Computers and Electronics in Agriculture 219 (2024) 108794

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Exploration of data for analysis using boundary line methodology

C. Miti [a,b,*], A.E. Milne [b], K.E. Giller [c], V.O. Sadras [d], R.M. Lark [a]

[a] School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, UK
[b] Net Zero and Resilient Farming, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK
[c] Plant Production Systems, Wageningen University, P.O. Box 430, 6700AK Wageningen, The Netherlands
[d] South Australia Research and Development Institute, The University of Adelaide, Waite Research Precinct, GPO Box 397, Adelaide, SA 5000, Australia

ARTICLE INFO

Keywords:
Boundary line
Convex hull
Standard deviation
Peel density

ABSTRACT

The boundary line model has been proposed for interpretation of the plot of a biological response (such as crop yield) against a potentially-limiting variable from observations in a large set of scenarios across which other factors show uncontrolled variation. Under this model the upper bound of the distribution of data represents the limiting effect of the potential factor on the response. Methods have been proposed to fit this model, but we propose that an initial exploratory data analysis step is needed to evaluate evidence that (i) the model is plausible and (ii) that any limiting upper bound is exhibited by the data set (which could, in principle, not include any cases where the factor is limiting). We propose a statistic based on the density of observations in upper sections of early convex hull peels of the data plot. We evaluate this approach using various data sets, some of which have been used for boundary line analysis in previous studies.

1. Introduction

Biological responses in nature, such as yield of arable crops, are often driven by multiple factors (Cossani and Sadras, 2018), and so the relationship between a response variable, such as yield of a dry-land crop, and a single factor which, in principle, might influence this variable, such as soil phosphorus concentration, have a complex joint distribution. Webb (1972) observed that there may exist a maximum limit which the biological response to a given level of factor does not exceed which he referred to as the boundary line. A boundary line, therefore, gives the maximum possible biological response for a given level of the factor and may be an appropriate model for a biological response in the most conducive environment where other factors are not limiting. Any data points that fall below the boundary line are due to the limiting effects of factors other than the factor of interest.

The boundary line model has attracted attention since Webb (1972) proposed it, and has been widely used in studies that relate biological responses to different environmental and non-environmental factors e.g. in yield gap analysis (Casanova et al., 1999; Fermont et al., 2009; Wairegi et al., 2010), studies of biogenic nitrous oxide emission from soil (Lark and Milne, 2016) and of plant physiology (Buckley, 2017; Shao et al., 2023) and ecology (Su et al., 2022). However, its interpretations only hold if the upper (in some cases the lower) margin of the scatter plot represents a limit and not just the contingent margins of a particular data set. It has been recognized that boundary line models are often used without any justification (Sadras, 2020). Two questions arise when one considers fitting a boundary line model to a particular data set. First, is it biologically or agronomically reasonable to postulate that some upper bound exists on the joint distribution of a response variable y and a potentially limiting factor x. There may be prior biological grounds to expect this, but that will not always be the case. Second, even if a boundary might exist in principle, does the data set cover a sufficiently wide range of conditions such that this boundary is exhibited, that is to say, there is a significant number of cases where the limiting effect of x is expressed because no other factors are limiting. Further practical questions arise: what parametric form of boundary model is appropriate and what values of those parameters could be used as starting points to fit the model?

Most statistical methods used for boundary analysis provide no basis to evaluate evidence that a bounding function is part of a plausible model for the distribution of response variables and others of interest. Lark and Milne (2016) give an example of one, where the evidence can be assessed in terms of the maximized likelihood once a model has been fitted. However, we propose that an exploratory method to examine data to make an initial assessment of the plausibility of a boundary model would facilitate the use of boundary line methods, addressing both the questions of model plausibility and data suitability discussed above, as well as the practical questions for model fitting.

Milne et al. (2006b) proposed such a method which we refer to as the convex hull peel count method (CHPC). The convex hull of

* Corresponding author at: School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, UK.
E-mail address: chawezi.miti@nottingham.ac.uk (C. Miti).

https://doi.org/10.1016/j.compag.2024.108794
Received 30 January 2023; Received in revised form 19 December 2023; Accepted 26 February 2024
Available online 4 March 2024

a bivariate set of data is the smallest convex subset of observations which enclose all observations. The convex hull provides a basis for the procedure of 'peeling' such a data set. The convex hull of the full data set is its first peel, and the second peel is the convex hull of the remaining data once the observations in the first peel are removed, and so on. Milne et al. (2006b) proposed that a data set in which a limiting boundary constrains the possible distribution of observations, and which covers a sufficient range of conditions for that limitation to be exhibited, will have a larger number of observations in the first few peels than would be expected for an unbounded 'null' alternative model such as the bivariate normal distribution. This test can be made more sensitive by focusing on the upper portion of the peel (the subset which is convex upward) when an upper boundary is expected.

Milne et al. (2006b) found that the CHPC method was somewhat insensitive, and that it failed to provide evidence for a boundary in a joint distribution in cases where this could subsequently be justified by likelihood based criteria. We propose that a more sensitive test would not be based on the number of convex hull peel points but rather on their density in the space of the x/y scatterplot. If the joint distribution is bounded over some section, and this boundary is well-exhibited by our data because in a significant number of instances other factors are all not limiting, then we would expect multiple cases to be clustered around this boundary, and for the data set to contrast in this respect from expectations if it were a realization of an unbounded distribution such as the multivariate normal.

The objective of this study is to develop and present a statistical exploratory method that provides evidence of the limiting effect of a boundary in a joint data set on a biological response and a potentially limiting factor, based on the density of points in its convex hull peels. This exploratory method can help the data analyst justify the fitting of a boundary line model to a data set. We also show how these peels might then be used for the initial selection of boundary models. This will be illustrated with different biological data sets, some of which have previously been used in boundary analysis studies.

2. Materials and methods

2.1. Development of method: Determination of peel density and testing its significance

Our method consists of three steps (1) a check on marginal normality and identification and removal of outliers from the dataset, (2) identification of boundary points in successive peels of the dataset, and (3) testing the peel concentration in the upper bounds of the data to see if it is significantly greater than that expected from a bivariate normal dataset of the same size and similar basic summary statistics.

In the first step, we examine the marginal distribution of the x and y variables with histograms and summary statistics. In this case, x represents the independent variable of interest e.g soil nutrient concentration while y represents the biological response e.g crop yield. If a boundary model applies to our data, then the y variable might not appear normal, at least in the upper tail. We therefore, do not expect our data necessarily to appear normal. However, we use plots and summary statistics to evaluate whether it is plausible to regard the variable as drawn from a normal process, perhaps with an upper censoring limit. Variables such as soil nutrient concentrations are commonly positively skewed, and this is a deviation from normality which might influence our diagnostic while not reflecting the presence of a bound. Given our interest in a bounded model we compute, along with the conventional coefficient of skewness, the octile skewness (Brys et al., 2002). If this takes values outwith the range $[-0.2, 0.2]$ then we consider a transformation of the variable (Rawlins et al., 2005). It is also necessary to have an objective procedure to identify and remove outliers from the dataset. For this we use the bagplot, a multivariate equivalent of the univariate boxplot (Rousseeuw et al., 1999). A bagplot has four main components which include, (1) a depth median (equivalent to the median in a boxplot) which is represents the centre of the dataset, (2) a 'bag' that contains 50% of the data points (equivalent to the interquartile range, $Q_3 - Q_1$, where Q_j is the jth quartile, in a univariate boxplot), (3) a 'fence' that separates probable outliers (equivalent to $Q_1 - 3\times(Q_3 - Q_1)$ and $Q_3 + 3\times(Q_3 - Q_1)$ for upper and lower outer fences in a univariate boxplot), and (4) a loop indicating the points outside the bag but inside the fence (see Fig. 1). A bagplot is constructed around a plot of x against y variables of a dataset and all points that fall outside the outer fence are taking to be outliers and are therefore, removed from the dataset. Skewness is checked by looking at the shape of bag and loop (Rousseeuw et al., 1999). In this study we used the bagplot function from the aplpack library in R to compute the bagplot of our data, observations from outside the loop were discarded as outliers.

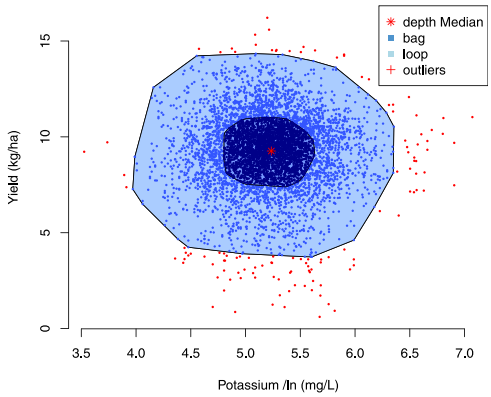

Fig. 1. An example of a bagplot components for a bivariate dataset of wheat yield against soil potassium concentration (Lark et al., 2020).

In the second step, the data points in the outer peels (n = 10) of a dataset are identified using the convex hull method (Skiena, 2008) (see Fig. 2). Because the vertices in a peel are a convex set, one can order them uniquely clockwise or anti-clockwise from an arbitrary first vertex.

Let $\mathbf{v}_i = [x_i, y_i]^{\mathrm{T}} \in V$ denote the ith out of n vertices in a peel of a data set, where V is the set of all vertices in the peel. Let $\acute{V} \subset V$ denote the subset of these vertices where

$$\mathbf{v}_j \in \acute{V} \Rightarrow x_j = \min_{i=1,n}(x_i), \tag{1}$$

We then denote by $\mathbf{v}_l$ the vertex such that

$$\mathbf{v}_l \in \acute{V} \text{ and } y_l = \max_{\mathbf{v}_k \in \acute{V}}(y_k). \tag{2}$$

We call $\mathbf{v}_l$ the first vertex in the clockwise ordering of the upper peel. Similarly the last vertex in the upper peel is $\mathbf{v}_m$ where $\grave{V}$ denotes the subset of vertices in the peel where

$$\mathbf{v}_j \in \grave{V} \Rightarrow x_j = \max_{i=1,n}(x_i), \tag{3}$$

We then denote by $\mathbf{v}_m$ the vertex such that

$$\mathbf{v}_m \in \grave{V} \text{ and } y_m = \max_{\mathbf{v}_k \in \grave{V}}(y_k). \tag{4}$$

Any vertex $\mathbf{v}_i$ belongs to the upper peel set $\hat{V} \subset V$, where the indices i are ordered clockwise, and $l \le i \le m$. This is illustrated in Fig. 2

The upper peel set $\hat{V}$ can be divided into a left and right subset, $\hat{V}_{\mathrm{l}}$ and $\hat{V}_{\mathrm{r}}$ respectively. If the set of vertices in the upper peel with the maximum value of y is denoted by $\dot{V}$ then the mean value of the corresponding values of x,

$$\dot{x} = \mathrm{mean}_{\mathbf{v}_i \in \dot{V}}(x_i), \tag{5}$$

and, for any $\mathbf{v}_i \in \hat{V}$,

$$\mathbf{v}_i \in \hat{V}_{\mathrm{l}} \iff x_i \le \dot{x}$$

$$\mathbf{v}_i \in \hat{V}_{\mathrm{r}} \iff x_i > \dot{x}. \tag{6}$$

If $\hat{V}_{\mathrm{l}}^m$ and $\hat{V}_{\mathrm{r}}^m$ denote, respectively, the left and right upper sections of the mth peel of a data set, then our analysis was based on the combined subsets of sections from the first ten peels:

$$\hat{\mathcal{V}}_l = \bigcup_{m=1,\ldots,10} \hat{V}_{\mathrm{l}}^m \tag{7}$$

and

$$\hat{\mathcal{V}}_r = \bigcup_{m=1,\ldots,10} \hat{V}_{\mathrm{r}}^m \tag{8}$$

Our proposal is that evidence for a limiting boundary in the left or right upper sections of a scatter plot of data can be evaluated by the dispersion of the vertices in these respective subsets, compared with the same statistic for a bivariate normal random variate of the same size with the same parameters. Our proposed statistic is the standard deviation of the Euclidean distances between the left and right subsets and the centroid of the full data set, $\mathbf{m} = [\bar{x}, \bar{y}]^{\mathrm{T}}$, where the average values of the x and y variables over all observations are $\bar{x}$ and $\bar{y}$ respectively. The Euclidean distances between a vertex $\mathbf{v}_i$ and the centroid is given by

$$d_i = \left\{ (\mathbf{v}_i - \mathbf{m})^{\mathrm{T}} (\mathbf{v}_i - \mathbf{m}) \right\}^{\frac{1}{2}}. \tag{9}$$

For each data set we computed the standard deviation of the values d_i for all $\mathbf{v}_i \in \hat{\mathcal{V}}_l$ and the same statistic for $\mathbf{v}_i \in \hat{\mathcal{V}}_r$. If this value is smaller than the corresponding value for the first ten peels of a bivariate normal random variate of the same length as the data set, and with the same covariance matrix, then this is evidence for a greater concentration of vertices in the upper bound (left or right section) of the data set. To test the strength of this evidence we used a Monte Carlo method to obtain a distribution of the test statistic for the case of the multivariate normal null distribution (Mecklin and Mundfrom, 2005).

The sample covariance matrix of the data set was estimated and then used to compute a realization of a bivariate normal random variate of the same length as the data. This was done using the mvrnorm function from the MASS library for the R platform (R Core Team, 2022; Venables and Ripley, 2002). The bagplot was then used to exclude any simulated values which would be identified as outliers according to the criteria we used in the analysis of the real data. The first ten peels of the simulated values were removed, and the left and right upper subsets were extracted for each in turn and combined into corresponding left and right upper bound sets. The standard deviation of the Euclidean distance from the data centroid for each set was then calculated. This was repeated 10 000 times. This number was determined based on the procedure suggested by Percival and Walden (2000). In this procedure, the number of simulations are sufficient if the condition,

$$(\rho - \alpha)^2 > \frac{4M'(M - M')}{M^3} \tag{10}$$

is satisfied. In Eq. (10), M is the number of simulations, M' is the number of times the simulated test statistic exceed the actual observed test statistic, ρ is proportion of the number of times the simulated statistic exceeds the actual statistic and the total number of simulations and α is the critical probability value (0.05 in this case). The empirical distribution of this statistic was then used to compute an approximate p-value for the null hypothesis that the original data had concentration of observations in the upper peels comparable to a normal random variate.

The peel density results in the left and right sections of the data guide on which model to fit by checking the structure of points in the peels when significant clustering is observed. Initial guess parameters to be used in statistical boundary line modelling can be obtained by fitting an appropriate model to the boundary points.

A graphical example of the process of determining the successive peels in a data set is given in Fig. 2 using a scatter plot of a simulated bivariate (x, y) dataset with means, $\mu_x = 0, \mu_y = 0$, correlation $(x, y) = 0$, and covariance 1 (Fig. 2a). The convex hull of the data set is determined as the first peel with the vertices indicated with a red solid round dots (Fig. 2b). These vertices are then split into two, the left, and right sections with respect to the maximum value of y, $y_{\max}$, in the peels (Fig. 2c). Vertices with the value of x_i less than the x value which corresponds to $y_{\max}$ in the peel are denoted as left section (points with solid red circles) while vertices with values of x_i greater than the x value which corresponds to the $y_{\max}$ are denoted as right section (points with solid blue boxes). Vertices in successive 10 peels of the data set are categorized as the combined left and right sections (Fig. 2d). The clustering in the left and right sections can then be determined.

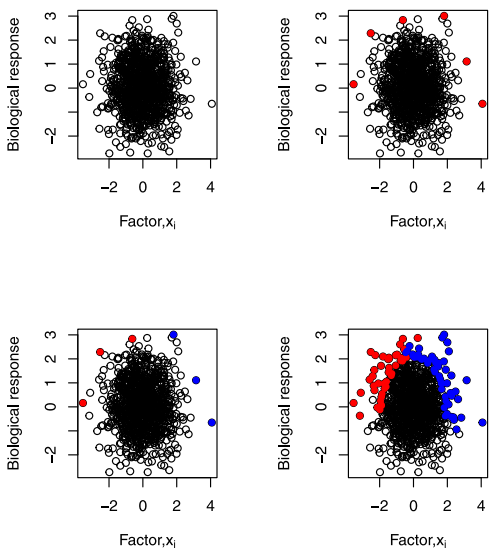

Fig. 2. The process of determining the successive peels in a dataset. (a) A scatterplot of a simulated bivariate (x, y) normal dataset. (b) The convex hull of the dataset determined as the first peel indicated with a red solid round dots . (c) The peel split into the left (red), and right (blue) sections. (d) Vertices in successive 10 peels of the dataset.

2.2. Sample size requirement

The size of sample required to detect a boundary has received little attention. In this case we are concerned with the task of testing a null hypothesis that the observations in the peels of a data set, specifically in the upper increasing or decreasing sectors, are distributed as would be expected for observations from a bivariate normally distributed random variate. Sample size requirement for such inference is the addressed by power analysis. In power analysis we consider a minimum effect size of interest, for which we wish to be able to reject the null hypothesis at a desired level e.g. $p \leq 0.05$ or $p \leq 0.01$. The power is the probability that an underlying effect size would be detected at the specified level, and depends on sample size. In this case we consider the effect size as the proportion of observations which lie on the boundary. This can be specified in a model case where the boundary effect is modelled by censoring a bivariate distribution of variables at some boundary defined as a function $y = f(x)$. Power analysis can be done by simulating a data set from such a censored distribution, with an added measurement error in the response variable, and then running the inference procedure described above. Power is estimated by repeating this procedure multiple times, and noting the proportion of cases in which the null hypothesis is rejected at the specified level. This is the estimate of power.

To demonstrate this process, we simulated a bivariate normally distributed data set of 2000 points that relates a potential limiting soil variable, x and response variable crop yield, y. To create a boundary,

where all other factors are assumed to be non-limiting, a limiting exponential function of the form

$$y = y_0 + y_{\max}(1 - e^{\left(\frac{-x}{c}\right)}) \quad (11)$$

was fitted to the data as the boundary line where y_0 is the intercept, $y_{\max}$ is the maximum possible response value of y and c describes how quickly the response approaches $y_{\max}$. Similar boundary line models have been fitted to data relating crop yields to a soil nutrient e.g. soil P in previous boundary line studies (Fermont et al., 2009; Kintché et al., 2017; Wairegi et al., 2010). The parameters of the function were adjusted so that there were 30%, 20% or 10% data points in the original sample which were above the boundary line. All points that lie above the boundary line were adjusted to the corresponding boundary values, given x. This allowed us to have three datasets with different concentration of points near the boundary. A random error, which represents the measurement error, was added to the response variable y by randomly sampling from a normal distribution with mean 0 and a standard deviation given as a percentage of the mean of y. Three possible percentage values for this error, 2%, 5% and 10% were used. This gave us three possible measurement errors for y.

For each combination of the concentration of points near the boundary and measurement error, the data was sampled with replacement n number of times, where n is equal to a given data size, and the p-value for peel clustering was determined as described in Section 2.1. This was repeated 1000 times and as such, 1000 p-values were determined for a given n. The power was then determined as the proportion of p-values less or equal to the desired significance level (0.05) in the left and right sections of the data. A confidence interval for the power was calculated using the method proposed by Blaker (2000). This process was repeated for varying data sizes, n, equal to 100, 300, 500, 700, 900, 1000 and 2000. A power of 80% is usually considered sufficient in most experimental studies (Scheiner and Gurevitch, 2001). Therefore, a data size that gives 80% power is considered appropriate to detect boundary when it exists at a given significance level. This allowed us examine a minimum data size required to detect evidence for a boundary given a measurement error and proportion of points limited by the boundary.

2.3. Description of the experimental and field data used

We illustrate our method with seven data sets, some of which have been used in previous boundary line analysis studies. These are described below.

2.3.1. Dataset 1: Wheat yield vs. evapotranspiration

Dataset 1 was compiled by Sadras and Angus (2006) and comprises measured wheat yield and estimated evapotranspiration (ET) from sites in China, the Mediterranean regions of Europe, North America, and Australia. For more details about this dataset refer to Sadras and Angus (2006). Yield is associated with factors other than evapotranspiration. However, for a given evapotranspiration, yield is biologically bounded by a conserved upper limit of the biomass-transpiration ratio, and the theoretical harvest index (Foulkes and Reynolds, 2015; French and Schultz, 1984). In our analysis, we examined evidence of peel clustering on the upper bound for yield in response to ET to reflect the bounding relationship.

2.3.2. Dataset 2 and 3: AgSpace Agriculture Ltd wheat yield and soil property survey data

Dataset 2 and 3 was compiled by AgSpace Agriculture Ltd and comprises measures of wheat yield and selected soil properties including potassium (K) and phosphorus (P), which were measured across England in different management units, in this case in 2016. AgSpace Agriculture Ltd conducts soil sampling for its customers on the basis of pre-identified management zones within each field. The management units were delineated by experienced soil scientists using a free survey and each management unit formed the basis for the sampling zone. Twenty four soil cores to depth of 15 cm were collected in each sampling zone and combined into a bulk sample. From the bulk sample, a subsample was then taken for laboratory analyses for P and K. The Olsen's method was used to extract P while 1M ammonium nitrate was used to extract K. The result was treated as the estimate of the sampling zone. The mean wheat yield was measured for each zone for the year 2015 to 2017. The dataset used in this study is based on measurements done in 2016. For more details about this study, refer to Lark et al. (2020). In our analysis, we examined existence of a boundary on the upper edges of yield response to P (and K) which is taken as the yield response when other factors are not limiting.

2.3.3. Dataset 4: Leaf stomatal conductance of broad bean plants

Dataset 4 is based on a pot experiment that was conducted in 2003 at Silsoe, Bedfordshire, UK, to evaluate the effect of soil water status on stomatal conductance. Broad beans plants (*Vicia faba*) were planted in compost and then transplanted into 200 one litre pots with soils of varying textural classes after germination and grown in a glass house. The variation in texture gave rise to varying soil moisture conditions. Simultaneous leaf stomatal conductance (mmol m^{-2} s^{-1}) and volumetric soil water content (%) measurements were made using an AP4 porometer (Delta-T Devices Ltd, Burwell, Cambridge, UK) and a Theta Probe (Delta-T Devices) respectively on a regular basis during the growth period. For more information on the data see Milne et al. (2006b). Leaf conductance is dependant on stomatal opening which is affected by water status amongst other environmental factors. It is thus expected that conductance will be maximum when the stomata are fully open. In this study, we examined the relationship of leaf conductance and volumetric soil water content. We expect an upper boundary in this relationship to represent leaf conductance when other environmental factors are not limiting. We note that various studies by plant physiologists have used the boundary line concept to relate process models of stomatal function to corresponding data (e.g. Buckley, 2017).

2.3.4. Dataset 5: Leaf stomatal conductance of winter wheat plants

Dataset 5 is based on a pot experiment that was conducted at Silsoe, Bedfordshire, UK, to evaluate the effect of soil water status on stomatal conductance. A winter wheat crop (*Triticum aestivum* var. Consort) was grown in the season 2002/2003 in a field with soils of varying textural classes. After germination, three wheat plants were transplanted into each of the 200 one litre pots filled with soils of varying textural classes from the field they were initially planted. Simultaneous measurements of stomatal conductance (mmol m^{-2} s^{-1}) and volumetric soil water content (%) were taken at an interval of three weeks using an AP4 porometer and a Theta Probe respectively. For each pot, conductance measurements were made on six leaves and the mean of three moisture content measurements was assumed to be the associated moisture content value. For more information on the data see Milne et al. (2006b). Just as in dataset 4, we examined the relationship of leaf conductance and volumetric soil water content. Similarly, we expect an upper boundary in this relationship to represent leaf conductance when other environmental factors are not limiting.

2.3.5. Dataset 6: Vegetation index of winter wheat plants

Dataset 6 is based on an experiment conducted at Silsoe Research Institute, Bedfordshire, UK, during the season 2000/2001 to study wheat response to nitrogen spatial variations. This was a randomized block design having 465 plots with five different rates of nitrogen fertilizer on an 11.6 ha field. At the end of the growing season, the crop was harvested and local yield response to nitrogen functions were estimated at nodes of a square grid of 10-m sides. As described by Lark and Wheeler (2003), the local yield response functions are of the form $Y = a + bRN$. Yield, Y, therefore, increases with nitrogen rate (N) to an asymptote, which they called the local asymptotic yield (LAY). In a follow up study by Milne et al. (2006b), they evaluated the possibility of predicting local asymptotic yield at an early stage in the season using

Table 1
Summary statistics of the response and independent variables in datasets 1 to 7.

Dataset	Size (n)	Variable	Mean	Median	sd	Skewness	Octile skewness
1	691	ET (mm ha^{-1})	289.62	281.54	83.70	0.54	0.15
1	691	Yield (ton ha^{-1})	2.41	2.27	1.08	0.53	0.12
2	6358	P (mg kg^{-1})	25.96	22	14.39	1.84	0.36
2/3	6358	Wheat yield	9.25	9.36	1.86	−0.48	−0.06
3	6358	K (mg kg^{-1})	198.34	183	84.52	2.36	0.21
4	1438	Leaf conductance (mmol m^{-1} s^{-1})	16.77	14.2	9.72	1.23	0.36
4	1438	Moisture content (%)	13.69	12.9	5.19	0.91	0.14
5	3430	Leaf conductance (mmol m^{-1} s^{-1})	22.80	15.55	20.52	1.56	0.52
5	3430	Moisture content (%)	15.35	13.2	9.08	0.47	0.30
6	200	NDVI	0.65	0.66	0.06	−0.84	−0.29
6	200	Local asymptotic yield (ton ha^{-1})	8.43	8.14	1.145	0.48	0.33
7	188	SOC (%)	1.07	1.07	0.16	0.08	0.08
7	188	Soil clay (%)	26.03	25.33	3.75	0.82	0.22

Table 2
The probability (p-value) of getting an sd value less than that of a normal bivariate joint distribution on the left (l) and right (r) sections of datasets 1 to 7.

Dataset	Variables	sd_l	$\bar{sd}_l$	p-value$_l$	sd_r	$\bar{sd}_r$	p-value$_r$
1	Wheat yield vs. ET	54.229	64.824	0.019	79.908	57.511	0.999
2	Wheat yield vs. log P	1.045	1.181	0.019	1.115	1.276	0.013
3	Wheat yield vs. log K	1.208	1.294	0.097	1.335	1.390	0.229
4	Leaf conductance vs. log moisture content	2.450	4.162	0.000	4.970	3.201	1.000
5	log leaf conductance vs. log moisture content	0.332	0.448	0.000	0.286	0.340	0.096
6	Local asymptotic yield vs. NDVI	0.609	0.692	0.125	0.688	0.714	0.379
7	Inv-SOC vs. log soil clay	0.085	0.075	0.839	0.090	0.0854	0.709

spectral reflectance in the visible red and near-infrared region which were measured at the time of second nitrogen applications. For this, they used Skye Instruments type SKR1800 dual channel radiometers (Skye Instruments, Llandrindod, Powys, UK) fitted with narrow band interference filters centred at 660 nm and 730 nm. These were mounted on a 24-m boom at 4-m intervals. NDVI was calculated from these measures. For more details on the data see Lark and Wheeler (2003). We expect this relationship to be limited by an upper boundary that shows the potential yield and so, we evaluated evidence of boundary existence in this dataset.

2.3.6. Dataset 7: Soil carbon and clay content of soils at the Broadbalk wheat experiment site

The Broadbalk wheat experiment at Rothamsted, Harpenden, UK is one of the oldest continuous agronomic experiments in the world. It was set up in 1843 to tests long-term effects of fertilizer and cropping treatments (for more details see Powlson (1994). Here we consider paired measurements of soil organic carbon (SOC) and clay content that were taken on plots from the Broadbalk experiment as part of a study by Watts et al. (2006)). In this study SOC was measured on 188 plots from Broadbalk and clay contents on a subset of these (131 plots in total). The missing clay values were estimated by linear interpolation. Prior to cultivation in autumn 2000 the soil was sampled to a depth of 23 cm using a 19-mm-diameter gouge auger. A total of 18 samples were taken per plot, which were then bulked. Total C was determined by combustion and inorganic C ($CaCO_3$-C) was determined by manometry (Martin and Reeve, 1955). SOC was calculated and expressed as percentage (g SOC per 100 g soil). The clay content (%) of the soil was determined by sieving and sedimentation. Our conjecture is that the clay protects the organic matter against bacterial degradation, and so SOC cannot fall below a clay-content-determined threshold (Milne et al., 2006a) and hence we expect some bounding effects at the lower bounds of the dataset. Due to the fact that the limiting response of SOC to clay content has a lower boundary rather than an upper boundary, which our method tests, this dataset has been inverted by multiplying the soil organic carbon content by −1 to create a new variable called 'Inv-SOC' and thus the relationship between soil clay content and Inv-SOC is expected to have an upper boundary (Fig. 3(g)).

3. Results

Table 1 shows the summary statistics mean, median, standard deviation, skewness and octile skewness of the variables in the different datasets. The variables conductance, in datasets 4 and 5, soil P concentration in dataset 2, Soil K concentration in dataset 3, NDVI and LAY in dataset 6, and the clay content in dataset 7 have an octile skewness outside the range of [−0.2,2] and hence indicate skewness. The skewness of these variables can also be observed in the exploratory bagplots of datasets 1 to 7 presented in Fig. A.1 in Appendix A.

In dataset 2 and 3 (Figs. A.1(b) and A.1(c)), the depth median leans to the left of bag while datasets 4 and 5 (Figs. A.1(d) and A.1(e)), the depth median is leaning towards the bottom left of the bag plot. A log-transformation was done on these variables bring them to normality. For NDVI and LAY in dataset 6 transformation did not improve the normality. Exploratory histograms have also been presented in Appendix B for these variables and their transformations. Using the bagplot, outliers were observed and removed from datasets 1, 2, 3, 4 and 5 (see Appendix A). No outliers were observed for datasets 6 and 7. The scatter plots in Figs. 3(a) to 3(g) show the relationships between the response variables and the independent variables for datasets 1 to 7 respectively showing the boundary points in the left and right sections of the datasets.

Table 2 shows the results of the hypothesis tests for evidence of clustering of peels at the upper bounds of the seven datasets. The sd_l and sd_r are standard deviations on left and right sections respectively, $\bar{sd}_l$ and $\bar{sd}_r$ are the means of the obtained sd of the left and right sections from the 10 000 simulations. Three scenarios are possible for any dataset, (1) No evidence of a boundary in the dataset, (2) there is evidence of a boundary on one side of the scatter plot (left or right sections) or (3) there is evidence of a boundary on both sides of the dataset. For dataset 1, which relates wheat yield to ET, the p-value of the left section is 0.019, indicating that there is evidence of the existence of a boundary on the left side of the dataset while the right side shows no evidence (p-value > 0.05). This is a similar for dataset 4 and 5, relating beans log stomatal conductance to volumetric soil water content and wheat log stomatal conductance to log volumetric soil water content respectively, as well as the dataset 7 which relates log soil clay content and Inv-SOC. The dataset 2, relating log P concentration to wheat yield, shows evidence of a boundary existence in both the left and right sections

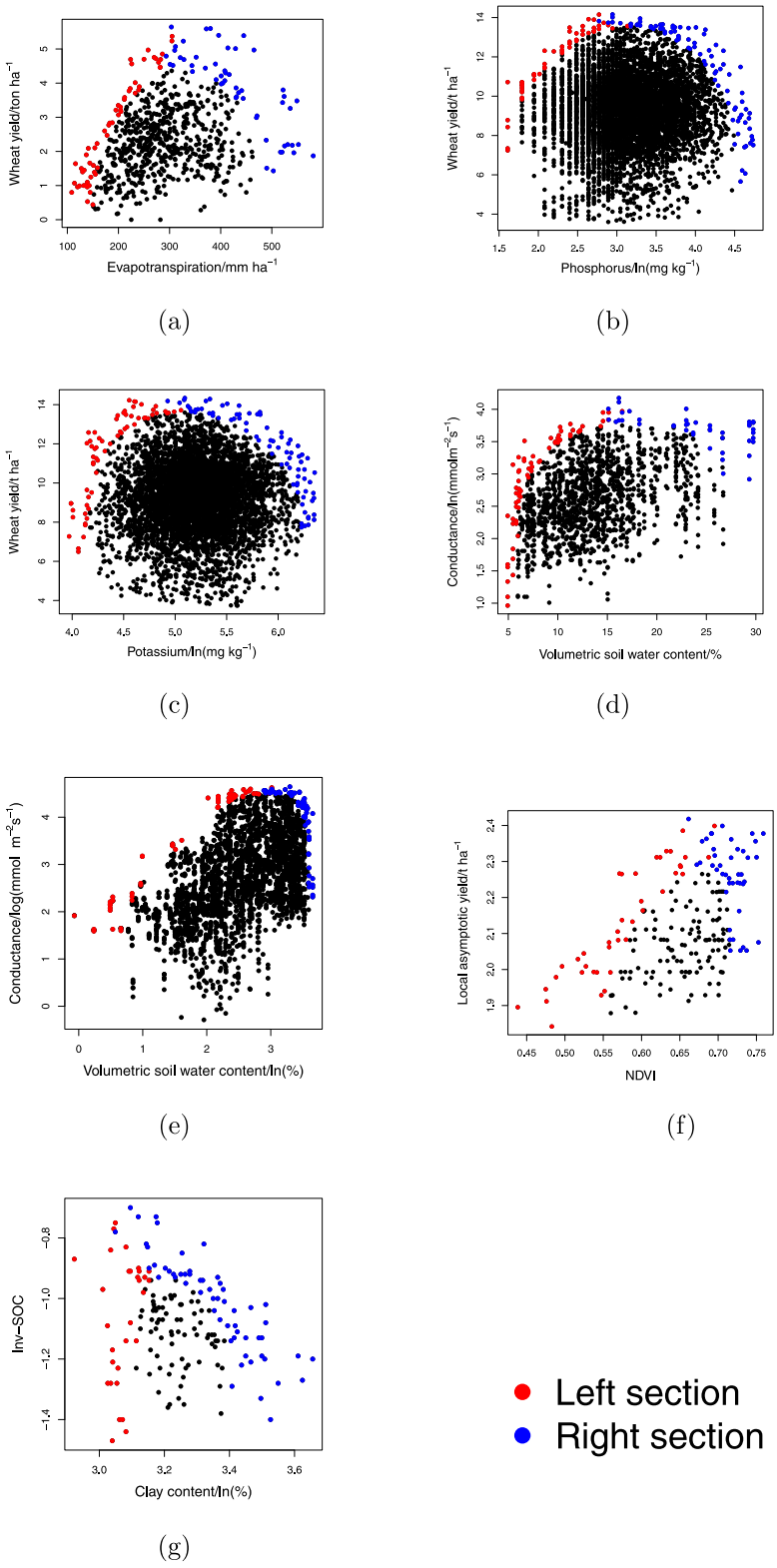

Fig. 3. Scatter plots of (a) wheat yield against evapotranspiration, (b) wheat yield against log phosphorus, (c) wheat yield against log potassium and, (d) beans log stomatal conductance against soil water content, (e) wheat log stomatal conductance against log volumetric water content, (f) local asymptotic yield against NDVI and (g) Inv-SOC against log soil clay content, showing the boundary points in the left and right sections of the datasets 1 to 7 respectively.

of the data (p-values < 0.05). Datasets 3 and 6 (relating log soil K concentration to wheat yield, and NDVI to the local asymptotic yield) do not exhibit evidence of a boundary in both the left and right sections of the datasets (p-values > 0.05).

The Fig. 4 shows two examples of the appropriate form of models that can be fitted to the datasets relating ET and wheat yield (Fig. 4(a)), and log soil phosphorus concentration and wheat yield (Fig. 4(b)) as guided by the peel density results in the left and right sections of the

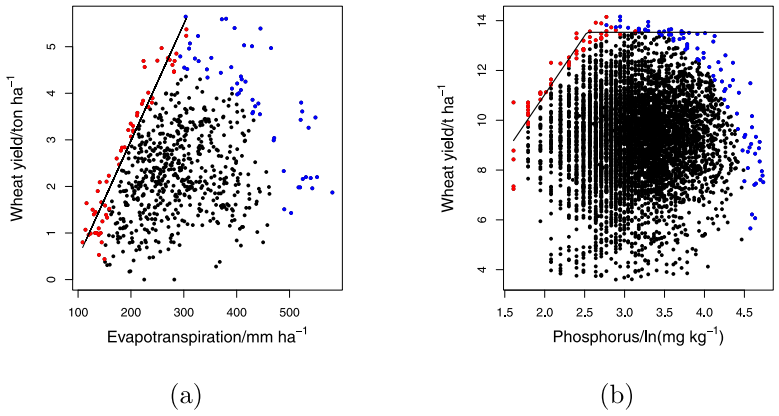

Fig. 4. Fitting appropriate forms of boundary model to datasets 1 (a) and 2 (b) based on peel density results.

datasets. A linear model of the form

$$y = mx + c \tag{12}$$

was fitted to the dataset 1 (Fig. 4(a)) where c represents the y-intercept and m is the slope. The coefficients of the fit are $c = -2.04$ and $m = 0.025$. A linear plus plateau broken stick model of the form

$$y = \begin{cases} mx + c, & \text{if } x \leq bp \\ bpx + c, & \text{if } x > bp \end{cases} \tag{13}$$

was fitted to dataset 2 (Fig. 4(b)), where c representing the y-intercept, m is the slope and bp is the x value at which the equation changes from linear to plateau. The coefficients of the fit are $c = 1.57$, $m = 4.73$ and bp = 2.53.

The results of the power analysis from the simulated data set are shown in Fig. 5. These results are for the left section only as evidence of boundary was only observed in the left section of the simulated dataset. At a concentration of 10% of points near the boundary, a power of 80% was attained at a sample size of 800 when measurement error was 2% of mean yield while a power of 80% was achieved at 900 and 1000 data size for measurement errors of 5% and 10% respectively (Fig. 5(a)). Increase in concentration of data points from 10% to 30% reduced the data size required to achieve a power of 80% (Fig. 5(b)). At a measurement error of 5%, a 10% concentration of data points at the boundary achieved 80% power with a data size of 800 while this was achieved at a data size of 750 and 650 for a concentration of 20% and 30% respectively.

4. Discussion

An exploration of a dataset with objective and repeatable statistics should be a first step in boundary line analysis. In previous studies the decision to fit a boundary model has been based on visual inspection of the data, and in most cases once the model is fitted there is no basis for *post hoc* assessment of the boundary-based interpretation. This will limit the validity and practical value of the model itself. For example, dataset 6 (Fig. 3(f)), from which we expected to predict the local asymptotic yield from NDVI measurements at the upper bounds of the data, looks to have a limiting response of local asymptotic yield to NDVI which may take the form of a rising linear function from point {0.45,6} to point {0.7,10.5}. However, the test shows that there is insufficient evidence to support a boundary-based interpretation. It might be better to fit a predictive model with additive effects of other potential limiting factors where these can be measured, or it might be necessary to collect more data from a wider range of conditions to exhibit a biological bound convincingly. This is a similar case to dataset 7 (Fig. 3(g)) which relates Inv-SOC and log soil clay content. Clay protects SOC from microbial degradation by forming organo-clay compounds which reduce the SOC loss, the greater the clay content, the greater the SOC is expected (Singh et al., 2018). Therefore, there is a limiting response such that a given amount of clay content will hold a minimum amount of SOC otherwise it will always be above that minimum. It is, therefore, expected that the test will pick a boundary in the right section of the scatter plot (recall the data was inverted by multiplying SOC by −1). However, the test on this dataset does not give sufficient evidence of the existence of this boundary in the right section, where our prior expectation of bounded behaviour holds. It is possible in this case that the data, coming from a single field, albeit a variable one, do not represent sufficiently varied environmental conditions to exhibit the lower bound of interest.

For dataset 3 (Fig. 3(c)), we expect to have a response of increasing yield with soil K concentration up to a given level of K (K_{peak}) that produces maximum yield. Beyond K_{peak}, yield will not increase further but reaches a plateau. We might expect a reduction in yield at some point beyond K_{peak}, perhaps because within-field regions with severe limitations from factors other than available K tend to accumulate this nutrient in the soil because of small rates of offtake by the crop, or if a large concentration of K reduces the retention of other cations like magnesium leading to its deficiency. Although the visual inspection shows some form of a relationship of which the yield initially increases with K and then reduces after some point {5.2,14}, the test shows that there is no significant peel clustering in both the left and right sections of the dataset.

As we have noted above, such negative outcomes do not necessarily preclude the boundary-line interpretation for a relationship between variables. It may be, for example, that the boundary is not exhibited in the particular data set because of other limiting factors, or that the data set is too small to provide evidence for a relatively complex model, and more data are needed. If the boundary line is to be fitted as an explicitly statistical model (e.g. Lark and Milne, 2016), then it may be justified to proceed and to evaluate internal evidence for this model before applying it. However, where this is not done, as in most studies on yield gap analysis or wider boundary line analysis, then in cases such as our datasets 3 (relating log K concentration to wheat yield) and 6 (relating NDVI to local asymptotic yield), then a boundary model is hard to justify. We think that boundary line methods, as applied in a range of fields, would gain credibility if this approach were used to justify the fitting and interpretation of boundary response models.

Datasets 1, 4 and 5 show evidence of a boundary effects in the left sections only. Both visual assessment of the plots in Figs. 3(a), 3(d) and 3(e) and peel density test are consistent in this. Visual interpretation

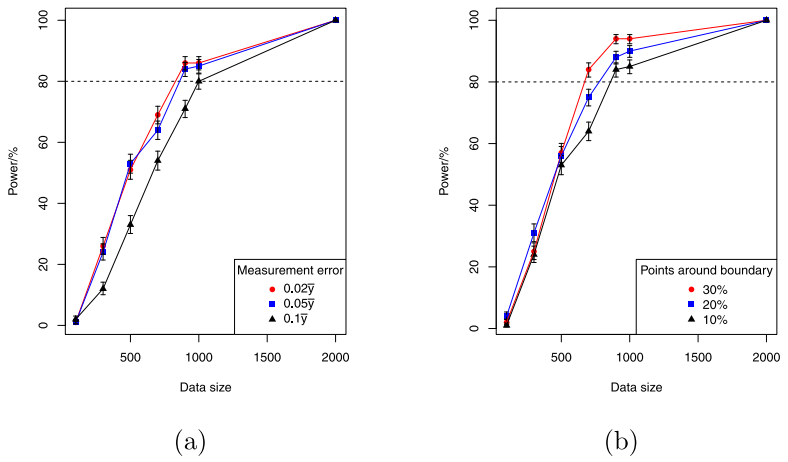

Fig. 5. Detection power with confidence interval for (a) varying measurement error and (b) varying concentration of points at the boundary of a dataset at 0.05 significance level.

of Fig. 3(a) (dataset 1), suggests that a linear boundary yield response applies from an ET of 100 mm ha^{-1} to 300 mm ha^{-1}. However, above the ET of 300 mm ha^{-1}, there does not seem to be a well defined boundary. Earlier research by French and Schultz (1984) showed a positive relation between transpiration and the crop dry matter production. However, it is also expected that some data points will fall below this established relationship because of other biotic and abiotic factors affecting dry matter production. This positive relation between the ET and wheat yield is maximized at the upper edges in the left section of the dataset as confirmed by the positive test of peel clustering.

For dataset 4 and 5 (Figs. 3(d) and 3(e)), visual inspection suggests some form of limiting response (boundary) which is characterized by an initial increase in stomatal conductance with increasing soil moisture content and a plateau thereafter. Stomatal conductance is influenced by plant water status i.e. increase in moisture content increases stomatal conductance. However, the porosity of a leaf is controlled by opening of the stomata, which not only respond to plant water status but also other environmental factors (Lavoie-Lamoureux et al., 2017). These environmental factors may reduce the conductance below the expected as permitted by the plant water status. Therefore, the relationship between conductance and the water status of a plant is expected to have a boundary at the upper limits at which the conductance is maximum when the stomata are fully open i.e. when not restricted by other environmental conditions. Stomatal conductance will fall below this maximum when environmental conditions limit stomatal opening. This is confirmed by the test in both the beans and wheat leaves which indicate a presence of boundary in the left section. For dataset 5 (Fig. 3(e)), the left section shows rise and plateau which is in agreement with the theoretical basis. For dataset 4 (Fig. 3(d)), it does not show evidence of peel clustering at the plateau of constant conductance although a visual interpretation might suggest this. This may result from there being fewer observations data points on the right side of the scatter plot (wetter soil) compared to the left side and insufficient to exhibit an upper bound. There are notably more observations with soil water content below 17.5% v/v than above. As a result we have insufficient data to exhibit a plateau conductance and our data are not at the boundary in right section of the scatter plot. Though there are some points in the right section of dataset 5 (Fig. 3(e)) which may suggest that conductance will reduce with increased moisture content above 3.5, there is no biological explanation for this. Neither of these datasets showed evidence of a boundary when the method of Milne et al. (2006b), which considers the number of vertices in the peel, was used. In our proposed method, the data is split into two sections (left and right) and each section is tested separately. This aids detection of structure in the different parts of the data scatter and therefore increases the sensitivity for detecting a boundary.

For dataset 2 (Fig. 3(b)), visual inspection suggests that there is a linear limiting response of yield to log P concentration from point {1.8,10.8} to point {2.6,13.8} which reaches a plateau thereafter at yields of about 14 t ha^{-1}, meaning that there is no further increase in yield with increased log P at this stage. This is confirmed by the test which has a positive test for boundary occurrence in both the left and right sections of the data. We expect that yield will increase with P and then plateaus at some soil P content (P_{peak}). There maybe some negative effect of increasing soil P concentration above some given value beyond the P_{peak} if there is some indirect effect of P on yield e.g. too much P can inhibit the development of soil organisms like mycorrhizal fungi which have symbiotic relationships with plant roots and are necessary for healthy plant growth. However, this is usually not very common and hence the linear plus plateau model illustrated in Fig. 4(b) is an appropriate model for this data. Another possible interpretation was raised for the corresponding data set on soil K above, if there is a consistent limiting effect of some other factor in a region of a field, then an immobile nutrient such as P may accumulate in the soil there because take-off by the crop is small. In this part of the boundary line the soil P itself is thus a proxy for other limiting factors.

In these cases where the evidence of a boundary is provided, one is confident to fit the boundary line model to these datasets.

In boundary line analysis, there is a need for one to choose an appropriate model to fit a dataset after it has been established that a boundary model is plausible (Lark et al., 2020; Milne et al., 2006b,a). Various models are available to fit to datasets. Some datasets may conform to models that show a decrease in response variable, y, with an increase in the independent factor, x, e.g. relationship between timing of first weeding operation and crop yield (Fermont et al., 2009), others datasets may conform to models that exhibit a linear rise in response variable as the independent factor increases. This will, however, not increase to infinity as biological response will always reach a limit. Therefore, some models will show an increasing response with an independent factor until they reach a maximum, at which point the response will decrease with increase in the independent factor e.g. the response of soil nitrous oxide emission to soil water filled pore space (Schmidt et al., 2000) while some models will show an increase in response with factor until it reaches a maximum after which an increase in factor will not result in any increase in response resulting in a plateau of response variable e.g. the response of soil nitrous oxide emission to soil nitrate content (Schmidt et al., 2000). An appropriate model for a particular dataset must thus be chosen if the results of the analysis are to be reliable and of practical use.

The division of the vertices in the peels into left and right sections, as given in the method we propose, provides guidance on what model one can fit the data. Taking dataset 1 as an example (Fig. 3(a)), from the scatter plot of yield (t ha^{-1}) against ET (mm ha^{-1}), one may be tempted to fit a non-linear or broken stick boundary line model of increasing yield with increasing ET from the point {105 mm ha^{-1}, 1.7 t ha^{-1} } up to the point {300 mm ha^{-1}, 5 t ha^{-1}} and have a horizontal function of yield for ET greater than 300 mm ha^{-1}. However, the results from the test show evidence of a boundary only in the left and not in the right sections, this indicates that it may be better to fit a linear model of increasing yield with increasing ET without the horizontal section as the data may not have reached the point of constant yield with increasing ET (Fig. 4(a)). This agrees with the model that was suggested by the authors that used this dataset in previous boundary line analysis study (FAO and DWFI, 2015) . Conversely, if you take dataset 2 (Fig. 3(b)), which shows evidence of a boundary in both the left and right sections, the broken stick model might be a better model. The left side of the scatter shows a more linear relationship from the point {1.8 mg kg^{-1}, 11 t ha^{-1}} up to the point {2.5 mg kg^{-1}, 13 t ha^{-1}} while the right side shows more of a flat relationship between the yield and the log-transformed P concentration (Fig. 4(b)), hence, a broken stick model that consist of a linear and plateau component would be ideal for this dataset. This agrees with the model that was suggested by the authors that used this dataset (Lark et al., 2020). The decision of selecting an appropriate model should, however, be made by taking into account other considerations like the theoretical basis and plausibility of the suggested model. Although the boundary points used to check for the bounding effects are not necessarily the points to which a boundary line is to be modelled, the coefficients of a model fit to these points can provides the initial starting values (coefficients) for fitting statistical boundary line models like the bivariate censored model proposed by Milne et al. (2006a).

The exploratory method we propose is intended for analysis of biological data sets where one factor is thought to limit the response of another, for example crop yield, in response to a soil nutrient concentration. However, we recognize that the number of peels used for this analysis, which was set to 10 as default, might not be possible for some datasets, especially those containing fewer data points. For such datasets, the number of peels tested may be reduced to an appropriate number else the whole dataset may be considered as boundary points.

This study presents some novel results on the sample size required for a boundary line analysis. These are based on a particular hypothetical scenario. From the power analysis on the simulated data set, the data size required to achieve a power of 80% is affected by the measurement error and concentration of points at the boundary. The larger the concentration of data points near the boundary, the larger the power to detect the bounding effect. Conversely, as the measurement error increases, the power to detect the boundary reduces. Large measurement error obscures the boundary, and so reduces power. In the simulated case, between 650 to 800 data are required to detect a boundary in a dataset ($p \leq 0.05$) where 10 to 30% of data are on the boundary (apart from measurement error) and the measurement error is 1%–10% of the mean of response variable. This indicates that boundary line analysis is a tool for analysis of 'big-data'. In real cases sample size might be investigated based on an estimate of measurement error and a prior view on the proportion of sites at which a factor should be limiting in order to be of practical relevance. As with other statistical hypothesis testing methods, the larger the data size, the better as it reduces the margin of error and increases the reliability of the results. The method of Milne et al. (2006b), which is based on the number of vertices in a peel, may also be used as a complementary test for smaller datasets.

5. Conclusion

We provide an exploratory tool for determining evidence of the existence of a boundary in a dataset which also gives guidance to the suggestion of an appropriate type of model that one can fit a dataset. This tool provides an objective test for plausibility of the boundary model and therefore, the basis for fitting boundary line to a dataset and interpret them biologically. This has been a missing element in most boundary analysis procedures. This methodology additionally enables the selection of the starting values for fitting boundary line model when using the bivariate censored model. Simulation studies on this methodology show that several hundred observations are required for this method. Given our observation that a data set must be large enough to exhibit the boundary, this is not surprising and emphasizes that boundary line analysis is a tool for the assessment of big data sets. We recommend further works to improve the power analysis methodology we have proposed by accessing other factors that affect the effect size in boundary detection. Data sets 1 and 2 which have been used for boundary line analysis in previous studies were confirmed to show evidence of a boundary and the boundary line model forms fitted are in agreement to what the results of our exploratory analysis suggest. We recommend that future boundary analysis studies should carry out this initial exploratory data analysis step so as to justify the of the fitting boundary line models to data if there is evidence of bounding effect.

CRediT authorship contribution statement

C. Miti: Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **A.E. Milne:** Methodology, Supervision, Writing – review & editing. **K.E. Giller:** Supervision, Writing – review & editing. **V.O. Sadras:** Resources, Writing – review & editing. **R.M. Lark:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

We acknowledge Rothamsted Research, Silsoe Research Institute and AgSpace Agriculture limited for allowing us to use the various datasets in this study. Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom. AEM acknowledges support from the Growing Health Institute Strategic Programme (BB/X010953/1; BBS/E/RH/230003C). We further acknowledge the University of Nottingham Future Food Research Beacon and Rothamsted Research Programme in International Agricultural Development for providing the funding for this Ph.D. studentship.

Appendix A. Bagplots for the variables in datasets 1 to 7

See Fig. A.1.

Appendix B. Histograms for the variables in datasets 1 to 7

See Figs. B.1–B.7.

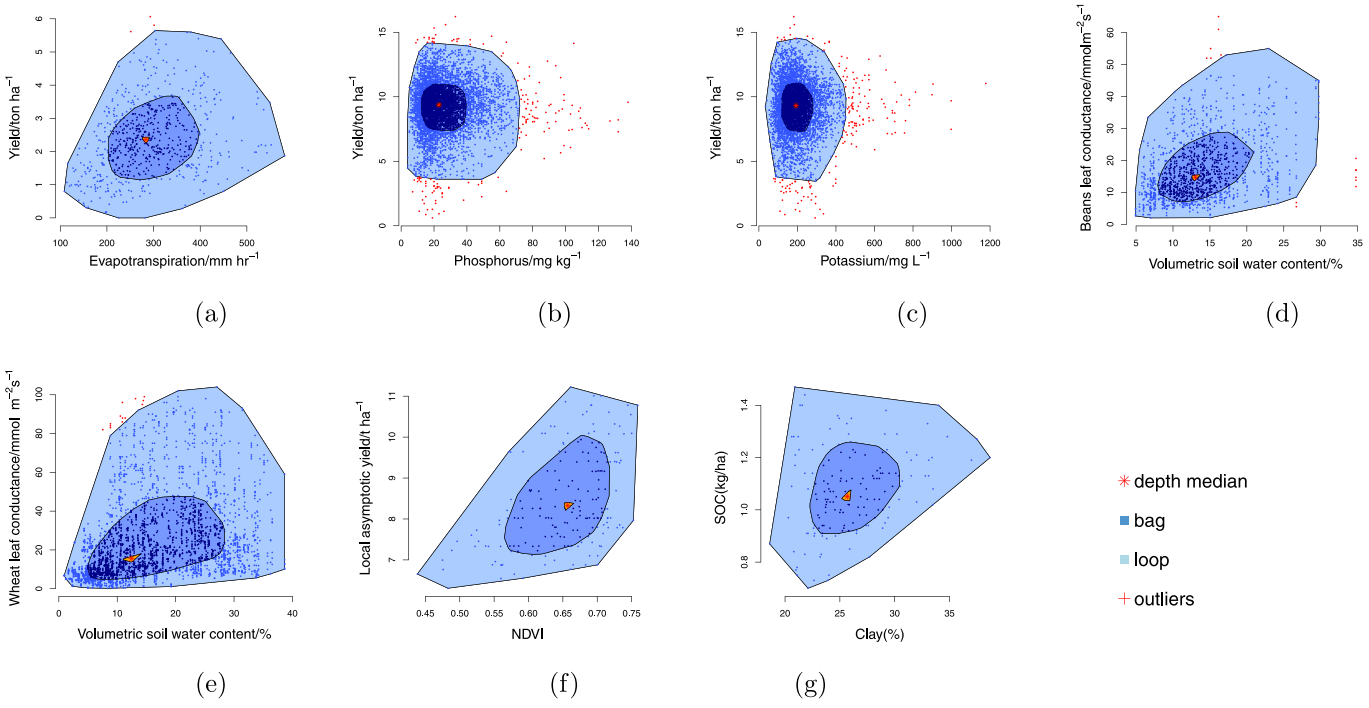

Fig. A.1. Bagplots of the datasets (a) wheat yield against evapotranspiration, (b) wheat yield against phosphorus, (c) wheat yield against potassium and, (d) beans stomatal conductance against soil water content, (e) wheat stomatal conductance against volumetric water content, (f) local asymptotic yield against NDVI, and (g) SOC against soil clay content.

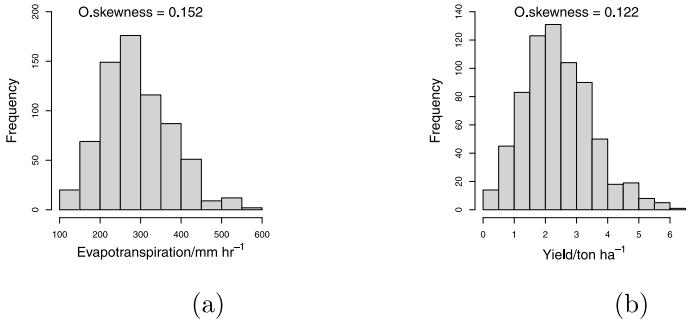

Fig. B.1. Histograms of the (a) evapotranspiration and (b) wheat yields from dataset 1.

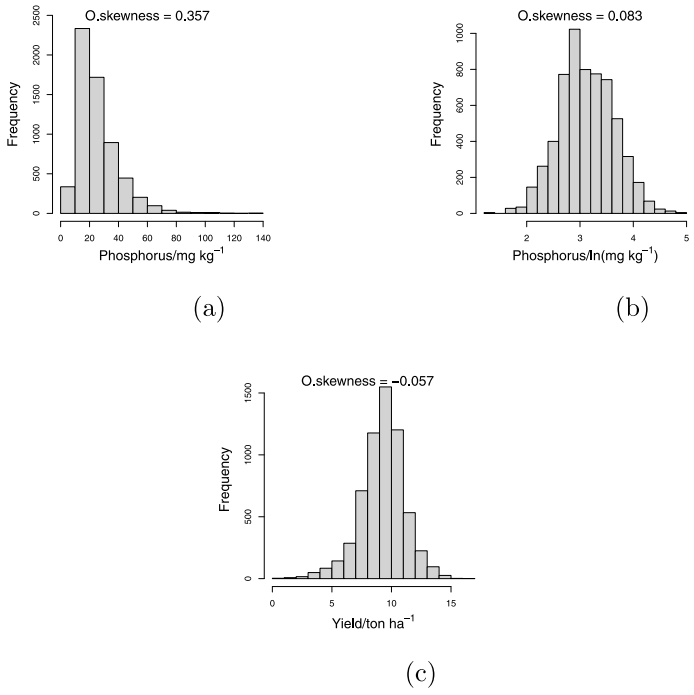

Fig. B.2. Histograms of the (a) soil phosphorus concentration, (b) log soil phosphorus concentration and (c) yield from dataset 2.

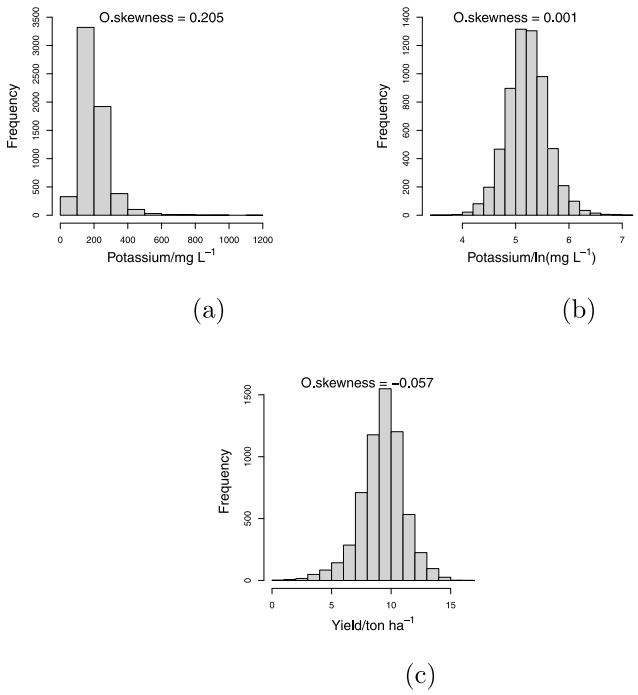

Fig. B.3. Histograms of the (a) soil potassium concentration, (b) log soil potassium concentration and (c) yield from dataset 3.

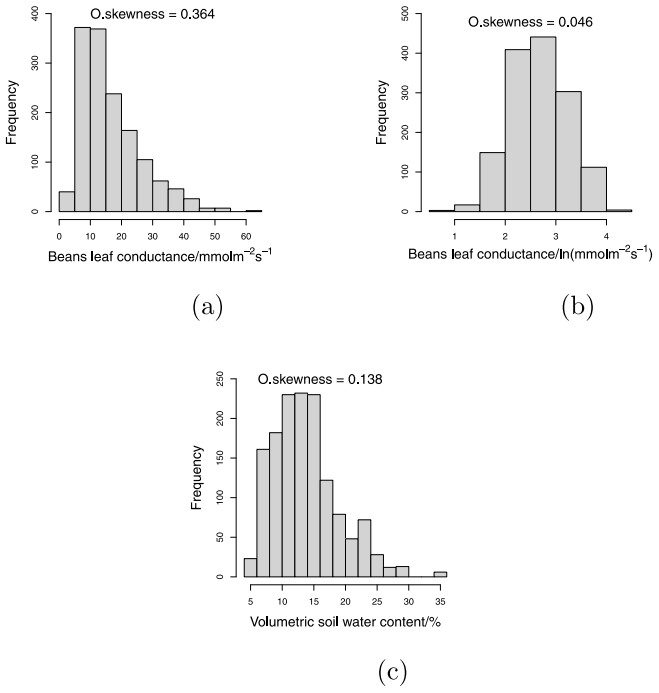

Fig. B.4. Histograms of the (a) bean leaf conductance, (b) natural log of bean leaf conductance and (c) soil volumetric moisture content from dataset 4.

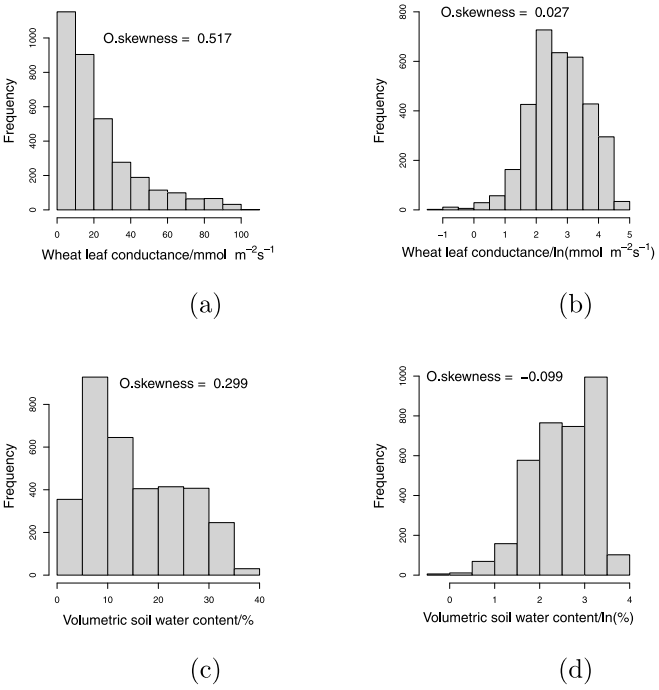

Fig. B.5. Histograms of the (a) Wheat leaf conductance, (b) natural log of wheat leaf conductance, (c) soil volumetric moisture content and (d) natural log of soil volumetric moisture content from dataset 5.

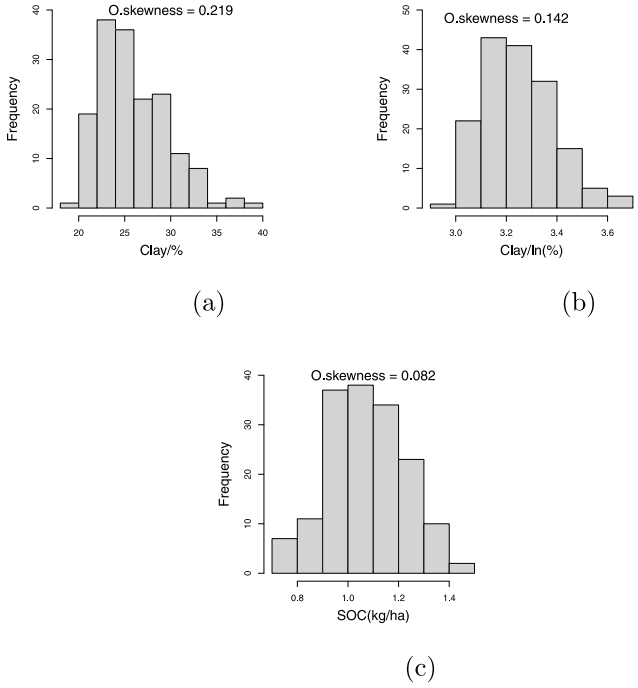

Fig. B.6. Histograms of the (a) soil clay content, (b) natural log of clay content and (c) soil organic content from dataset 6.

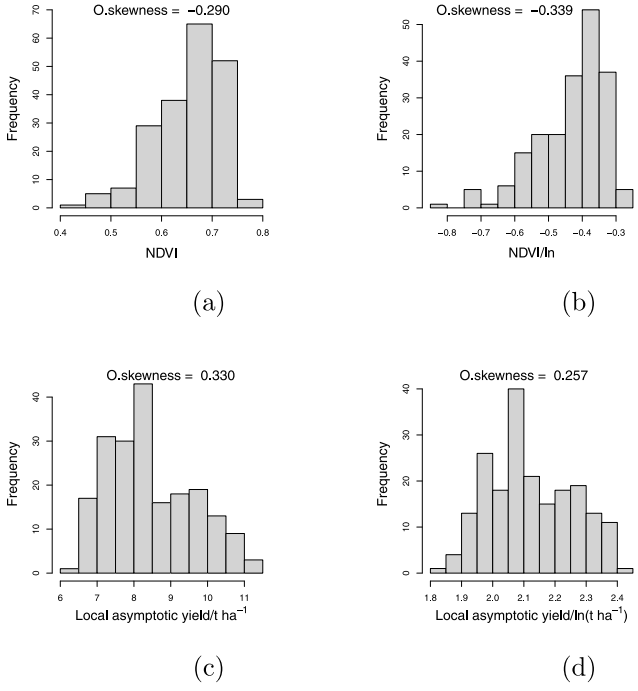

Fig. B.7. Histograms of the (a) NDVI, (b) natural log of NDVI, (c) local asymptotic yield and (d) natural log of local asymptotic yield from dataset 7.

References

Blaker, H., 2000. Confidence curves and improved exact confidence intervals for discrete distributions. Canad. J. Statist. 28 (4), 783–798.

Brys, G., Hubert, M., Struyf, A., 2002. A comparison of some new measures of skewness. In: Dutter, R., Filzmoser, P., Gather, P. (Eds.), Developments in Robust Statistics. Physica-Verlag, Heidelberg, pp. 98–113.

Buckley, T., 2017. Modelling stomatal conductance. Plant Physiol. 174, 572–582.

Casanova, D., Goudriaan, J., Bouma, J., Epema, G.F., 1999. Yield gap analysis in relation to soil properties in direct-seeded flooded rice. Geoderma 91, 191–216.

Cossani, C.M., Sadras, V.O., 2018. Water–nitrogen colimitation in grain crops. Adv. Agron. 150, 231–274. http://dx.doi.org/10.1016/bs.agron.2018.02.004.

FAO, DWFI, 2015. Yield gap analysis of field crops: Methods and case studies, by Sadras, V.O.,Cassman, K.G.G.,Grassini, P., Hall, A.J., Bastiaanssen, W.G.M., Labrte, A.G., Milne, A.E., Sileshi, G., Steduto, P.. FAO Water Rep. 41.

Fermont, A.M., van Asten, P.J., Tittonell, P., van Wijk, M.T., Giller, K.E., 2009. Closing the cassava yield gap: An analysis from smallholder farms in East Africa. Field Crops Res. 112, 24–36. http://dx.doi.org/10.1016/j.fcr.2009.01.009.

Foulkes, M., Reynolds, M., 2015. Breeding challenge: improving yield potential. In: Sadras, V.O., Calderini, D.F. (Eds.), Crop Physiology, Second Edition Elsevier, ISBN: 978-0-12-417104-6, pp. 397–421,

French, R.J., Schultz, J.E., 1984. Water use efficiency of wheat in a mediterranean-type environment 1. The relation between yield, water use and climate. J. Agric. Res. 35, 743–764.

Kintché, K., Hauser, S., Mahungu, N.M., Ndonda, A., Lukombo, S., Nhamo, N., Uzokwe, V.N., Yomeni, M., Ngamitshara, J., Ekoko, B., Mbala, M., Akem, C., Pypers, P., Matungulu, K.P., Kehbila, A., Vanlauwe, B., 2017. Cassava yield loss in farmer fields was mainly caused by low soil fertility and suboptimal management practices in two provinces of the Democratic Republic of Congo. Eur. J. Agron. 89, 107–123. http://dx.doi.org/10.1016/j.eja.2017.06.011.

Lark, R.M., Gillingham, V., Langton, D., Marchant, B.P., 2020. Boundary line models for soil nutrient concentrations and wheat yield in national-scale datasets. Eur. J. Soil Sci. 71, 334–351. http://dx.doi.org/10.1111/ejss.12891.

Lark, R.M., Milne, A.E., 2016. Boundary line analysis of the effect of water-filled pore space on nitrous oxide emission from cores of arable soil. Eur. J. Soil Sci. 67, 148–159. http://dx.doi.org/10.1111/ejss.12318.

Lark, R.M., Wheeler, H.C., 2003. A method to investigate within-field variation of the response of combinable crops to an input. Agron. J. 95 (5), 1093–1104.

Lavoie-Lamoureux, A., Sacco, D., Risse, P.-A., Lovisolo, C., 2017. Factors influencing stomatal conductance in response to water availability in grapevine: a meta-analysis. Physiol. Plant. 159 (4), 468–482.

Martin, A.E., Reeve, R., 1955. A rapid manometeic method for determining soil carbonate. Soil Sci. 79 (3), 187–198.

Mecklin, C.J., Mundfrom, D.J., 2005. A Monte Carlo comparison of the type I and type II error rates of tests of multivariate normality. J. Stat. Comput. Simul. 75 (2), 93–107.

Milne, A.E., Ferguson, R.B., Lark, R.M., 2006a. Estimating a boundary line model for a biological response by maximum likelihood. Ann. Appl. Biol. 149, 223–234. http://dx.doi.org/10.1111/j.1744-7348.2006.00086.x.

Milne, A.E., Wheeler, H.C., Lark, R.M., 2006b. On testing biological data for the presence of a boundary. Ann. Appl. Biol. 149, 213–222. http://dx.doi.org/10.1111/j.1744-7348.2006.00085.x.

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis, vol. 4, Cambridge University Press.

Powlson, D., 1994. Quantification of nutrient cycles using long-term experiments. In: Long-Term Experiments in Agricultural and Ecological Sciences. CAB International, Wallingford, UK, pp. 97–115.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/.

Rawlins, B.G., Lark, R.M., O'Donnell, K.E., Tye, A., Lister, T.R., 2005. The assessment of point and diffuse soil pollution from an urban geochemical survey of Sheffield, England. Soil Use Manage. 21, 353–362.

Rousseeuw, P.J., Ruts, I., Tukey, J.W., 1999. The bagplot: a bivariate boxplot. Amer. Statist. 53 (4), 382–387, Retrieved from https://doi.org/10.1080/00031305.1999.10474494.

Sadras, V.O., 2020. On water-use efficiency, boundary functions, and yield gaps: French and Schultz insight and legacy. Crop Sci. 60, 2187–2191. http://dx.doi.org/10.1002/csc2.20188.

Sadras, V.O., Angus, J.F., 2006. Benchmarking water-use efficiency of rainfed wheat in dry environments. Aust. J. Agric. Res. 57, 847–856, Retrieved from https://doi.org/10.1071/AR05359.

Scheiner, S.M., Gurevitch, J., 2001. Design and Analysis of Ecological Experiments. Oxford University Press.

Schmidt, U., Thöni, H., Kaupenjohann, M., 2000. Using a boundary line approach to analyze N_2O flux data from agricultural soils. Nutr. Cycl. Agroecosyst. 57, 119–129.

Shao, W., Li, M., Su, Y., Gao, H., Vlček, L., 2023. A modified jarvis model to improve the expressing of stomatal response in a beech forest. Hydrol. Process. 37 (8), e14955.

Singh, M., Sarkar, B., Sarkar, S., Churchman, J., Bolan, N., Mandal, S., Menon, M., Purakayastha, T.J., Beerling, D.J., 2018. Stabilization of soil organic carbon as influenced by clay mineralogy. Adv. Agron. 148, 33–84.

Skiena, S.S., 2008. The Algorithm Design Manual. Springer.

Su, N., Zhao, Y., Ding, G., Duan, W., 2022. Relationships between key dryland ecosystem services: A case study in ordos, China. Front. Earth Sci. 10, http://dx.doi.org/10.3389/feart.2022.937491.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer, New York, Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/, ISBN 0-387-95457-0.

Wairegi, L.W., van Asten, P.J., Tenywa, M.M., Bekunda, M.A., 2010. Abiotic constraints override biotic constraints in East African highland banana systems. Field Crops Res. 117, 146–153. http://dx.doi.org/10.1016/j.fcr.2010.02.010.

Watts, C., Clark, L., Poulton, P., Powlson, D., Whitmore, A., 2006. The role of clay, organic carbon and long-term management on mouldboard plough draught measured on the Broadbalk wheat experiment at Rothamsted. Soil Use Manage. 22 (4), 334–341.

Webb, R.A., 1972. Use of the Boundary Line in the analysis of biological data. J. Hortic. Sci. 47, 309–319.