Journal of Agricultural Science (2004), **142**, 553–560. © 2004 *Cambridge University Press*
doi:10.1017/S0021859604004642 *Printed in the United Kingdom*

Boundary-line analysis of field-scale yield response to soil properties

T. M. SHATAR AND A. B. MCBRATNEY*

Australian Centre for Precision Agriculture, Faculty of Agriculture, Food & Natural Resources, McMillan Building A05, The University of Sydney, Sydney, NSW 2006, Australia

(*Revised MS received 25 August 2004*)

SUMMARY

An algorithm to fit boundary lines, using cubic smoothing splines, was written and used to identify yield responses to changes in soil properties. This method involves fitting a curve that represents the maximum yield response to each predictor value, which represents the yield potential at each soil property value. Boundary-line yield responses to individual soil properties were found to differ from responses found by fitting curves through the data scatter. The effects of correlated variables appeared to be lessened using the boundary line approach. Multivariate boundary-line models, based on the Law of the Minimum, were found to be useful for the identification of site-specific causes of yield variation and yield potentials. The boundary line was found to be a useful complement to more traditional data analysis techniques.

INTRODUCTION

Empirical or statistical modelling is commonly used to investigate yield responses to changes in the crop-growing environment. However, regression techniques provide only limited insight into observed relationships; it has proven difficult to separate responses to causal factors from responses to variables correlated with those causal factors in both single- and multiple-predictor models. Another limitation is that regression through the data scatter represents the average response. While this is appropriate for traditional, uniform management which is based on averages (Lark 1997), increasingly, site-specific yield responses are of interest.

An alternative to traditional, statistical models, the boundary line, was first presented by Webb (1972). It facilitates isolation of single-factor yield responses from data in which yields have been affected by multiple factors and does not represent merely the average response. Rather than fitting regression lines through the data scatter, Webb's approach is to fit a line above the scatter of the data points. This line represents the maximum potential yield, or best performance, for that input level. It is assumed (for a sufficiently large dataset) that these are the maximum potential yields in the absence of any other limiting factors (Elliott & de Jong 1993) and that any points falling below it are limited by another variable. In particular, by identifying the maximum obtainable yields at a range of predictor values, yield potentials can be identified. The yield potential is an important concept in site-specific management in that it represents the maximum yield that can be obtained at a site, given the constraints to production imposed by unmanageable, yield-affecting factors. It therefore plays a significant role in determining management strategies.

Lark (1997) has suggested using the boundary line for analysis of site-specific data collected from growers' fields and other authors have also pointed out its usefulness for identifying yield response to single factors from data in which yields have been affected by multiple factors.

A major impediment to wide-spread adoption of the boundary-line technique as a data analysis tool is the difficulty of fitting boundary-line curves and the lack of reproducibility of results (Schnug *et al*. 1996). Authors who have used this kind of analysis have usually fitted models by eye (e.g. Webb 1972) and commonly drawn the boundary line by hand (Schnug *et al*. 1996). Fixen & Grove (1990) suggested that the response could be better elucidated by fitting the boundary line according to a statistical procedure, which would also enable reproducibility of results.

* To whom all correspondence should be addressed. Email: Alex.McBratney@acss.usyd.edu.au

https://doi.org/10.1017/S0021859604004642 Published online by Cambridge University Press

Several authors have developed their own algorithms for fitting boundary lines, since standard statistical packages do not provide a means of fitting a curve to the maximum response. In the present study too, an algorithm was written to fit the boundary line. However, the methodology used to fit the boundary line is unlike those used in previous studies. In the absence of a theoretical framework for evaluating yield responses to environmental factors, the use of splines to fit the boundary line was explored in the present study.

The specific objectives were to:

(1) develop an algorithm to fit boundary lines to continuous yield responses using splines,
(2) use boundary-line analyses to
   a. interpret field-scale yield responses to identify yield-maximizing optimum values and identify the environmental factors responsible for yield variation
   b. identify yield potentials as determined by environmental factors.

MATERIALS AND METHODS

The dataset used to illustrate the boundary-line methodology was described by Shatar & McBratney (1999). Yield and soil data were collected from a 17 ha field on a commercial farm in Moree, in northern New South Wales (NSW), Australia. Soil sampling locations are shown in Fig. 1*a*. Yield data were collected in 1996 during harvest of a sorghum crop by continuous monitoring. Yield was monitored continuously using an AgLeader® mass flow impact yield monitor in conjunction with a Fugro Starfix® real-time differential global positioning system (GPS). Yield data were processed and predicted onto a 5 m grid using local ordinary kriging with a 20 m block size and an exponential variogram. A map of sorghum yields is presented in Fig. 1*b*.

Using the available yield data, soil sampling was targeted to span the entire range of observed yields. Soil samples were collected immediately following sorghum harvest in 1996 and analysed for physical and chemical properties, including moisture-holding capacity (w) at potentials of -33 kPa and $-1{\cdot}5$ MPa, pH, organic carbon content (OC), available P, cation exchange capacity (CEC) and exchangeable Ca, Mg, Na and K. The Ca/Mg ratio and exchangeable sodium percentage (ESP) were calculated. Soil data from a total of 110 locations were available for analysis.

Unless otherwise stated, all statistical analyses were performed using S-PLUS statistical software (Statistical Sciences 1995). Kriging was performed using the program VESPER (Minasny *et al*. 1999) and maps were produced using ArcGIS software (ESRI 2001).

Fitting methodology

Most of the procedures in the literature used to fit boundary lines include the following steps:

– grouping of data points according to their predictor values,
– removal of outliers,
– identification of the "maximum-yield subset"; the data subset representing maximum yields at each predictor value and
– fitting of a curve to the maximum-yield subset.

The fitting of the boundary line necessarily involves the subdivision of data into discrete groups in which the predictor values are the same and the greatest yield value observed can be chosen for inclusion in the data subset to which the boundary line is fit. Subdivision was relatively simple for Webb (1972), since the study involved achene counts. The data therefore naturally fell into discrete groups. In contrast, continuous yield responses are not naturally grouped, yet fitting of the boundary line has required subdivision of the x-axis into artificial categories. For example, Casanova *et al*. (1999) subdivided the x-axis into ten arbitrary, hard groups.

The removal of outliers is critical to boundary-line fitting, even more so than in traditional regression analysis. The position of the final boundary line is dependent upon a relatively small amount of data, therefore the presence of outliers can have a greater impact on the final model than if it were generated by a traditional least-squares approximation. Both Webb (1972) and Schnug *et al*. (1995) selected the data points to be included in boundary-line analysis by comparing yields within each predictor group and by comparison with successive maximum yield values, discarding those that did not fit the observed trend. This method assumes that the data should follow some pre-determined shape and resulted in missing yield values for some predictor values.

Identification of the maximum yield at each predictor value has also been approached in different ways. Webb (1972) and Schnug *et al*. (1996) selected the highest-yielding data point within a group. This method is highly dependent on the way in which the "group" is defined and, as stated previously, x-axes have typically been arbitrarily subdivided. Casanova *et al*. (1999) calculated the upper 95% confidence interval of the distribution of yield in each category and used this value as the maximum yield for the group. This approach removes outliers but makes assumptions about the shape of the yield distribution and assumes that sufficient data are available within each group to estimate this distribution.

The final fitting of a curve to the boundary-line curve has commonly been done by hand. When statistical models have been used, they have been limited to linear (straight-line) regression (Casanova *et al*.

https://doi.org/10.1017/S0021859604004642 Published online by Cambridge University Press

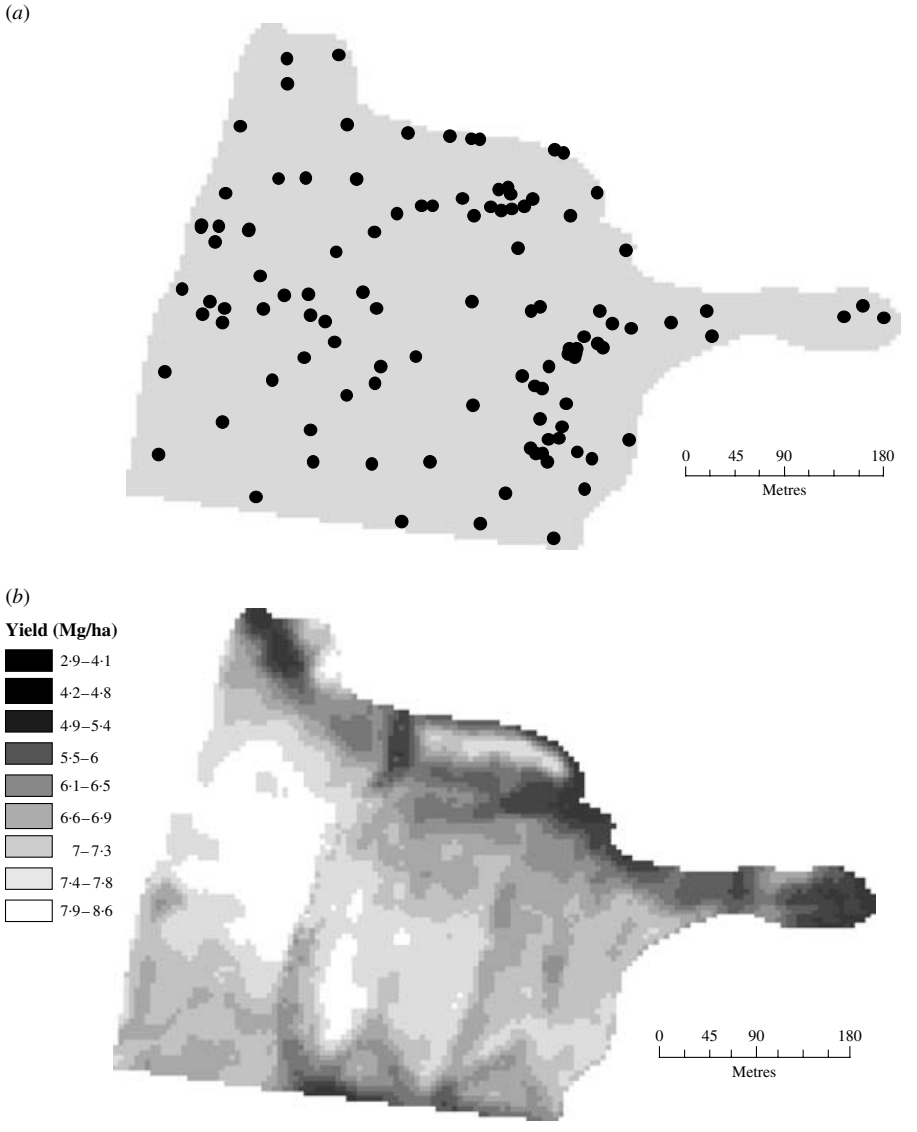


Fig. 1. (*a*) Map of soil sampling locations and (*b*) map of 1996 sorghum yields within the study field.

1999) or quartic polynomials (Schnug *et al*. 1995, 1996). By adopting more flexible curves, it may be possible to better represent the diversity of yield-response shapes reported in the literature.

The basic steps used by previous researchers to fit boundary lines were followed here. However, different methods were used.

Mahalanobis outliers were identified from the dataset as a whole. Those data points with the largest 5% of Mahalanobis distances (Mardia *et al*. 1979) were removed from the dataset before the fitting of the boundary line.

When dealing with prediction from continuous variables, if too many subdivisions are created, each single data point potentially represents the maximum yield at that predictor value. If too few subdivisions are created the underlying trend may be overlooked. This problem is akin to the problem of selecting a suitable degree of smoothing when using smoothers to analyse data. In effect, a smoothing approach was used to extract a data subset to which the boundary line was fitted. For each unique predictor value, data points falling within a certain range of that value were subset and the maximum response value identified.

https://doi.org/10.1017/S0021859604004642 Published online by Cambridge University Press

Essentially, a moving window was placed on each unique predictor value. The size of the window was fixed but there were no hard divisions between predictor values since there was a great deal of overlap in the data subsets used to identify maximum yields. The size of the window could be modified but was set at a default of one tenth of the data range.

A spline (with four degrees of freedom) was then fitted to the data subset. The method of generating the data subset resulted in it containing as many observations as there were unique predictor values. In many cases, the final dataset had a similar number of observations as the initial dataset since the predictor variables were continuous.

Boundary-line analysis of yield response

Single factor boundary-line models of yield response were created. The range of the boundary-line model was limited to the range of the predictor values because extrapolation of non-linear models can be very unreliable. The robustness of the modelling technique and its sensitivity to slight changes in the dataset was assessed empirically using bootstrapping (Efron & Tibshirani 1993). The original dataset was resampled using bootstrapping; individual data points were selected randomly, with replacement, to create new datasets with the same number of observations as the original. This was repeated 1000 times (as suggested by Efron & Tibshirani 1993) and a boundary line was fitted to each dataset. The average prediction and 95% confidence intervals were estimated from the distribution of results and plotted.

The relative value of a second environmental variable was represented on the graph by the size of the data points. This may aid in identification of the factors that cause yields to fall below the boundary line, that is, below the potential yield that could be reached at each predictor value if no other variables imposed a limitation on yield.

Identification of the site-specific factors limiting yield production within a field

Boundary lines were fitted to each yield and predictor variable combination. These individual boundary-line responses were combined in order to create a multivariate model, in the manner of Casanova *et al.* (1999), which assumes a von Liebig-type response. That is, the model assumes that the response of the crop to growth factors is based on the Law of the Minimum popularized by Justus von Liebig (von Liebig 1863). The Law states that the amount of crop growth or yield is determined by the yield-influencing factor present in the (relative) minimum amount; that yield will vary with changes in this attribute, until it is no longer the limiting factor, and that changes in other yield-influencing factors will not affect yield.

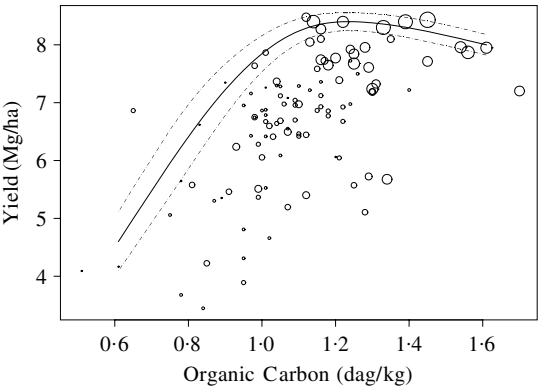


Fig. 2. Sorghum yield response to soil OC content. Circle sizes are proportional to soil exchangeable K content (mg/kg). Dotted lines indicate 95% confidence intervals.

The minimum yield predicted by any of the individual, single-response boundary lines at each location was taken to be the yield prediction at that location.

A map of the differences between yield predictions (yield potentials) and actual yield (yield prediction − actual yield) at all locations was created by calculating differences at sampled locations and interpolating results onto a 5 m grid, using block kriging. A map of the variables limiting production at each location was created by creating a 5 m grid of the study area and assigning to each raster cell the limiting factor of the nearest sampled location.

RESULTS

The data scatter in Fig. 2 indicates that sorghum yield values increased over the range of measured OC contents and the boundary line shows a strong, initial positive response to increases in soil OC content. However, the boundary line also shows that this response tapered off as OC contents of 1·2 dag/kg were approached and that there was some indication that yields were depressed very slightly as OC contents rose further. The confidence intervals show that there was little deviation from the average response. The size of the data points are proportional to soil exchangeable K content and show that higher-yielding sites had higher soil exchangeable K content; points falling closer to the boundary line were generally higher in K than those not meeting their yield potential determined by the soil OC content. The graph also shows a positive relationship between soil OC and K contents. These results indicate that yields at locations with soil OC contents greater than 1·2 dag/kg did not vary greatly due to OC levels and that much of the apparent relationship between yield and OC was really attributable to increases in other

https://doi.org/10.1017/S0021859604004642 Published online by Cambridge University Press

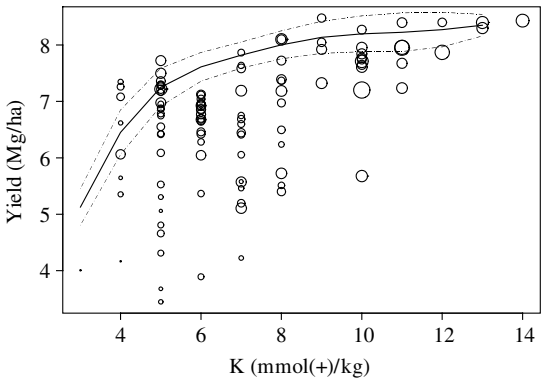


Fig. 3. Sorghum yield response to soil exchangeable K content. Circle sizes are proportional to soil organic carbon content (dag/kg). Dotted lines indicate 95% confidence intervals.

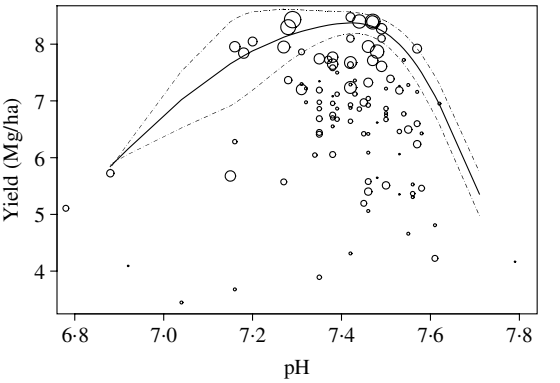


Fig. 4. Sorghum yield response to soil pH. Circle sizes are proportional to soil exchangeable K content (mmol(+)/kg). Dotted lines indicate 95% confidence intervals.

environmental factors, correlated to soil OC content, such as nutrient levels or water availability.

Similarly, Fig. 3 shows that the relationship between yield and soil exchangeable K content appeared to be strongly positive over the entire range of measured soil K contents. The boundary line indicates that the strongest yield response occurred where soil K contents were less than 6 mmol (+)/kg. At higher soil K contents, yield responses were less pronounced. Again, the confidence intervals indicate that the model was relatively robust. The size of the data points are proportional to the soil OC content. The effects of soil OC content in reducing yields below soil exchangeable K potentials are not obvious from the graph, perhaps because of the relatively small yield response to changes in soil OC content above 1·2 dag/kg; a very large OC content was not necessary to achieve higher yield potentials.

The boundary-line model of sorghum yield response to soil pH (Fig. 4) indicates that yields were maximized between pH values of approximately 7·3 and 7·5. Outside this range, yields exhibited a strong response to changes in soil pH. However, the model was greatly affected by small changes in the dataset at pH values below the optimum and results are likely to be unreliable within this range. The confidence intervals were widest at pH values where relatively few data points were available and where there was great variability in yield at those few points. The size of the data points are proportional to soil K content and indicate that sites that had greater soil exchangeable K content were found closer to the boundary line.

A soil Fe content of approximately 8 mg/kg appeared to be optimum for sorghum yield in this field (Fig. 5). Beyond this optimum, yields declined slightly. Fe deficiency is known to occur in calcareous, high pH soil, such as those types found in the study field (Berg *et al.* 1993); uptake is known to be

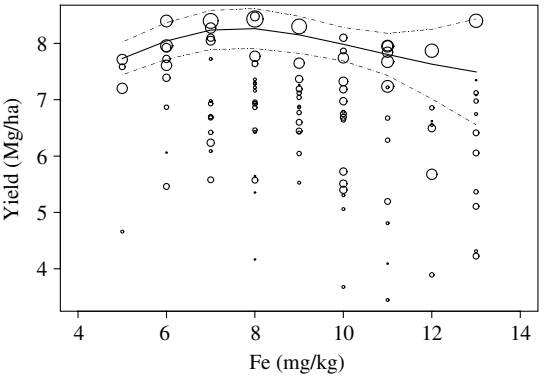


Fig. 5. Sorghum yield response to soil Fe content. Circle sizes are proportional to soil exchangeable K content (mmol(+)/kg). Dotted lines indicate 95% confidence intervals.

adversely affected under these conditions (Zaharieva & Romheld 1991). However, higher soil Fe contents have been associated with reduced uptake of exchangeable cations and other micronutrients (Kashirad *et al.* 1978). This may explain the shape of the response curve. The response observed was confined to only those sites which were already high yielding. The range of soil Fe values was sufficient to cause yield variation between approximately 7·5 and 8 mg/ha but does not explain why lower yields may have occurred; soil Fe contents did not prevent yields from reaching over 7·5 mg/ha at any site within the field.

In this case, the confidence intervals indicate more variation in model results, and therefore poorer reliability of the boundary line, at the largest predictor values. The size of the data points is indicative of soil exchangeable K content. A clear trend is evident; data points closer to the Fe boundary line had higher

https://doi.org/10.1017/S0021859604004642 Published online by Cambridge University Press

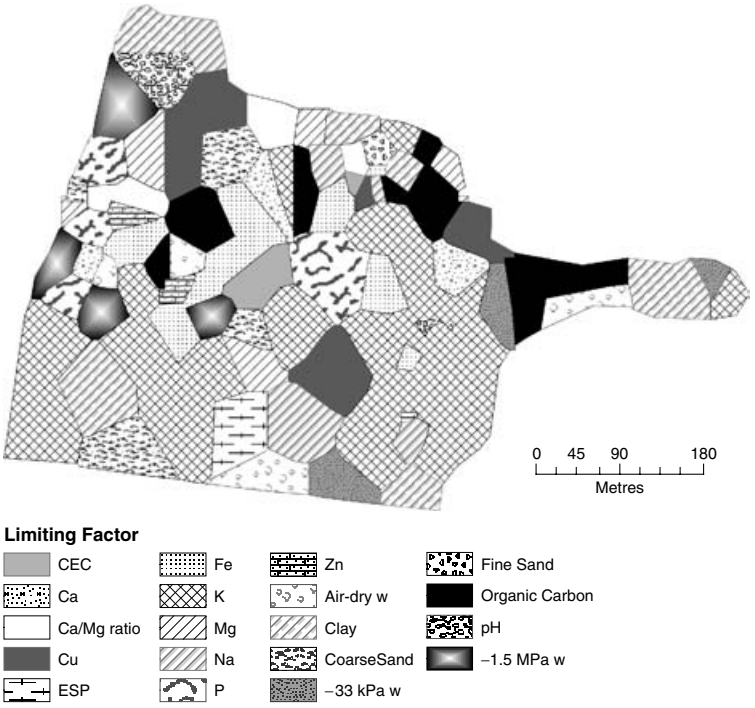


Fig. 6. Yield-limiting soil factors, as determined from the multivariate boundary-line model.

exchangeable K contents. There appears to be no trend between soil exchangeable K and Fe.

Figure 6 shows the predictor variables that limited yield at each sampled location in the field, as determined by the multivariate boundary-line analysis. Soil exchangeable K content appears to have been the major yield-limiting factor and affected yields within a large area of the field, concentrated in the southern portion. Soil Fe and OC contents also appeared to affect yields at a number of locations. Soil Fe content appeared to limit yields at higher yielding sites and may indicate interference with K uptake.

Figure 7 shows that the yield potentials determined by the multivariate boundary line model were not attained near field boundaries. Perhaps other, unmeasured factors caused yield reductions in these areas. In the high-yielding area near the western boundary, yields were somewhat under-predicted. This may indicate that the assumptions of boundary-line analysis were not met; the boundary-line necessarily assumes that the highest yields at each soil value represent the maximum obtainable yields. It may also be indicative of error in the boundary-line fit or of the von Liebig hypothesis.

DISCUSSION

The results presented show promise for the use of the boundary-line technique for analysis of yield responses and that a boundary-line analysis has some advantages over other techniques. For example, because the multivariate model considers each point in the field separately, the results are truly site-specific; potential yields are a function of the most yield-limiting factor at that site rather than being based on the whole-field trend of a combination of variables.

Unlike most multivariate models, the form of the specification of the multivariate boundary-line model ensures results are easy to interpret because at each single location a single variable responsible for yield variation is identified. However, while this makes interpretation easier, it may be an over-simplification of the reality of the response and ignores interactions between variables. This was evident in the difference map which showed that there were regions within the field where yields were under-predicted by the boundary-line model.

Unlike most multivariate modelling techniques, a separate process of variable selection is not required. Because all individual yield-response functions are compared and the maximum predicted yield at each site identified, a single predictor variable is chosen at each location. In this case, selection is based upon prediction of the smallest yield value. In the resulting model, although a number of single boundary-line responses are generated and compared, any single yield prediction is a function of a single response curve as opposed to the combination of a multitude

https://doi.org/10.1017/S0021859604004642 Published online by Cambridge University Press

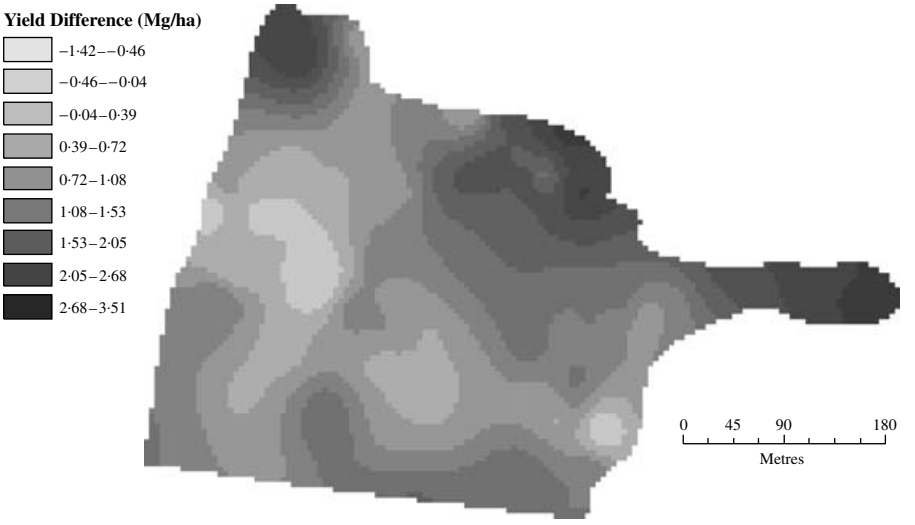


Fig. 7. Map of differences between potential sorghum yields, as predicted by the multivariate boundary-line model, and actual sorghum yields.

of predictors. As a result, the model may be considered parsimonious.

In an earlier work (Shatar & McBratney 1999), models of single and multivariate yield-responses were created from this dataset, using traditional regression approaches. These models identified whole-field trends; at any location, yield was modelled as a function of all the predictor variables and the model had the same form throughout the field.

Both methods showed a positive yield response to soil exchangeable K content. However, the boundary-line response was not as dramatic and yields increased only slightly as soil exchangeable K contents increased above 8 mmol/kg. The response fit through the data scatter showed increased yields over the entire range of measured soil K contents. Soil exchangeable K content was strongly correlated to a number of other environmental variables, including soil concentrations of other nutrients and moisture-holding capacity. This may explain why yields appeared to be more strongly influenced by changes in soil K content in the regression model than the boundary-line model. This also indicates that boundary-line responses are less sensitive to the influence of correlated variables than regression methods that fit curves through the data scatter and are therefore easier to interpret.

Shatar & McBratney (1999) concluded that yields in the field were limited primarily by soil moisture availability and somewhat by soil K content and pH. The results of the boundary-line analyses show that yields may have been affected by moisture supply at a number of locations, as reflected in responses to measurements of soil OC content, texture and soil moisture-holding capacity, but that soil nutrient availability, particularly of soil exchangeable K and Fe, also affected yields at many locations. Both methods indicated that sorghum yields were maximized within the pH range of 7·3–7·5, however, the corresponding yields within this range were significantly higher in the boundary-line model.

The differences in results obtained using the different analysis tools indicate the importance of using any data analysis technique only as an aid to interpretation. For meaningful explanations, like all results obtained using empirical techniques, results must be interpreted with knowledge of soil and agronomic principles. Schnug *et al.* (1995) also emphasized this point. While not a replacement for techniques that model the average yield response, rather than the best yield response, the boundary line is a complementary tool that may assist in the interpretation of yield-response data, particularly when datasets contain large numbers of predictor variables.

The location of the boundary line is dependent on relatively few data points so is unlikely to be very robust. The confidence intervals generated showed mixed results; in some cases, the models appeared relatively robust, in others, the confidence intervals widened dramatically, particularly where data were sparser and more variable. The boundary-line approach also assumes that the yields on the upper-boundary of the dataset are representative of the maximum attainable yields at that growth factor level, which may not be true.

It has been proposed that data from different sites and years could be pooled together and modelled with the boundary line because the effects of other variables are removed. This may be true of soil variables which are directly related to crop yield, such as

https://doi.org/10.1017/S0021859604004642 Published online by Cambridge University Press

nutrient availability. It would be expected that crop requirements of any particular nutrient, for a specific yield goal, would be fairly constant and determined by the crop genotype. However, many of the variables routinely used to investigate yield variation in site-specific studies are indirect measures of soil nutrient- and/or moisture-availability. Responses to these measures can change over time, depending on other factors. For example, yield response to topographic variables has been shown to change with variation in other environmental factors. Over 3 years, Sudduth *et al*. (1997) reported strongly negative, strongly positive and less strongly negative soybean yield responses to increased elevation. These differences were attributed to differences in water availability; in drier years, areas at low elevation benefited from additional water but, in wetter years, yields were limited by excess water.

The algorithm written to automate the fitting of boundary lines to data performed well. The graphs of the boundary line representations of yield response showed that the algorithm developed was able to fit curves on the upper-boundary of the yield response. However, the methodology developed should be tested on different datasets to further evaluate its usefulness as a data-analysis tool.

The spline was able to represent all the yield-response shapes encountered. In many cases, using the linear regression used by Casanova *et al*. (1999) would have been inappropriate and the quartic polynomial of Schnug *et al*. (1995) may not have been suited to representation of plateaux.

REFERENCES

Berg, W. A., Hodges, M. E. & Krenzer, E. G. (1993). Iron deficiency in wheat grown on the Southern Plains. *Journal of Plant Nutrition* **16**, 1241–1248.

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.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. San Francisco: Chapman and Hall.

Elliott, J. A. & de Jong, E. (1993). Prediction of field denitrification rates: a boundary-line approach. *Soil Science Society of America Journal* **57**, 82–87.

ESRI. (2001). *ArcGIS 8.2*. Redlands, California, USA: ESRI.

Fixen, P. E. & Grove, J. H. (1990). Testing soils for phosphorus. In *Soil Testing and Plant Analysis* (*3rd Edn*) (Ed. R. L. Westerman), pp. 141–180. WI, USA: Soil Science Society of America Inc.

Kashirad, A., Bassiri, A. & Kheradnam, M. (1978). Response of cowpeas to applications of P and Fe in calcareous soils. *Agronomy Journal* **70**, 67–70.

Lark, R. M. (1997). An empirical method for describing the joint effects of environmental and other variables on crop yield. *Annals of Applied Biology* **131**, 141–159.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

Minasny, B., McBratney, A. B. & Whelan, B. M. (1999). *VESPER version 1.0*. Australian Centre for Precision Agriculture, McMillan Building A05, The University of Sydney, NSW, 2006. (http://www.usyd.edu.au/su/agric/acpa)

Schnug, E., Heym, J. & Achwan, F. (1996). Establishing critical values for soil and plant analysis by means of the Boundary Line Development System (Bolides). *Communications in Soil Science and Plant Analysis* **27**, 2739–2748.

Schnug, E., Heym, J. & Murphy, D. P. (1995). Boundary line determination technique (BOLIDES). In *Site-Specific Management for Agricultural Systems* (Eds P. C. Robert, R. H. Rust & W. E. Larson), pp. 899–908. Madison, WI: ASA-CSSA-SSSA.

Shatar, T. M. & McBratney, A. B. (1999). Empirical modeling of relationships between sorghum yield and soil properties. *Precision Agriculture* **1**, 249–276.

STATISTICAL SCIENCES (1995). *S-PLUS Guide to Statistical and Mathematical Analysis*. Seattle, WI: StatSci.

Sudduth, K. A., Drummond, S. T., Birrell, S. J. & Kitchen, N. R. (1997). Spatial modeling of crop yield using soil and topographic data. In *Precision Agriculture '97, Proceedings of the 1st European Conference on Precision Agriculture* (Ed. J. V. Stafford), pp. 439–447. Oxford: BIOS Scientific Publishers.

von Liebig, J. (1863). *The Natural Laws of Husbandry*. London: Walton and Maberly.

Webb, R. A. (1972). Use of the Boundary Line in the analysis of biological data. *Journal of Horticultural Science* **47**, 309–319.

Zaharieva, T. & Romheld, V. (1991). Factors affecting cation-anion uptake balance and iron acquisition in peanut plants grown on calcareous soils. *Plant and Soil* **130**, 81–86.

https://doi.org/10.1017/S0021859604004642 Published online by Cambridge University Press