are often made using yield monitors built into the harvester (Grisso et al., 2009). This can lead to noisier or more variable experimental data and less certain conclusions. New skills are also required of the interested parties (farmers, advisors, suppliers, or scientists), to design the experiments, and collect, pre-process, analyse, interpret and store the experimental data. The design, implementation, analysis and interpretation of farmers' crop experiments generally require a collaboration between these interested parties who we collectively refer to as 'the experimenters'.

These challenges are being addressed through various research projects and initiatives in the UK (Sylvester-Bradley et al., 2017) and elsewhere (Cook et al., 2013). For example, farmers' networks and workshops are being organised to teach the required skills and facilitate the sharing of experiences and results. Other projects are concerned with determining the precision of farmers' crop experiments and establishing best practice. There is a trade-off between the complexity of the experimental design and the precision of the results (Whelan et al., 2012). For example, researchers have successfully used chess- or checker-board experiments (Pringle et al., 2004) to quantify the response of crop yields to different nitrogen fertiliser rates and to explore the spatial variation of these responses (Kindred et al., 2015). However, farmers often cannot accept the disruption that arises from accommodating such large and intricate experiments (Griffin et al., 2014). They generally prefer simpler designs where, for example, non-standard treatments are applied to a small number of strips across the field (Hicks et al., 1997). Alternatively, different uniform treatments are applied to large contiguous regions of a field. Also, whilst researchers would be likely to recognise and account for known spatial trends when designing field experiments, farmers are less certain of how best to optimise plot layouts, and require advice regarding the interpretation of the data and the conclusions that can be drawn. If farmers' crop experiments are to help drive farming progress it will be important to establish the most appropriate level of complexity for an experiment so that reliable management recommendations can be developed with minimal disruption to the farm.

Farmers' crop experiments are generally used to estimate the expected yield across the field under each treatment (e.g. Hicks et al., 1997) in order to test decisions and innovations that might be adopted on the farm. However, in some circumstances (e.g. when there are zones of distinct soil types in the field) it can be cost effective to vary the management plan within the field. In these circumstances the experiment must provide information about the within-field variation of the effectiveness of each treatment. When a chess-board experimental design is applied then multivariate kriging methods can be used to map the variation in the treatment effect (e.g. Kindred et al., 2015). Lawes and Bramley (2012) considered how simpler designs could be used to learn about such within-field variation. In an otherwise uniformly treated field, Lawes and Bramley (2012) positioned a single strip with a luxury N treatment such that it traversed management zones. They then looked at how the difference in yield between the treated strip and the adjacent control strip varied along the strip and determined the statistical significance of the differences through a series of t-tests. Rudolph et al. (2016) modified this approach to include a parametric model of the correlation amongst the measurements of the response variable. In an analysis of a nitrogen fertiliser strip trial they were able to identify significant jumps in a vegetation index at the boundary between different treatments.

The objectives of this study are to critically review the effectiveness of farmers' crop experiments with reference to examples from UK farms. We consider four types of sensor measurements of crop performance. These are yield monitors on commercial combine harvesters, a yield monitor on a plot combine harvester, Normalized Difference Vegetation Index (NDVI) measurements from an airborne multispectral sensor and NDVI measurements from a handheld sensor. We consider the practical limitations of each sensor, including the presence of artefacts within the data they produce, and the degree to which they can be overcome. We develop and describe a statistical framework for the analysis of farmers' experiments and use this to determine the precision and robustness of the exemplar experiments.

2. Methods

2.1. Overview

We explore the issues of farmers' experimentation with reference to six experiments on UK arable fields. The crop performance sensors, especially the yield monitors, tend to lead to noisy datasets that include artefacts and variation that is unrelated to the actual yield variation. We develop strategies that can be used to remove such artefacts and the statistical methods required to analyse the data. If classical statistical methods were applied the uncertainty associated with the results would be underestimated since these methods do not account for the spatial correlation amongst the data (Griffin, 2010). We describe model-based statistical methods (Diggle and Ribeiro, 2007) that can account for this correlation and determine the confidence with which we can interpret the results. In the case of data from yield monitors there tends to be stronger correlation in the longitudinal (along the swath) than the lateral direction. A product sum model (De Cesare et al., 2001), more commonly used to represent correlation in space and time, is used to account for this. After comparing statistical methodologies, we determine the precision that is achieved from different experimental designs.

2.2. Data types and pre-processing

2.2.1. Grain yield data from commercial harvesters

The sensor data which are most pertinent to farmers are georeferenced yield data, as measured for cereal crops by yield monitors fitted to commercial combine harvesters. We intended the yield data collected for this study to reflect the quality of data that might be expected from UK farmers' crop experiments. Therefore, data were received directly from the farmers concerned. No special requests were made to the

farmers regarding the calibration or operation of their combine harvesters. The combine harvesters were all fitted with a global positioning system (GPS) to record its location and to enable conversion of measurements of the grain flow or volume per unit time to a mass of harvested grain per unit area.

There are many processes or practices that can lead to noise appearing in maps of yield measurements. These may be inherent for any particular monitoring system or arise through differences in their operation. Errors introduced by the GPS can distort the area over which the yield is averaged and lead to the locations of yield measurements being poorly assigned. Other errors relate to poor calibration of the yield monitor or the operation of the combine harvester. Yield values will be erroneously small where the harvested swath width is less than the combine harvester header width (unless the operator and/or the yield monitoring system has automatically corrected or even over-corrected the width and yield). Artefacts can appear in the yield maps if the velocity of the combine harvester varies. Particularly extreme artefacts occur if the combine harvester stops since in this circumstance a flow of grain might still be recorded but the GPS signal indicates that an area of 0 m^2 is being harvested. At the start of each combine harvester run it can take a number of seconds (15 according to Grisso et al., 2009) until grain reaches the tank inside the combine harvester and a stable yield is registered. Similar delays continue to occur as the combine harvester progresses along the swath leading to errors in the locations at which yield measurements are assigned. Grain mixing within the harvesting mechanism also causes smoothing of any abrupt changes in yield along the swath. When adjacent runs of the combine harvester occur in opposite directions these location errors can lead to distortions to features in the yield maps that cross multiple passes.

Farmers and their advisors are generally aware of these potential artefacts in yield maps (Griffin, 2010). Different filters have been designed to identify and remove erroneous measurements (e.g. Sudduth and Drummond, 2007; Sun et al., 2012). We established a similar set of filters to remove artefacts from the yield monitor data collected in this study. These filters removed measurements:

1 flagged as headland or as erroneous by the combine harvester operator or experimenters.
2 within 15 m of a change in experimental treatment within a row
3 where the combine harvester velocity was outside of the normal operating range.
4 where the combine harvester was not moving in the expected direction. All of the experiments in this paper relate to parallel farm traffic rows.
5 at locations where the combine harvester has already cut the grain. This filter also serves to identify where rows are less than the combine harvester width apart and therefore the swath is not a full header width.
6 within a specified number of seconds of when the GPS time signal indicated that there was a break in the recorded data. Such a break might have resulted from one of the first four filters. This filter also served to remove data from the start and end of each row.
7 from rows where the average yield was less than expected.
8 which were outside of the expected yield values.
9 that are extreme relative to their neighbours. The method for identifying such local outliers is described by Marchant et al. (2010).

Many of these filters required specified thresholds (e.g. the ranges of combine harvester velocity or yield values that could be expected in normal operating conditions). These thresholds varied from field to field. They were chosen to remove measurements that the experimenter believed to be implausible and were based on knowledge of the experiments and fields concerned and then adjusted by trial-and-error to produce realistic yield maps. Ideally the majority of removed measurements would be as a result of the first six filters where the cause of the artefacts can be explained in terms of the operation of the combine harvester. However, extreme values which can potentially lead to misleading experimental results can still remain after the application of these filters.

The final step in the data pre-processing procedure was to apply a time-shift to the data to compensate for the time-delay in grain reaching the yield monitor sensor. A series of time shifts ranging between minus 30 and plus 30 s was applied to the data. The average correlation between proximal measurements from adjacent rows harvested in opposing directions was then calculated. The time-shift which led to the largest correlation was applied to the data. Both positive and negative time shifts were permitted because the yield monitor software might have already applied an overly severe time delay to the data. Further details of this time delay correction are provided by Muhammed et al. (2017).

2.2.2. Yield monitor data from a plot combine harvester

A combine harvester yield monitor typically records the crop yield every few seconds. Each datum does not relate to the yield at a precise location but instead relates to the average yield across a small region traversed by the combine harvester during this time interval. We refer to this region as the spatial 'footprint' of the measurement. One might assume that the footprint is a rectangle with width equal to the swath width and length equal to the distance the combine harvester has moved within the time interval. In reality, the footprint is likely to correspond to a shape with curved ends since grain cut by the edge of the header might take longer to reach the yield sensor than grain that is cut at the middle of the header (Lark et al., 1997; Whelan and McBratney, 2002). In either case, the spatial precision of the yield monitor data is limited by the width of the combine harvester. Commercial combine harvesters in the UK typically have header widths of at least 9 m and one method to increase the precision of a farmers' crop experiment is to reduce this width.

A Sampo 2010 plot combine harvester with a header width of 2.1 m was therefore modified by Trials Equipment Limited (www.trialseq.co.uk) for continuous yield monitoring. Two hoppers were attached to the side of the combine harvester (Fig. 1). After the grain had been cut, it was transported by a conveyor-belt to one of these hoppers. Each hopper was supported by calibrated load cells which recorded the mass of grain within the hopper 10 times every second. Each hopper had a capacity corresponding to around 10 kg of grain; once a hopper was full, a flap switched the supply of grain to the second hopper and the grain in the first hopper was emptied into the main tank of the plot combine harvester. The time series of recorded weights from each hopper were merged to produce one continuous signal corresponding to the hopper that was receiving grain at any instant. The difference in successive weight measurements was then calculated and this

Fig. 1. The Sampo 2010 plot combine harvester adapted for the continuous collection of harvested grain weights, locations and hence grain yields.

corresponded to the mass of grain entering a hopper within the 0.1 s time period. A GPS with 1 Hz uptake rate was also fitted to the plot combine harvester. As with commercial combine harvester yield monitors, the GPS was used to determine the area covered during each time interval and the location from which the grain had been cut.

The plot combine harvester typically operates at a speed of around 1 km/h. Thus, in 0.1 s it moves around 0.03 m. One might therefore hope that it could produce yield measurements with a spatial footprint of $2.1\text{m} \times 0.03\text{m} = 0.063\ \text{m}^2$. However, these raw yield measurements were very noisy. Much of this noise was mechanical and resulted from the large vibrations as the combine harvester moved and threshed the grain. Further noise resulted from the GPS being insufficiently precise to measure the few centimetres traversed in 0.1 s. Therefore, it was necessary to smooth the data and the average yield was calculated for 3 m long blocks. Thus, each smoothed measurement corresponded to a footprint of $6.3\ \text{m}^2$.

This smoothed yield data was then pre-processed using the series of filters applied to the yield monitor data from the commercial combine harvester.

2.2.3. Normalized difference vegetation index data

The NDVI relates to the greenness or ground cover of a crop and is calculated from multispectral reflectance data. The NDVI is the ratio of the reflectance in the near-infrared region of the measured spectrum minus the reflectance in the visible region, and the sum of the reflectance from these two regions. Thus, it varies between -1 and 1, with a NDVI of 1 corresponding to a green crop with full ground-cover.

The NDVI does not relate directly to the final yield and hence to the profitability of the crop. However, the NDVI can be used during the growing season to assess the fertiliser requirements of the crop. It can also be used as an indicator of whether an experimental management intervention has increased or decreased progress towards full ground cover.

We consider two sources of multispectral and hence NDVI data. One source is a handheld sensor that measures the average NDVI for an area of about $0.2–0.5\ \text{m}^2$ beneath the sensor. The exact area varies according to the height of the crop. The second is an airborne multispectral sensor which can be used to measure the NDVI across a field with a pixel size of < 10 cm. This airborne NDVI image was obtained when there was no cloud cover.

Many of the difficulties in obtaining spatially accurate yield data do not apply to NDVI data, therefore, there is no need to apply the extensive series of yield monitor filters described above. However, artefacts can occur in NDVI images, particularly those obtained from airborne sensors due to factors such as the observation angle, the speed of the vehicle, the angle of solar incidence and within field variation of slope, humidity and canopy structure. The majority of such artefacts are not easily identifiable within an image and therefore would add to the unexplained variation in the data and the uncertainty of the final results of experiments. Artefacts in uncropped areas can generally be identified. The handheld measurements were limited to areas where there was good crop cover. The airborne image did include uncropped areas (e.g. the farm traffic wheelings), so these pixels were removed from the image by setting a lower threshold on the NDVI measurements.

2.2.4. Experimental data

We consider data from farmers' crop experiments at six sites in England. Our focus is on the precision of these experiments rather than their findings. We are not concerned with the relative merits of the different experimental treatments and the specific local conditions that might influence their effectiveness. With this in mind, and to protect the anonymity of the farms involved, we only give the most general information about the experimental protocols, treatments and sites (Table 1 and Figs. 2–11). The quoted approximate site areas refer to the areas from which data were gathered and do not include headlands but do include other portions of data that were removed in the pre-processing steps. In each case, the data relate to a winter wheat crop. The experiment at Site D formed part of the 'LearN' project (Kindred and Sylvester-Bradley, 2014).

3. Statistical theory

The aim of each of the experiments was to determine whether the experimental treatments had an effect (positive or negative) on the response variable. In the description that follows we refer to the response variable as yield although we note that other variables, such a NDVI, might be used. To simplify our discussion of the effectiveness and precision of the experiments we designate one treatment as the standard or control treatment, numbered treatment 1 in the equations that follow. In experiments with a nil treatment (e.g. the fungicide and phosphorus-input trials) we designate this as the standard treatment. In the other experiments we designate the treatment covering the largest area as the standard. We then ask whether any of the other treatments have an effect relative to this standard.

The major challenge in answering this question is in separating the underlying variation in crop yield from that caused by the experimental treatments. In standard field trials, each treatment is replicated and allocated at random to different plots within the experimental area and classical or design-based statistical methods can be used to determine the probability that any difference in the measured yields under each treatment could have occurred by chance (The University of Reading, 2000). The underlying variation in the field is assessed by quantifying the variability between yields measured in plots under the same treatment.

In contrast, farmers' crop experiments tend to follow a more systematic design which is inconsistent with the design-based statistical methods (Brus and De Gruijter, 1997). Also, there might be too few plots to assess the underlying yield variability and this variability is likely to be spatially correlated. Thus there might be large and contiguous portions of the experimental area where the underlying yield is larger than the norm. If we conduct a split-field design and record the average yield for each half of the field, we have insufficient evidence to assess whether the observed difference between the two portions of the field has occurred because of the underlying variation or because of the different treatment.

We therefore need to use a model-based statistical approach (Diggle and Ribeiro, 2007). Here the variation and correlation between the individual yield measurements are quantified in a statistical model. We use a linear mixed model (LMM; Lark et al., 2006) that divides the spatial variation of yield into fixed and random effects. The fixed effects correspond to the treatment effects. The random effects correspond to the underlying variation or the residual variation that has not been explained by the fixed effects. Once we have estimated such a model we can use it to determine whether any observed difference between yields under different treatments could be explained by the underlying yield variation in the field.

The LMM is written:

$$\boldsymbol{z} = \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{z}$ is the length n vector of response variable measurements, n is the number of measurements, $\mathbf{M}$ is the size $n \times t$ fixed effects design matrix, $\boldsymbol{\beta}$ is a length t vector of fixed effects coefficients, $\boldsymbol{\varepsilon}$ is the length n vector of random effects and t is the number of different experimental treatments. The product $\mathbf{M}\boldsymbol{\beta}$ constitutes the fixed effects. All of the entries of the first column of $\mathbf{M}$ are equal to 1. For the remaining columns, the i th entry of the j th column of $\mathbf{M}$ is 1 if the i th yield observation underwent treatment. Otherwise the entries of these columns are 0. Thus β_1, the first entry of $\boldsymbol{\beta}$, corresponds to the mean of the fixed effects for the standard treatment and the other β_j correspond to the adjustment to this mean for each of the other treatments.

The elements of $\boldsymbol{\varepsilon}$ are generally assumed to be realized from a second order stationary Gaussian random function with zero mean and

Table 1
Details of the six experimental sites. Sensors include commercial combine yield monitor (CCYM) and plot combine yield monitor (PCYM).

Field	Area (ha)	Experiment type	Design	Data sources
A	16	Variety	Split-field	CCYM and airborne NDVI
B	11	P fertiliser	Alternate rows receive zero P; other rows receive one of four P rates.	CCYM
C	8	Fungicide	Three fungicide treatments allocated at random within six replicate blocks. Includes treatment variation within-rows	CCYM
D	9	N fertiliser rate	Standard N rate applied for majority of field. Standard rate plus 60 kg/ha and standard minus 60 kg/ha applied to sets of adjacent rows.	CCYM
E	0.4	Fungicide	Random allocation of three treatments within six replicate blocks	PCYM and handheld NDVI
F	0.3	Fungicide	Random allocation of three treatments to three replicate blocks	PCYM and handheld NDVI

covariance matrix $\mathbf{\Sigma}$. The random effects can be spatially correlated. We assume that the covariance for any two elements of the vector of random effects is a function of the distance between the locations of the two yield measurements. Many suitable covariance functions exist but we focus on the flexible nugget and Matérn function (Webster and Oliver, 2007):

$$c(h) = \begin{cases} c_0 + c_1 \text{ if } h = 0 \\ c_1 G(h) \text{ for } h > 0 \end{cases}, \quad (2)$$

where

$$G(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\sqrt{\nu}h}{a}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}h}{a}\right). \quad (3)$$

Here, Γ is the Gamma function and K_{ν} is a modified Bessel function of the second kind of order ν and h is the distance separating the two measurements. This function has four parameters which must be estimated from the data. These are the nugget variance c_0, the sill variance c_1, the distance parameter a and the smoothness parameter ν. The covariance function is often expressed as a variogram:

$$\gamma(h) = c_0 + c_1 - c(h). \quad (4)$$

There are two commonly used approaches to estimate the parameters. The method of moments (Webster and Oliver, 2007) allocates pairs of observations to a series of lag bins based upon their separation distance. The average semi-variance for the pairs in each bin is then calculated and the covariance function is fitted to these averages, generally using a least squares estimator. The method is undemanding to compute and can therefore be applied quickly. It does however require the practitioner to make a series of subjective decisions about how the data are arranged into different lag bins. We favour the maximum likelihood estimator (Diggle and Ribeiro, 2007; Lark et al., 2006) which

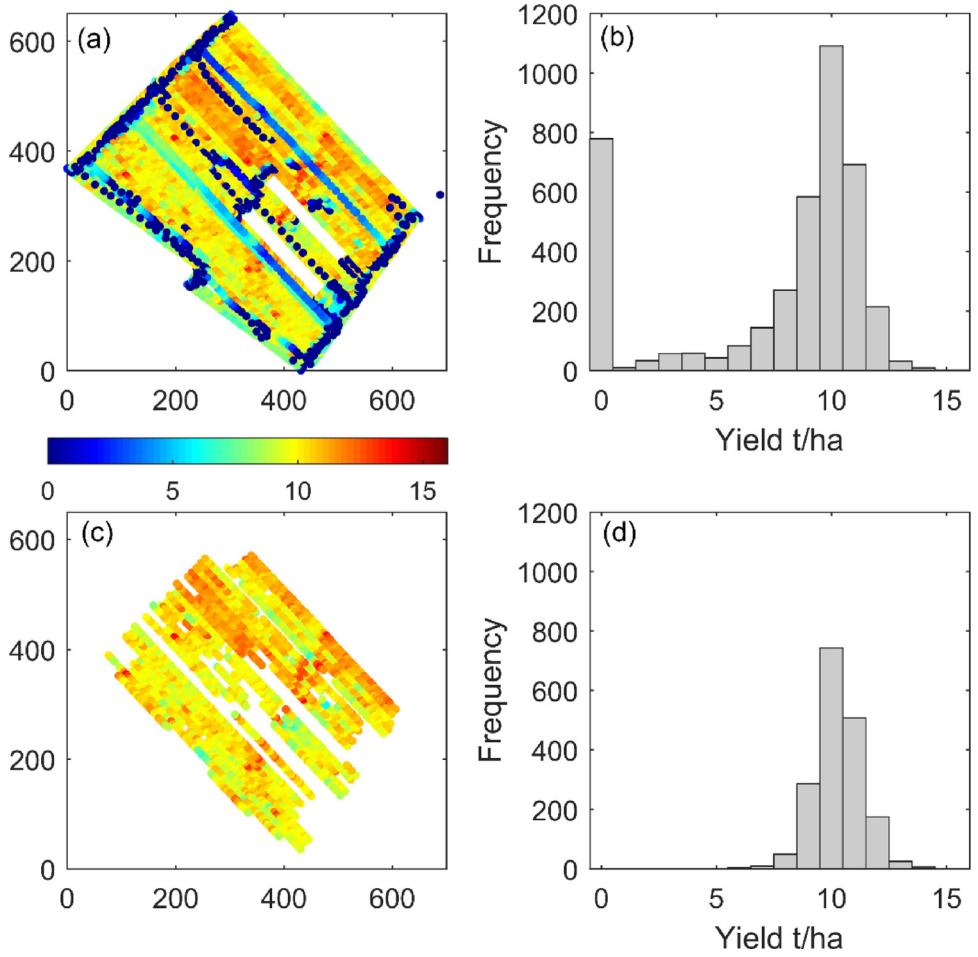

Fig. 2. (a) scatter plot of raw commercial yield monitor data from Experiment 1 (t/ha); (b) histogram of raw commercial yield monitor data from Experiment 1; (c) scatter plot of cleaned commercial yield monitor data from Experiment 1; (d) histogram of cleaned commercial yield monitor data from Experiment 1.

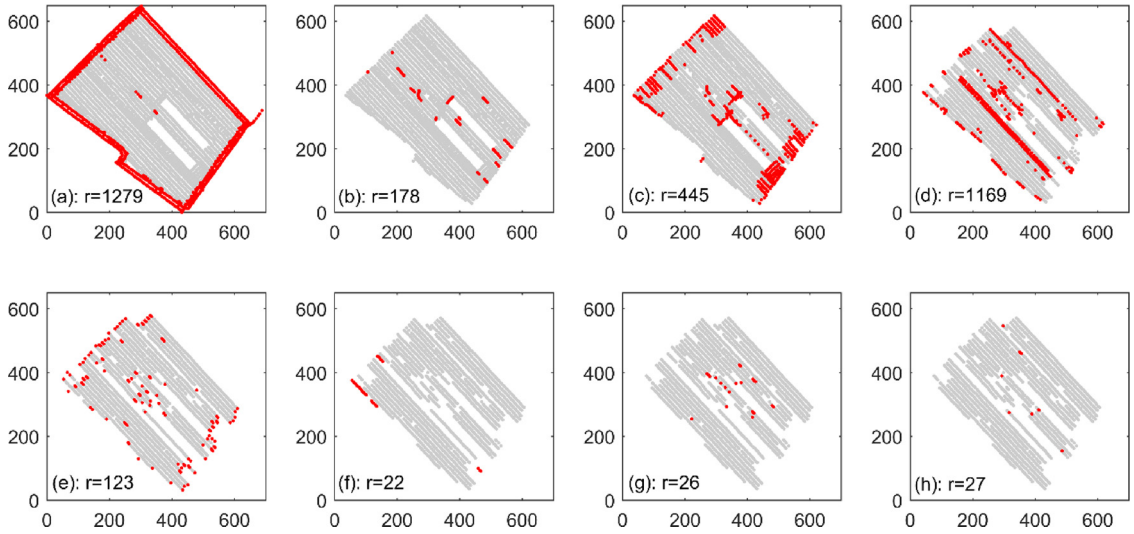

Fig. 3. Removed yield monitor observations (red dots) and remaining observations (grey dots) by the following filters (a) flagged by operator, (b) combine harvester velocity, (c) direction of travel, (d) already harvested location, (e) break in signal, (f) yield by line, (g) global threshold on yield and (h) local yield outliers. 'r' value indicates number of observations removed and the total number of observations was 4524. Coordinates are in m.

is considerably more computationally demanding and time-consuming but fully accounts for the spatial configuration of the data without requiring subjective choices. This estimator finds the parameter values which lead to the largest achievable value of L, the likelihood or probability that the observed data would be realised from the statistical model. Once these random effects parameters have been estimated it is possible to calculate the correlation between the random effects for any pair of locations and hence the size $n \times n$ covariance matrix $\mathbf{\Sigma}$.

Then the random effects parameters can be estimated using the formula:

$$\boldsymbol{\beta} = (\mathbf{M}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{z} \tag{5}$$

and the covariance matrix of these estimates is:

$$\mathbf{C} = (\mathbf{M}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{M})^{-1} \tag{6}$$

In summary, the expected effect of treatment $j = 2, \ldots, t$ is equal to β_j and this estimate has standard error $\sigma_j = \sqrt{C_{jj}}$ where C_{jj} is entry j, j of matrix $\mathbf{C}$. If we assume that these β_j are realised from a Gaussian distribution then this is sufficient information to calculate the probability density function of the treatment effect and the probability that the β_j are positive or larger than a threshold that would indicate it is worthwhile to adopt treatment j rather than the standard treatment.

The statistical significance of the treatment effect can be assessed using the z-score:

$$\zeta = \frac{\beta_j}{\sigma_j}, \tag{7}$$

to determine the probability that such an extreme estimate of the treatment effect might have occurred by chance. If there is no treatment effect then we would expect ζ to be realized from a random variable with a Gaussian distribution, zero mean and unit variance. We can find the probability that an extreme treatment effect would be estimated in the absence of a treatment effect by finding the probability that such an extreme ζ would be realised from a standardised Gaussian distribution. For example, if our null hypothesis is that the treatment has no effect on yield (i.e. if we conduct a two-sided test) then a $|\zeta| > 1.96$ would indicate that the probability of the observed data occurring under this hypothesis is less than 0.05. Similarly, if our null hypothesis is that the treatment effect is zero or negative (i.e. a one-sided test) then a $|\zeta| > 1.64$ would indicate that the probability of the observed data occurring under this hypothesis is less than 0.05. In this paper, we are primarily interested in the precision of the different experiments and the magnitude of the treatment effect which we can expect to identify. We use the σ_j as measures of the precision of the experiment. Statistical power is defined as the probability that the null hypothesis will be (correctly) rejected when it is false. It depends on the size of the treatment effect that is present. For a one-sided test, a statistical power of 0.8 can be achieved for a treatment effect of size $2.48\sigma_j$.

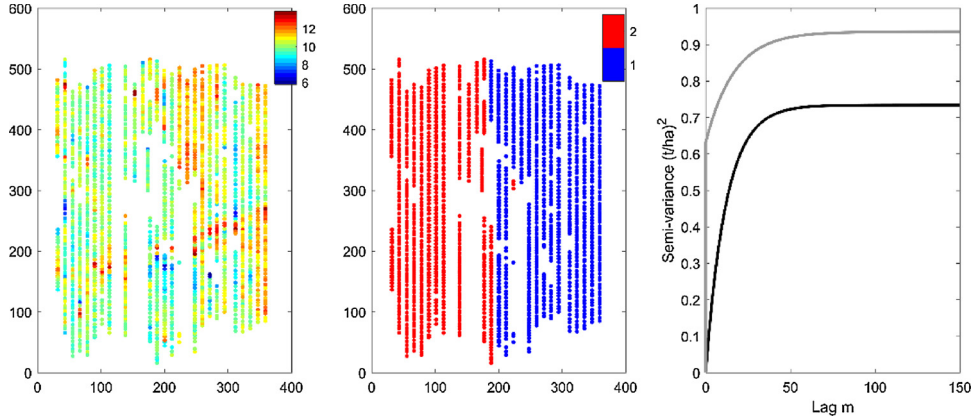

Fig. 4. (left) cleaned and rotated commercial yield monitor data from Experiment 1 (t ha^{-1}); (centre) experimental treatments; (right) estimated variograms in direction of travel (black) and perpendicular to direction of travel (grey). Coordinates are in m.

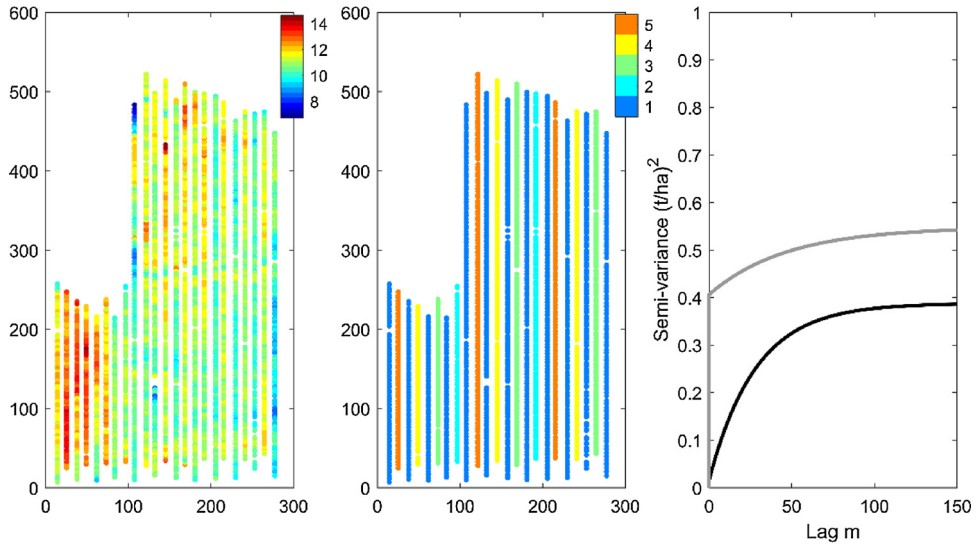

Fig. 5. (left) cleaned and rotated commercial yield monitor data from Experiment 2 (t ha^{-1}); (centre) experimental treatments; (right) estimated variograms in direction of travel (black) and perpendicular to direction of travel (grey). Coordinates are in m.

The estimated β_j and their uncertainty can be used to test the effects of decisions and treatments across the field. Rudolph et al. (2016) used a different statistical test, referred to as spatial discontinuity analysis, to conduct more localised studies of the effects of the different treatments. They particularly considered the boundary between the portion of the field under the standard treatment and that under treatment j for each. They used the estimated LMM and the observations from where the standard treatment was applied to predict the expected yield at the sites of observations under treatment j that were immediately adjacent to the boundary. They used the LU simulation method (Webster and Oliver, 2007) to simulate 1000 realisations of the yield that could have occurred at these sites if the standard treatment been applied. This approach ensured that in each realisation, the simulated values at each site were correlated according to the LMM. They then compared the observed value at each site to the set of simulated values. If the observed value was greater than 95% of the simulated values then this would indicate that the positive treatment effect was significant at the $p = 0.05$ level. When Rudolph et al. (2016) applied such a series of tests to the individual pixels adjacent to a treatment boundary in an image of a vegetation index for a field undergoing a nitrogen fertiliser trial, they did not see any significant treatment effects. This reflected the considerable noise in the image. However, when they compared the average simulated yields in blocks of multiple pixels along the boundary and compared these to the average observed values for these pixels, then a treatment effect was evident.

In some circumstances, additional terms must be included in the fixed effects to achieve reliable results. The use of the LMM implies that the random effects are stationary (Webster and Oliver, 2007). This means that the same mean and spatial covariance function are applicable throughout the field. This assumption is inconsistent with a trend in the underlying yield and, if such a trend is present, it must be included in the fixed effects. We would hope that the inclusion of the additional terms in the fixed effects will reduce the unexplained variation in the model and hence improve the precision of the experiment. However, it is also possible that the underlying trend coincides with and confounds the treatment effects. For example, consider the split-field design. It is possible that the same partition of the field was used for an experiment in a previous year. In this circumstance, it would be almost impossible to distinguish between the effect of this year's experiment and any continuing effect of the experimental treatment from the previous year. When a term is added to the fixed effects that acknowledges the potential for such a continuing effect we would therefore expect the standard errors of the treatment effects to increase. The localised test described by Rudolph et al. (2016) is particularly sensitive

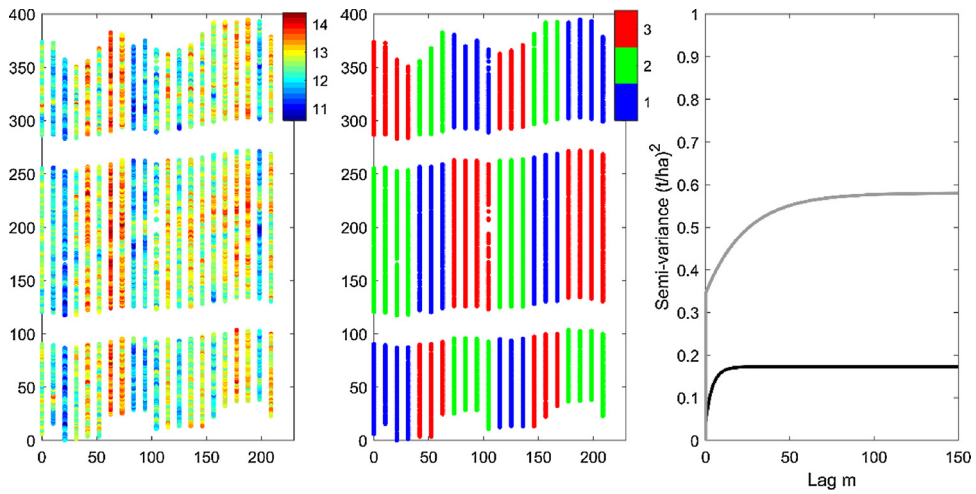

Fig. 6. (left) cleaned and rotated commercial yield monitor data from Experiment 3 (t ha^{-1}); (centre) experimental treatments; (right) estimated variograms in direction of travel (black) and perpendicular to direction of travel (grey). Coordinates are in m.

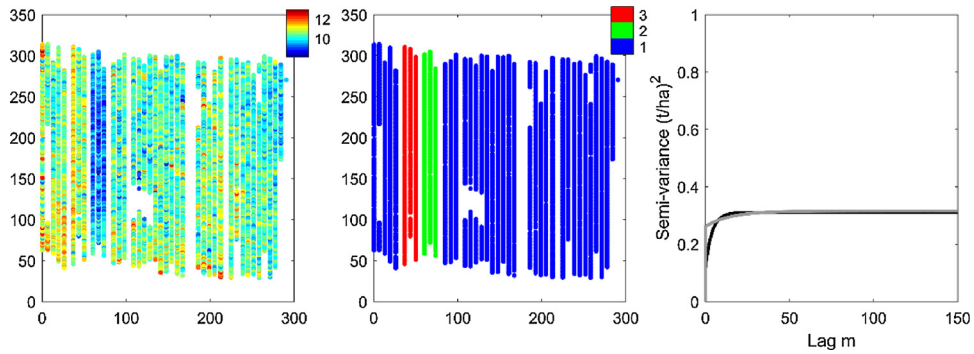

Fig. 7. (left) cleaned and rotated commercial yield monitor data from Experiment 4 (t ha^{-1}); (centre) experimental treatments; (right) estimated variograms in direction of travel (black) and perpendicular to direction of travel (grey). Coordinates are in m.

to violations of the assumption that the same special covariance function applies everywhere. If the random effects close to the treatment boundary are more variable than implied by the second order stationary model then this can lead to significant treatment effects being falsely identified.

We have thus far assumed that the spatial correlation amongst the observed yield measurements is isotropic (i.e. the covariance functions are the same in all directions). In reality, greater spatial correlation is often observed amongst yield monitor measurements collected on the same row than is observed perpendicular to the row. The variation in yield measurements along a row can be smoothed because the grain cut at different points on the header can take different amounts of time to reach the sensor. Conversely, various factors can cause additional variation in measurements from different rows. For example, the combine harvester might be travelling uphill for one row and downhill for the other, the swath might not be a full header width for one of the rows or one row might have reduced yield because it contained farm traffic wheelings. Therefore, an anisotropic spatial covariance model is required. The most commonly used anisotropic models stretch or contract the range of spatial correlation in certain directions (Webster and Oliver, 2007). However, this approach is rather restrictive since the sill variances in all directions are still identical. Li et al. (2016) demonstrated that the product sum covariance model (De Cesare et al., 2001) can lead to a more general representation of anisotropic spatial variation. The product sum model is more commonly used to represent spatial and temporal variation. It permits different covariance functions in time and space. Li et al. (2016) used it to distinguish between the correlation of soil salinity measurements separated vertically and horizontally. Here, we use the product sum model to represent the variation along and perpendicular to a combine harvester row. We rotate the coordinates so that the combine harvester is moving in the y direction. Then the product sum model is written:

$$c(h_x, h_y) = c_x(h_x) + c_y(h_y) + kc_x(h_x)c_y(h_y). \quad (8)$$

Here, h_x is the separation lag perpendicular to the combine harvester direction of travel and h_y is the corresponding lag in the direction of travel. The c_x and c_y are different covariance functions for each direction and $0 < k < 1/(\max\{c_x(0), c_y(0)\})$ is an additional parameter which gives flexibility to model the spatial covariance when both h_x and h_y are non-zero. If both c_x and c_y are nugget and Matérn functions then the product sum model has nine parameters. The nugget in the x direction reflects the difference between the within-row variance of the response variable in comparison to variance throughout the experiment. Small fluctuations in the x caused by inaccuracies in the GPS can prevent this parameter being estimated. Therefore, prior to estimating the product sum model we adjust the x so that each row runs in an exact straight line.

The Akaike Information Criterion (AIC; Akaike, 1973):

$$AIC = 2p - 2L \quad (9)$$

where p is the number of model parameters, can be used to test whether the use of the product sum leads to a significantly better model fit than the isotropic model. The preferred model is the one with the smallest AIC value and thus this criterion penalizes the additional parameters of the product sum model.

The experimenters should also confirm that the observed random effects are consistent with the Gaussian assumption (e.g. by making a visual inspection of the histogram of the random effects). When the assumption is inappropriate the data can be transformed so that they more closely resemble the Gaussian distribution or a more general LMM

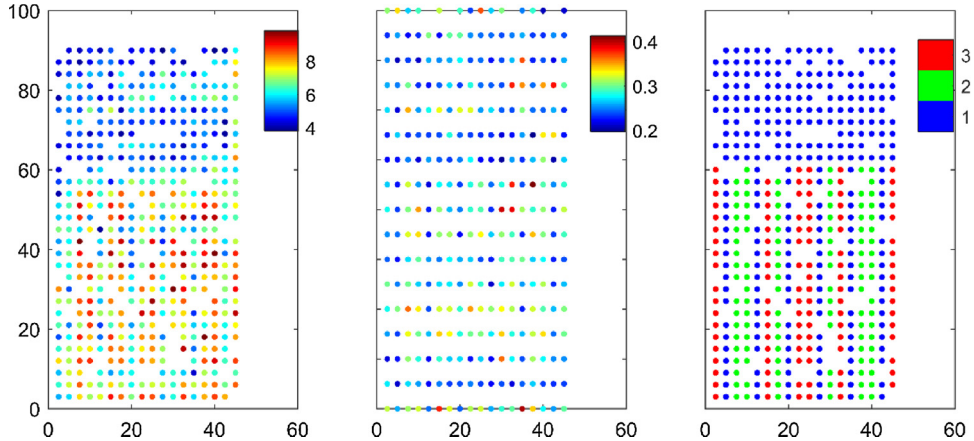

Fig. 8. (left) cleaned and rotated plot combine harvester yield monitor data from Experiment 5 (t ha^{-1}); (centre) cleaned and rotated handheld NDVI data from Experiment 5 (right) experimental treatments. Coordinates are in m.

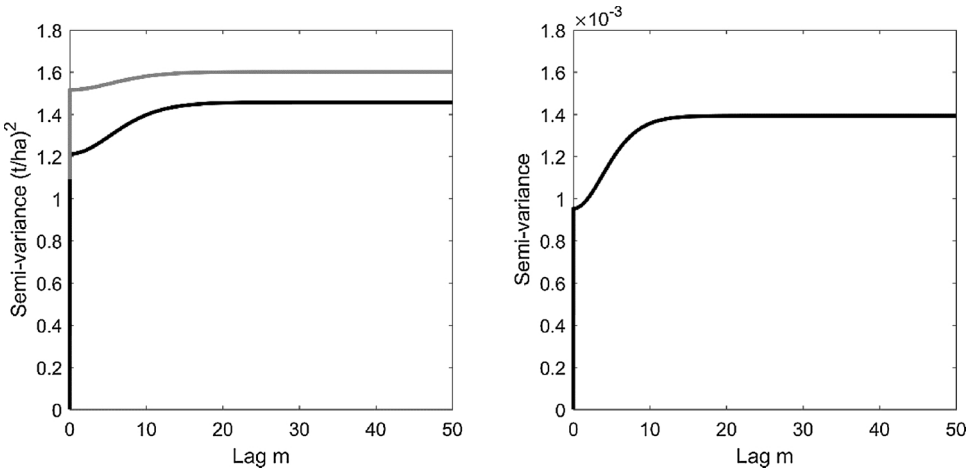

Fig. 9. (left) estimated variograms in direction of travel (black) and perpendicular to direction of travel (grey) for plot combine harvester yield monitor data from Experiment 5; (right) estimated variogram for handheld NDVI data from Experiment 5.

that can accommodate non-Gaussian residuals can be applied (Marchant et al., 2011). This approach of analysing farmers' crop experiments also requires that there are sufficient observations to be able to model accurately the underlying variation of the response variable. Webster and Oliver (2007) suggest that at least 100 observations of a soil variable located on a regular grid are required to estimate a standard variogram model. Yield monitors produce much larger datasets than this, but we do require that a sufficient number of rows of data are collected to calculate an accurate variogram in the direction perpendicular to travel of the combine harvester.

3.1. Analysis of the farmers' crop experiments

The six sites and four sensors provided nine distinct farmers' crop experiments (Table 2). For each experiment, we estimated the difference between the response variable for the control treatment and each of the other treatments featured in the experiment. We calculated the σ_j as described above and interpreted these as measures of the precision of the experiments. The precision of any single experiment will depend on the underlying variability of the response variable, the noisiness of the measured data and the experimental design. We recalculated the σ_j for each commercial yield monitor experiment using the spatial model that was fitted to the experiment with the least variable yield data. This permitted comparison of the precision of the different experimental designs for similarly variable data and we interpreted the resultant σ_j as the precision that could potentially be achieved from good quality yield monitor data.

We are also concerned about the robustness of the results of each

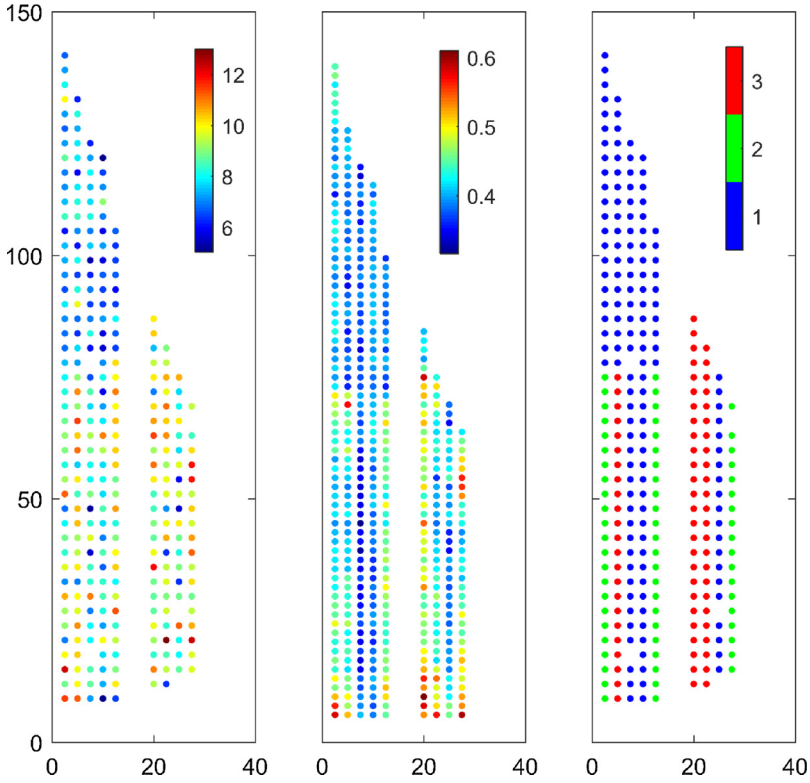

Fig. 10. (left) cleaned and rotated plot combine harvester yield monitor data from Experiment 6 (t ha^{-1}); (centre) cleaned and rotated handheld NDVI data from Experiment 6 (right) experimental treatments. Coordinates are in m.

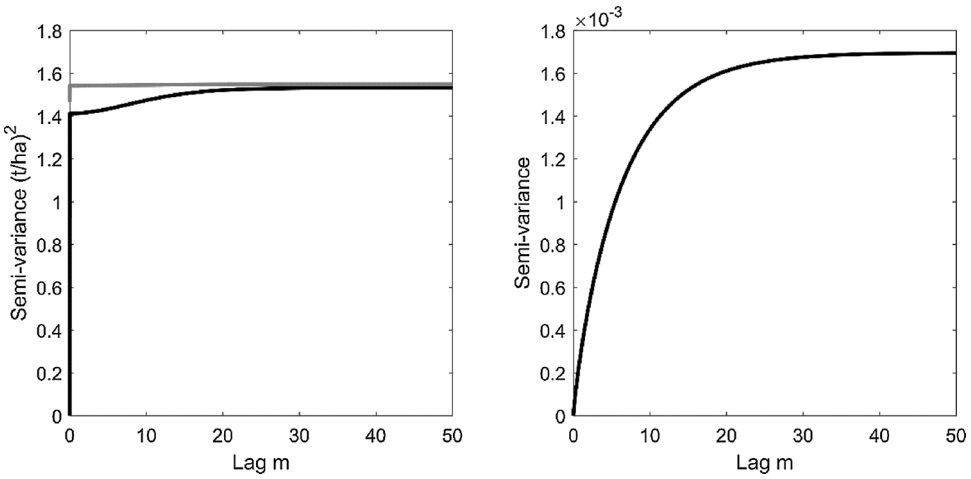

Fig. 11. (left) estimated variograms in direction of travel (black) and perpendicular to direction of travel (grey) for plot combine harvester yield monitor data from Experiment 6; (right) estimated variogram for handheld NDVI data from Experiment 6.

Table 2
Comparisons between responses to various experimental treatments in comparison to the response to the standard treatment for each experiment. Treatment reference (T_j), Treatment effects (β_j), standard errors (σ_j), standard errors when a lateral trend is accommodated in the fixed effects (σ_j incl. trend), standard errors using isotropic random effects (σ_j isotropic), standard errors using independent and identically distributed random effects (σ_j iid) and standard errors using the linear mixed model fitted to yield measurements from Experiment 4 (σ_j using M4). The highlighted standard errors correspond to the model that achieves the smallest AIC.

Exp.	Field	T_j	β_j	σ_j	σ_j incl. trend	σ_j isotropic	σ_j iid	σ_j using M4
Commercial combine harvester yield monitor (t/ha)								
1	A	2	−0.48	**0.21**	0.32	0.12	0.05	0.07
2	B	2	0.45	**0.32**	0.32	0.07	0.05	0.14
2	B	3	0.90	**0.27**	0.27	0.06	0.04	0.11
2	B	4	0.98	**0.27**	0.27	0.06	0.04	0.11
2	B	5	0.88	**0.27**	0.27	0.06	0.04	0.11
3	C	2	0.04	**0.03**	0.03	0.06	0.02	0.03
3	C	3	0.00	**0.03**	0.04	0.06	0.02	0.03
4	D	2	−0.93	**0.13**	0.13	0.08	0.03	0.13
4	D	3	0.40	**0.13**	0.13	0.08	0.03	0.13
Plot combine harvester yield monitor (t/ha)								
5	E	2	1.30	**0.23**	0.24	0.13	0.12	NA
5	E	3	1.64	**0.23**	0.23	0.13	0.12	NA
6	F	2	1.30	**0.26**	0.26	0.20	0.19	NA
6	F	3	1.64	**0.26**	0.26	0.20	0.18	NA
Aerial NDVI (dimensionless × 10^{-2})								
7	A	2	0.04	NA	0.95	**0.94**	0.09	NA
Handheld NDVI (dimensionless × 10^{-2})								
8	E	2	3.82	NA	0.56	**0.55**	0.55	NA
8	E	3	4.96	NA	0.56	**0.56**	0.55	NA
9	F	2	6.12	NA	0.47	**0.47**	0.43	NA
9	F	3	6.13	NA	0.47	**0.46**	0.41	NA

experiment. We therefore adjusted the LMM by adding a linear trend, perpendicular to the direction of travel of the combine harvester, to the fixed effects and then recalculated the σ_j. Any increase in the σ_j would indicate that the experimental design was not robust to such a linear trend. In the yield experiments we accounted for potential anisoptropy in the random effects by using the product sum model. We refitted these models using the isotropic covariance function (Eqn. 2) and independent random effects to explore the consequences of miss-specifying the random effects model. The isotropic covariance function was used in the random effects of the NDVI data and the consequence of miss-specifying independent random effects was explored.

The localised analysis method described by Rudolph et al. (2016) was specifically designed to identify and quantify the step changes in response that can be visually evident (e.g. Kindred et al., 2015) at the boundary between different management practices. We demonstrate and discuss this methodology in relation to one experiment where such a step change is evident.

4. Results

4.1. Pre-processing of commercial combine harvester yield data

The raw commercial grain yield data from Experiment 1 are shown in Fig. 2. It is clear that many of the observed yield values are considerably smaller than would be expected in UK conditions. Two rectangular areas within the field were not harvested by the combine harvester. They were instead used to test the plot combine harvester. Many of the artefacts result from leaving suitable crop for this test.

The observations removed by each filter and the cleaned data are shown in Figs. 3 and 4. Most of the artefacts were flagged by the combine harvester operator or experimenter, primarily because they were located in the headlands. A similarly large number of artefacts were the result of the combine harvester revisiting a site where the crop had already been harvested. Many of these observations occur where the combine harvester is tidying up around the area of the crop which was left to be harvested by the plot trial combine harvester. The direction filter removed a substantial number of observations at the end of rows. These observations would also have been removed by a subsequent filter because they occur soon after or before a break in the data. The number of observations removed purely because the yield was extreme was fairly small and these observations were dispersed across the field. Therefore, it seems unlikely that the filtering of the data led to the removal of observations from a genuinely low yielding portion of the field. The clean data are shown in the lower plots of Fig. 2. The histogram of the yield observations is consistent with the Gaussian assumption required by the statistical model.

The cleaned data from the other commercial yield monitor experiments are shown in Figs. 5–7. Again a substantial number of measurements have been removed from the headlands but otherwise there are few other regions in the fields where a large proportion of data are missing. Some complete rows of data were removed from Experiment 4 because the swath width was less than that of the combine harvester header. Also one patch of unusually low yields was removed (to the right of the treated area in Fig. 7).

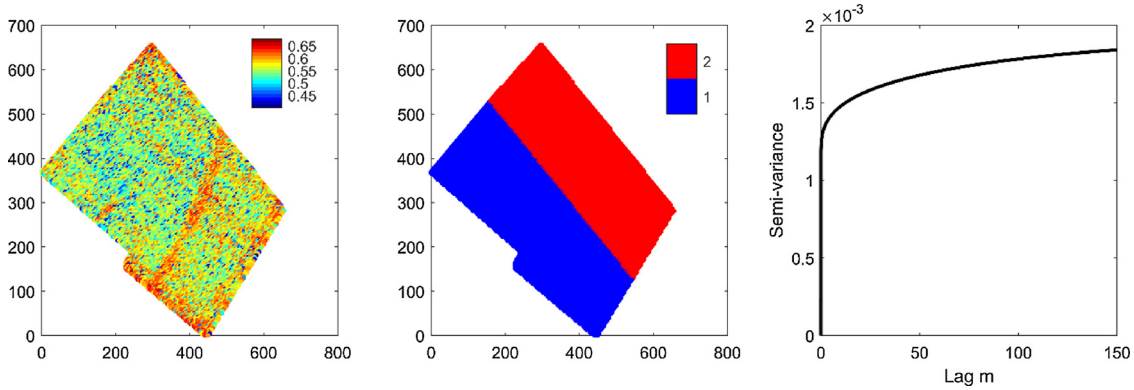

Fig. 12. (left) cleaned and aerial NDVI data from Experiment 1; (centre) experimental treatments; (right) estimated variogram. Coordinates are in m.

4.2. The precision of farmers' crop-experiments using commercial combine harvester yield monitor data

The pre-processed data from the commercial combine harvester yield monitors and the models of residual variation are shown in Figs. 4–7. In each case, the histograms of the yield data were judged to be consistent with the Gaussian assumption and in all cases the product sum covariance function led to the smallest AIC (Eqn. 9) and hence best fit to the data. The experiments differed in the magnitude of the variability contained in the random effects. Experiment 4 had the least residual variation. The sill variance of the estimated variogram function was around 0.3 $(t/ha)^2$ in both directions. In contrast, in Experiment 3 the variogram sill parallel to the swath was less than 0.2 $(t/ha)^2$ whereas perpendicular to the swath it was almost 0.6 $(t/ha)^2$. This was reflected by the stark differences (unrelated to treatments) that were evident in the yield measurements from the different rows of Experiment 3 (Fig. 6). The largest residual variation was observed in Experiment 1.

The precision of the experiments based on commercial combine harvester yield data are compared in Table 2. The largest σ_j of around 0.3 t/ha were recorded in Experiment 2 whereas the smallest σ_j of 0.03 t/ha was recorded in Experiment 3. When the linear trend was accommodated in the split-field design (Experiment 1) the σ_j increased from 0.21 to 0.32 t/ha. In the other experiments there was no substantial increase in σ_j indicating that they were more robust to the presence of such a trend.

When the covariance function was miss-specified to be isotropic, the σ_j for Experiments 1, 2 and 4 decreased by a factor of up to four and a further smaller decrease occurred when the random effects were treated as independent and identically distributed (iid in Table 2). Thus, the wrong choice of covariance function can lead to the uncertainty in the experiments being hugely under-estimated. In Experiment 3 where the treatments varied within rows the use of the isotropic covariance function led to an increase in the σ_j.

When a common spatial model was applied to each experimental design, the largest σ_j of 0.13 t/ha was seen for Experiment 4 which consisted of two sets of treated rows amongst the standard treatment. The σ_j decreased to around 0.11 t/ha for Experiment 2 where there were repeated strips of each treatment and further decreases to 0.07 t/ha for the split-field design which occupied the largest area. The smallest σ_j of 0.03 t/ha was seen for the block design where the treatment varied within each row.

4.3. The precision of experiments using plot combine harvester yield monitor data

The pre-processed and rotated yield monitor data from the experiments where the plot combine harvester was used are shown in Figs. 8 and 10. Differences in yield that reflect the different experimental treatments were clearly evident. In each case the product sum covariance function led to the lowest value of the AIC (Eq. (9)) and hence best fit to the data. The random effects appeared to be more variable than the corresponding data from the commercial combine harvester. The variograms attained a sill variance of around 1.5 $(t/ha)^2$ (Figs. 9 and 11). However, it should be remembered that the plot combine harvester measurements were based on a smaller footprint having a width of 2.1 m compared with the approximately 10 m width of the commercial combine harvester swath. The σ_j from the two experiments were around 0.25 t/ha which is comparable to the least precise commercial combine harvester experiments. Again, the smaller area of the plot combine harvester experiments should be taken into account when assessing the effectiveness of these experiments. The miss-specification of an isotropic covariance function led to the σ_j being substantially underestimated. The increases in σ_j when the fixed effects included a trend perpendicular to the swath were relatively small.

4.4. The precision of experiments using NDVI data

Treatment effects were evident in the handheld NDVI data from Experiments 8 and 9 (Figs. 8 and 10). In contrast, the patterns of variation in the aerial NDVI data from Experiment 7 were not obviously related to the different treatments (Fig. 12). In fact, the most obvious feature within the data ran perpendicular to the treatment boundary. The random effects for both the handheld and aerial NDVI data were similarly variable with a variogram sill of around 1.5×10^{-3}. The σ_j from Experiments 8 and 9 were larger than that from Experiment 7 reflecting the larger area of the split-field design.

4.5. Localised analyses of treatment effects

A jump in the observed yield is evident at the boundary between the standard treatment and the low nitrogen treatment (Treatment 2) in Experiment 4 (Fig. 7). Therefore, this boundary was used to demonstrate the localised analysis method of Rudolph et al. (2016). The shaded regions in Fig. 13 show the 90% prediction interval for yield at the measurement locations to the left of this boundary using the yield measurements and LMM corresponding to the standard treatment. The measured yields at these sites (where Treatment 2 was applied) are marked in red. The upper plot treats each measurement location separately whereas the lower plot averages the measured and observed yield across blocks of five adjacent locations. When the locations are considered individually around half of the measured values fall below the 90% prediction interval indicating that these measurements are significantly smaller than would be expected under the standard treatment. When the averages of blocks of five locations are considered, the prediction interval narrows and all of the measured yields fall below it. The exceptions to this occur at the start of the row. Fig. 14 shows how the standard deviation of the simulated measurements for each

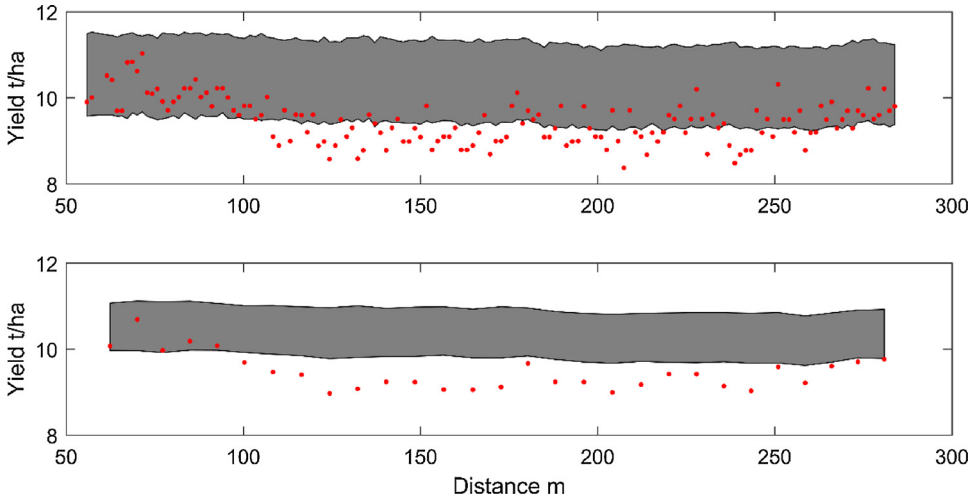

Fig. 13. Ninety percent prediction interval of yield at sites adjacent to treatment boundary under the model for the standard treatment (shaded area) and observed values at these sites under treatment 2 (red) for Experiment 4. Upper plot shows individual yield measurements and predictions whereas lower plot shows averages of five measurements.

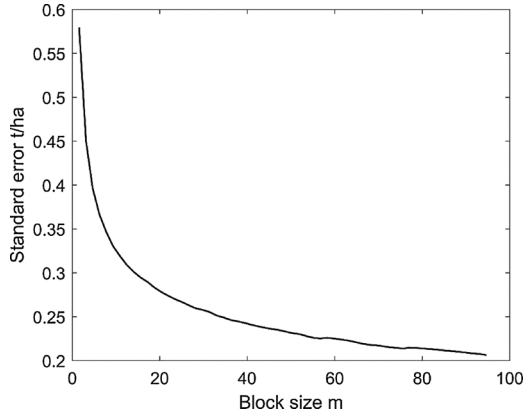

Fig. 14. Standard deviation of predictions of yield at sites adjacent to treatment boundary under the model for the standard treatment against the length of the block over which they are averaged.

block decreases as the block length increases reflecting the improvement in precision as the area of which the treatment effect is assessed increases.

5. Discussion

5.1. Different sources of data for farmers' crop experiments

In common with many other authors (e.g. Griffin, 2010) we have seen that yield monitor data are inherently noisy and contain artefacts which have the potential to confound the results of farmers' crop experiments. Many of these artefacts can be identified fairly easily and removed using the sort of filters described in this paper and by others (Sudduth and Drummond, 2007; Sun et al., 2012). However, this practice is not ideal since the decision about whether to remove an observation will always be subjective to some degree and there is a risk that some observations that relate to genuinely extreme yields will be erroneously removed. Also, any removal of observations corresponds to a loss of information from the experiment.

In our experiments, it was possible to determine the cause of the vast majority of extreme measurements that were removed from the datasets. They often arose from headlands, the combine harvester header not being full, erroneous measurements at the start and end of each row and changes of direction to avoid obstacles such as trees in the field. Where the causes of outliers can be explained it does lead to confidence that extremely high or low yielding portions of the field are not being removed in error.

There are further concerns about more subtle artefacts within the data. For example, the degree of smoothing of the yield observations along a single row is likely to vary with small fluctuations in the speed of the combine harvester but perhaps not to the extent that could be identified by filters. Some of the artefacts within the yield monitor data could be prevented through more careful calibration or operation of the combine harvester. However, such measures are likely to increase the time required to harvest the crop and could then represent a direct cost to the farmer (Griffin et al., 2014).

Our results do indicate that some sets of yield monitor data contain more noise than others once the actual yield variability has been taken into account. This suggests that further research into the sources of noise in yield monitor data is required. The results of such research could indicate why they differ and the extent to which improvements to the design of yield monitoring systems is likely to lead to more accurate yield maps and confidence in farmers' crop experiments. Often, the smoothing effects of the yield monitor on data collected within the same row mean that standard isotopic models of spatial correlation are inappropriate. More complex models, such as the product sum model (de Cesare et al., 2001), are required to quantify the larger variation in observations from different rows. Otherwise the model is likely to lead to overstated inferences about the significance of the yield differences in different rows. Noise and artefacts are also evident in yield data from the plot combine harvester. The noise here is possibly further exaggerated since smaller masses of grain are being measured. It appears inevitable that, at least in the short term, experimenters must accept that yield monitor data are imperfect.

Less noise was evident in the NDVI data. The main identifiable source of noise in NDVI measurements was inclusion of non-cropped regions within the image. In fine scale aerial images the corresponding readings can be easily identified. Other sources of noise resulting from the manner in which the airborne vehicle was flown and the weather conditions cannot be so easily identified and will therefore increase the amount of residual variation in the spatial model and the standard errors of the estimated treatment effects. A further disadvantage of NDVI data is that they do not relate directly to the yield or profitability of the crop. In some instances differences in NDVI might merely reflect differences in crop maturity that could disappear before harvest.

5.2. The precision of farmers' crop experiments

We have seen that the precision of farmers' crop experiments varies according to the experimental design, the experimental area and the inherent unexplained variability or noise amongst the experimental data. The sources of this unexplained variation are often not apparent. They could be the result of within-field variation in crop performance

or imperfections in the sensor used to measure crop performance. In either case this variability can be modelled and accounted for when assessing the uncertainty of the experimental results. When the same model of spatial variability is applied to different experimental designs we see the benefit of having different experimental treatments within a single row and to a lesser extent the benefit of disparate repetitions of each experimental treatment within the experiment. Split-field designs can also yield reasonably small standard errors provided the experimenters can be confident that the fixed effects do not vary between the two portions of the experiments. Such experiments were demonstrated to be compromised by any underlying trend in the fixed effects in a direction perpendicular to the experimental boundary, as commonly occurs in yield data from grain harvesters. Experiments where a treatment only appears in one continuous region, perhaps consisting of a few adjacent traffic-rows across the centre of the field, also have relatively large standard errors since in these circumstances it can be difficult to distinguish a treatment effect from underlying variation in the yield. However, such experiments are more robust to a lateral trend than the split-field design.

Yield experiments where the treatment does not vary within a row are particularly susceptible to substantially underestimated σ_j when the anisotropy in the measurements is not accounted for. This is because these experiments base estimates of the treatment effect on comparisons between measurements from different rows and therefore require that the between-row covariance is accurately estimated.

Many decisions considered by farmers have near-zero cost, so the level of precision achieved in their experiments is important to them, enabling them to accumulate increasing numbers of small yield gains. We have seen that, in ideal circumstances, it is possible to achieve standard errors in yield of less than 0.05 t/ha for experiments with multiple treatments in a row and less than 0.1 t/ha for experiments with multiple rows of each treatment. The largest standard error recorded in an experiment using the commercial combine harvester was 0.32 t/ha. This degree of precision is similar to that obtained from agricultural small-plot trials in the UK. For example, the least significant differences recorded in the Agriculture and Horticulture Development Board's recommended list experiments in 2017 (AHDB, 2017) were in the range 0.33–0.42 t/ha and Sylvester-Bradley et al. (2004) observed least significant differences ranging from 0.22 to 1.00 t/ha but generally between 0.4 and 0.5 t/ha for a series of nitrogen fertiliser trials. The standard errors quoted in this paper can be converted to least significant differences by multiplying by 1.96.

Standard errors of close to 0.25 t/ha were observed for experiments using the modified plot combine harvester but these experiments required a smaller area. The corresponding handheld NDVI experiments achieved standard errors of approximately 0.005 and a smaller standard error was achieved for the split-field NDVI experiment which occupied a larger area.

The precision of the localised analysis (Figs. 13 and 14) increased as the spatial resolution was reduced. When comparisons were made at individual observation sites then a standard error of 0.58 t/ha was estimated for the effect of the low nitrogen treatment in Experiment 4. A comparison over a 100 m block had estimated standard error of 0.2 t/ha.

Researchers often follow the scientific convention of assigning significance to a treatment effect when the probability of the estimate occurring in the absence of a real treatment effect is less than 0.05. Farmers seldom recognise this benchmark (McCown, 2002; Whelan et al., 2012) and are likely to be primarily concerned about the expected profitability of applying a particular management intervention and the uncertainty associated with this estimate. The probability density function that results from the statistical methods described here can be used to calculate the probability that the β_j are positive or greater than a threshold that signifies profitability. Hence these results can support farmers in their decision making process.

5.3. Recommendations for farmers' crop experiments

There is often a trade-off between the precision and the complexity and hence cost of a farmers' crop experiment. Precision can be improved by varying treatments within rows but this requires careful and potentially time-consuming application of the treatments. There is a similar trade-off associated with including disparate repetitions of each treatment within the experiment. More complex designs can also accommodate a larger number of hypothesis tests (e.g. comparisons between a larger number of treatments or interactions between treatments). In yield trials, careful operation of the combine harvester can also lead to improved precision, but this can increase the time required to harvest the crop. The experimenters must consider the costs and benefits of different approaches when deciding upon the appropriate level of complexity of the experiment. Griffin et al. (2014) describe a framework for determining the cost-effectiveness of farmers' crop experiments. The experimenters must also ensure that the experimental design is sufficient to produce accurate spatial models of the underlying variation of the response variable. We would recommend that the experimenters aim to include at least 10 rows of yield monitor data for the standard treatment and that each region of the non-standard treatment is bounded on at least one side by a row of the standard treatment.

The combine harvester operator should take all precautions to improve the quality of the yield data without unduly increasing harvest time (e.g. ensuring that the header is full for each combine harvester row and not cutting with the header straddling across treatment boundary). It can also be helpful if the operator flags any portions of the data which he or she knows to be unrepresentative of the true yield. Alternative sources of data, such as aerial NDVI surveys, cause fewer detrimental effects on the farm operation but the cost of performing the survey must be considered. Also, these data are less directly related to the profitability of the different treatments. Such ancillary data can however be very useful in confirming spatial trends in yield data and identifying anomalous areas that should be excluded from analyses.

The estimated standard errors of the experiment can decrease when the analysis accounts for underlying trends in the yield within the field. Such trends could be quantified through the application of digital soil mapping methodologies (Minasny and McBratney, 2016) or the use of crop yield models and simulation techniques (Carberry et al., 2009). If such trends are ignored then the precision of the experiment can be overstated. These problems are particularly likely to occur when the trends are in a direction perpendicular to the treatment strips or if they coincide with the experimental design. Experimenters should take care to ensure that such trends do not occur. They might consult yield maps from previous years in addition to other sources of spatial information such as soil maps. They should avoid conducting experiments in the same field each year in case the treatment effects persist. They should also use designs with disparate repetitions of one or more treatments since these are less likely to be confounded by an underlying trend.

The analysis of the farmers' crop experiments requires advice from expert statisticians to ensure that appropriate models have been fitted to the data and that therefore the estimated precision of the estimated treatment effects is accurate. Additionally, we have found value in ensuring a close and objective visual inspection of yield maps before yield data are analysed; spatial patterns may well be evident which are not identified by the spatial model.

If it appears appropriate from visual inspection, the localised analysis approach of Rudolph et al. (2016) can be used to assess whether there is a jump in the response at the boundary between two different treatments. The level of precision of these tests is comparable with the field-scale analyses and the tests are not unduly influenced by variation away from the boundary which is unlikely to result from the treatments being compared. However, experimenters should be aware that these tests are reliant on the assumption that the same spatial covariance function is applicable across the field and should not be applied if there is evidence that the variability of the response varies within the field.

Also, the approach can encourage experimenters to make a large number of comparisons between predicted and realised yields. For a single comparison, the measured response falling outside the prediction interval can indicate a significant treatment effect. However, if 100 comparisons are made it is not unreasonable to expect that five of the observations would fall outside the 95% prediction interval purely by chance and this false discovery rate (Benjamini and Hochberg, 1995) should be accounted for. Therefore, we recommend that the localised approach is only used to test a small number of focussed hypotheses (e.g. that there is a significant jump in the response for all of the points on the boundary where there is a particular soil type).

The problems resulting from confounding variation or trends within the field do not occur if a traditional plot experiment with appropriate randomisation and replication is performed. However, few farmers possess the skills, time or equipment to conduct such an experiment and the results might not be appropriate at the scale that is appropriate to farm management decisions. When analysing farmers' crop experiments the experimenters could choose to explore the treatment effect for specific management zones or soil types by including these factors in the fixed effects of the LMM. Such analyses of the interaction between the treatment effect and other environmental covariates are generally not possible with conventional random plot trials unless the design was specially adapted to accommodate such a comparison (Bramley et al., 2013).

We note that farmers' crop experiments can facilitate and be facilitated by the farmers' networks advocated by MacMillan and Benton (2014). If they address the same questions, and plan ahead, farmers in the same locality can share the results of the experiments that they have conducted. The results of farmers' experiments are also amenable to more formal meta-analyses to derive more generally applicable management recommendations.

6. Conclusions

Farmers' crop experiments can provide the agri-food industry with evidence at a relevant spatial scale regarding the most appropriate farm management practices. Such experiments are simpler and less time consuming to implement than traditional replicated plot trials. However, they do require assumptions about the spatial variation of the response variable of interest and expert statistical guidance is required to ensure that these assumptions are appropriate. Noise within yield monitor data can reduce the precision of the experimental results. Further research is required to determine how this noise can be minimised. There is a trade-off between the complexity of an experiment and its precision. Farmers and researchers must decide the most cost effective level of effort to devote to experimentation.

Acknowledgements

We are grateful to Murray Lark for useful discussions, to farmers Ron Gabain, Jeremy Margesson, Tom Banks and George Renner for applying the treatments and harvesting the experiments, supported by Steve Dudman (Frontier Agriculture), the Agriculture & Horticulture Development Board (for LearN trials) and David Wright (Spectrum Aviation). We acknowledge AHDB and Frontier Agriculture for funding the farm research networks that undertook the experiments (AHDB Projects 216-004 and 216-005) and financial support for coordination and analysis from Innovate UK in the Argonōmics project involving ADAS, AgSpace, BASF, BGS, Trials Equipment Ltd and VSN International. This paper was published with the permission of the Executive Director of the British Geological Survey (Natural Environment Research Council).

References

AHDB, 2017. AHDB Recommended Lists for Cereals and Oilseeds 2017/18, Summer 2017 Edition. Stoneleigh, UK Accessed September 2018. https://cereals.ahdb.org.uk/media/800462/ahdb-recommended-list-web.pdf.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Petov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory 267–281.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57, 289–300.

Bramley, R.G.V., Lanyon, D.M., Panten, K., 2005. Whole-of-vineyard experimentation: an improved basis for knowledge and decision making. In: Stafford, J.V. (Ed.), Proceedings of the 5th European Conference on Precision Agriculture. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 883–890.

Bramley, R.G.V., Lawes, R.A., Cook, S.E., 2013. Spatially distributed experimentation: tools for the optimization of targeted management. In: Oliver, M., Bishop, T., Marchant, B. (Eds.), Precision Agriculture for Sustainability and Environmental Protection. Earthscan, UK.

Brus, D.J., De Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–44.

Carberry, P.S., Hochman, Z., Hunt, J.R., Dalgliesh, N.P., McCown, R.L., Whish, J.P.M., Robertson, M.J., Foale, M.A., Poulton, P.L., van Rees, H., 2009. Re-inventing model-based decision support with Australian dryland farmers. 3. Relevance of APSIM to commercial crops. Crop Pasture Sci. 60, 1044–1056.

Cook, S.E., Cock, J., Oberthur, T., Fisher, M., 2013. On-farm experimentation. Better Crops Plant Food 97, 17–20.

De Cesare, L., Myers, D.E., Posa, D., 2001. Estimating and modelling space-time correlation structures. Stat. Probab. Lett. 51, 9—14.

Diggle, P.J., Ribeiro, P.J., 2007. Model-Based Geostatistics. Springer, New York.

Fisher, R.A., Wishart, J., 1930. The arrangement of field experiments and the statistical reduction of the results. Imperial Bureau of Soil Science. Tech. Commun. 10 (1), 23.

Griffin, T.W., 2010. The spatial analysis of yield data. In: Oliver, M.A. (Ed.), Geostatistical Applications for Precision Agriculture. Springer, New York.

Griffin, T.W., Mark, T.B., Dobbins, C.L., Lowenberg-DeBoer, J., 2014. Estimating whole farm costs of conducting on-farm research on Midwestern US corn and soybean farms: a linear programming approach. Int. J. Agric. Manage. 4 (1), 21–27.

Griffin, T.W., Dobbins, C.L., Vyn, T.J., Florax, R.J.G.M., Lowenberg-Deboer, J.M., 2008. Spatial analysis of yield monitor data: case studies of on-farm trials and farm management decision making. Precis. Agric. 9 (5), 269.

Grisso, R., Alley, M., McClellan, P., 2009. Precision Farming Tools: Yield Monitor. Virginia Cooperative Extension, Publication, Virginia, USA, pp. 442–502.

Hicks, D., Vanden Heuvel, R., Fore, Z., 1997. Analysis and practical use of information from on-farm strip trials. Better Crops 81, 18–21.

Kindred, D.R., Milne, A.E., Webster, R., Marchant, B.P., 2015. Exploring the spatial variation in the fertiliser-nitrogen requirement of wheat within fields. J. Agric. Sci. 153 (1), 25–41.

Kindred, D., Sylvester-Bradley, R., 2014. Using precision farming technologies to improve nitrogen management and empower on-farm learning. Aspects of applied biology 127. Precis. Decis. Profit. Crop. 173–180.

Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 57, 787–799.

Lark, R.M., Stafford, J.V., Bolam, H.C., 1997. Limitations on the spatial resolution of yield mapping for combinable crops. J. Agric. Eng. Res. 66, 183–193.

Lawes, R.A., Bramley, R.G.V., 2012. A simple method for the analysis of on-farm strip trials. Agron. J. 104 (2), 371–377.

Li, H.Y., Marchant, B.P., Webster, R., 2016. Modelling the electrical conductivity of soil in the Yangtze delta in three dimensions. Geoderma 269, 119–125.

Little, T.M., Hills, F.J., 1978. Agricultural Experimentation: Design and Analysis. John Wiley & Sons Ltd., West Sussex, England.

MacMillan, T., Benton, T.G., 2014. Engage farmers in research. Nature 509 (7498), 25–27.

Marchant, B.P., Saby, N.P.A., Jolivet, C.C., Arrouays, D., Lark, R.M., 2011. Spatial prediction of soil properties with copulas. Geoderma 162, 327–334.

Marchant, B.P., Saby, N.P.A., Lark, R.M., Bellamy, P.H., Jolivet, C.C., Arrouays, D., 2010. Robust analyses of soil properties at the national scale: cadmium content of French soils. Eur. J. Soil Sci. 61, 144–152.

McCown, R.L., 2002. Changing systems for supporting farmers' decisions: problems, paradigms, and prospects. Agric. Syst. 74, 179–220.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: a brief history and some lessons. Geoderma 264, 301–311.

Muhammed, S., Milne, A., Marchant, B., Griffin, S., Whitmore, A., 2017. Exploiting Yield Maps and Soil Management Zones. AHDB Project Report No. 565. AHDB, UK last accessed March 2017. https://cereals.ahdb.org.uk/publications/2017/january/31/exploiting-yield-maps-and-soil-management-zones.aspx.

Pannell, D.J., Marshall, G.R., Barr, N., Curtis, A., Vanclay, F., Wilkinson, R., 2006. Understanding and promoting adoption of conservation practices by rural landholders. Aust. J. Exp. Agric. 46, 1407–1424.

Pringle, M.J., McBratney, A.B., Cook, S.E., 2004. Field-scale experiments for site-specific crop management. Part II: a geostatistical analysis. Precis. Agric. 5, 625–645.

Rudolph, S., Marchant, B.P., Gillingham, V., Kindred, D., Sylvester-Bradley, R., 2016. Spatial discontinuity analysis' a novel geostatistical algorithm for on-farm experimentation. Proceedings of the 13th International Conference on Precision Agriculture.

Sudduth, K.A., Drummond, S.T., 2007. Yield editor: software for removing errors from crop yield maps. Agron. J. 99, 1471–1482.

Sun, W., Whelan, B., McBratney, A.B., Minasny, B., 2012. An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for

site-specific crop management. Precis. Agric. 14, 376–391.
Sylvester-Bradley, R., 1991. Modelling and mechanisms for the development of agriculture. Aspects of applied biology 26. Art Craft Modell. Appl. Biol. 55–67.
Sylvester-Bradley, R., Kindred, D.R., Blake, J., Dyer, C.J., Sinclair, A., 2004. Optimising Fertiliser Nitrogen for Modern Wheat and Barley Crops. HGCA Project Report No. 438. HGCA, Stoneleigh, UK.
Sylvester-Bradley, R., Kindred, D.R., Marchant, B., Rudolph, S., Roques, S., Calatayud, A., Clarke, S., Gillingham, V., 2017. Agronōmics: transforming crop science through digital technologies. Proceedings of the 11th European Conference on Precision Agriculture.
The University of Reading, 2000. Concepts Underlying the Design of Experiments. Statistical Services Centre. Accessed here 12/12/16. .
Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, 2nd edition. John Wiley and Sons, Chichester.
Whelan, B., McBratney, A.B., 2002. A parametric transfer function for grain-flow within a conventional combine harvester. Precis. Agric. 3, 123–134.
Whelan, B., Taylor, J., McBratney, A., 2012. A 'small strip' approach to empirically determining management class yield response functions and calculating the potential financial 'net wastage' associated with whole-field uniform-rate fertiliser application. Field Crops Res. 139, 47–56.