Boundary Line Methodology for Yield Gap Analysis of Farm Systems

Chawezi Miti

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

School of Biosciences
University of Nottingham

July 2025

Abstract

The growing global food demand necessitates improvements in agricultural productivity to achieve global food security. This can be attained by closing the current yield gaps in cropping systems. The boundary line methodology has emerged as a popular tool for determination of yield gaps and has been widely used in many agronomic studies on yield analysis. However, there is no standard procedure for fitting boundary line models and there is a lack of exploratory analysis tools to ascertain the suitability of data for fitting a boundary line model. Furthermore, whereas the fitting of other types of statistical model is supported by the availability of functions in statistical software no such tool is available for applying boundary line analysis. In this thesis I advance the application of the boundary line methodology in yield gap analysis. First, I undertook a comprehensive review of the use of the boundary line methodology in yield gap analysis. This review revealed inconsistencies in application, particularly in the boundary line fitting procedures. Heuristic methods, though widely used, rely on subjective decisions such as the choice of bin width and often lack a statistical basis for assessing the suitability of the boundary model, underscoring the need for standardized and reproducible approaches. My research addresses these gaps by proposing an exploratory data analysis method using convex hull peels, which assesses whether datasets exhibit boundary-limited responses. This method improves the reliability of boundary line analysis by preventing misapplications to datasets that do not support boundary constraints. I also present a comparative evaluation of multiple boundary line fitting techniques, including binning, boundary line determination technique, quantile regression, and the censored bivariate normal model. Results showed that the censored bivariate normal model provides more stable estimates of critical values than heuristic methods, enhancing the precision of yield-limiting factor identification. However, the original model lacked the ability to accommodate categorical independent variables. This limitation was addressed by developing a framework that extends the censored model to handle categorical data. In this study, data size was found to significantly influence boundary line model fitting. In many cases, smaller datasets did not provide sufficient statistical evidence to support the fitted boundary models. A significant outcome of my research was the development of the `BLA R` package, which integrates multiple boundary line methodologies, exploratory tools (including

those I developed), and interactive functions for enhanced usability. This open-source tool promotes transparency, reproducibility, and accessibility for researchers, aiding in the robust application of boundary line analysis in yield gap assessments. To test the useability of the `BLA R` package and to evaluate how users engage with various boundary line methods, I held stakeholder engagement workshops in Nairobi and Harare to elicit their opinions on preferred boundary line fitting methods. Although no specific method was favoured, participants emphasized the need for a more objective approach to fitting boundary line models in agronomic research. This research provides a critical foundation for improving yield gap assessments, supporting sustainable agricultural intensification, and contributing to global food security through data-driven decision-making.

Dedication

I dedicate this thesis to my wife, **Cecilia Chitimwango**, and my beloved daughter and son, **Walusungu Miti** and **Alinafe Miti**.

Acknowledgments

This research was supported by the Nottingham-Rothamsted Future Food Beacon International Doctoral Training Program. A massive thanks go my supervisors Prof. Richard Murray Lark, Prof Ken Giller and Dr. Alice Milne for all their efforts in guiding me throughout the process.

I extend my thanks to everyone who has been part of this journey. Joanna Smuga-Lumatz for all the assistance with the administrative part of my PhD, colleagues from CGIAR, Dr. Joao Vasco Silva and Dr. Siyabusa Mkuhlani who greatly help in organizing the workshops we conducted to test the `BLA R` package in its development phase, the School of Biosciences Research Ethics Committee, especially Professor Kate Millar, for the comments and suggestions on research ethic applications for the elicitation exercises conducted in this research, and all colleagues from the Agriculture and Environment Department of the School of Biosciences at the University of Nottingham for the help, encouragement and positive feedback on my works.

To my wife Cecilia, my daughter Walusungu Miti, my parents and siblings, thankyou for the support during the 4 years of my study. Above all, I want to thank my God for according me this opportunity to undertake and complete this PhD study.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Need for Closing Yield Gaps

The global population has been growing at an unprecedented rate, surpassing 8 billion in 2022 (United Nations, 2023). According to United Nations projections, this number could reach nearly 10 billion by 2050, driving a 60% increase in global food demand compared to 2007 levels (Alexandratos & Bruinsma, 2012). This rapid population growth underscores the urgent need to expand food production to ensure global food security.

The United Nations' Sustainable Development Goals (SDG) provide a comprehensive framework for addressing these challenges. Goal 2, 'Zero Hunger', aims to eradicate hunger, enhance food security, improve nutrition, and promote sustainable agriculture. Achieving this goal requires doubling agricultural productivity and increasing the incomes of small-scale food producers, all while ensuring sustainable food production systems. However, these efforts must contend with mounting constraints, including shrinking natural resources, climate change, and the expansion of urban areas, which is reducing the availability of arable land. One of the greatest challenges of our time is meeting the rising food demand without further encroaching on natural ecosystems such as forests, wetlands, and savannas (Rattalino Edreira et al., 2021).

One of the key strategies for addressing food security challenges is increasing agricultural productivity through improved crop varieties. Advances in biotechnology and breeding have led to the development of high-yield, drought-resistant, and pest-resistant crops, offering farmers the tools to adapt to climate change and enhance productivity sustainably. However, significant disparities persist between the potential yields of these crops and the actual yields achieved by farmers (Figure 1.1). These yield gaps, represent untapped opportunities to boost food production without expanding agricultural land. In many regions, closing these gaps could substantially increase global crop yields. For in-

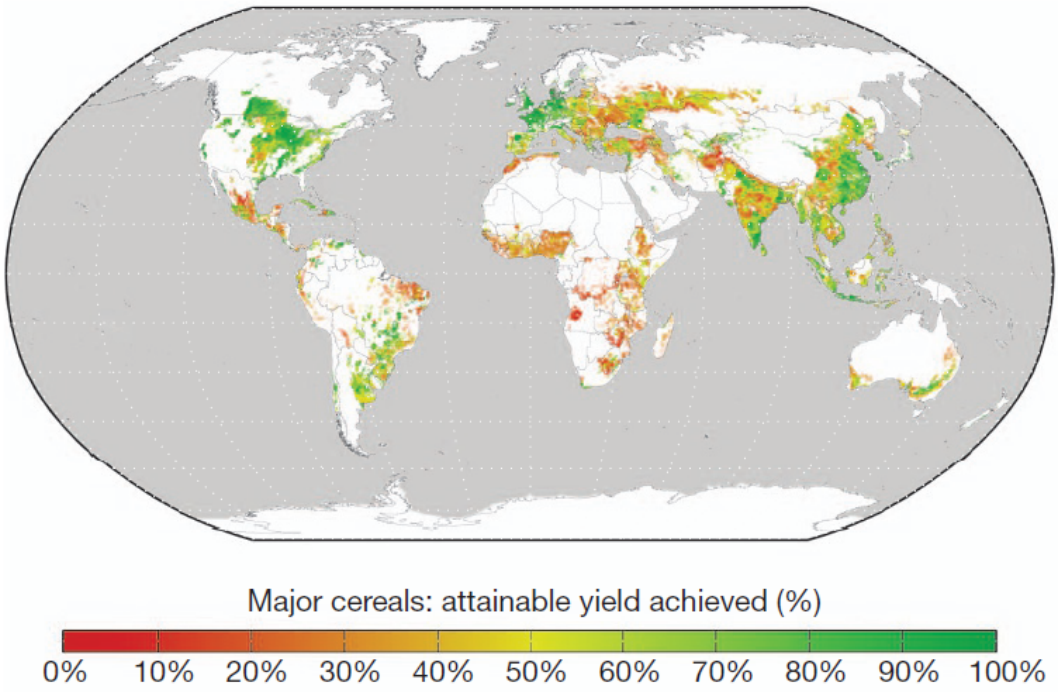


Figure 1.1: Percent of attainable yields achieved for maize, wheat and rice using the yield gap atlas (Mueller et al., 2012).

stance, Mueller et al. (2012) estimated that closing yield gaps to 100% of attainable yields could raise worldwide production by 45% to 70% for most major crops, with increases of 64% for maize, 71% for wheat, and 47% for rice.

Yield gaps are influenced by a range of factors, including suboptimal agricultural practices, limited access to high-quality inputs, pest and disease pressures, and climate variability. In sub-Saharan Africa, where smallholder farmers are central to food production, yield gaps are often driven by nutrient deficiencies (Tittonell & Giller, 2013). Many of these farmers also lack access to improved seed varieties, fertilizers, and modern irrigation technologies, which further constrains productivity. Reports indicate that small-scale farmers in sub-Saharan Africa achieve only 20–40% of the rain-fed potential yield (van Ittersum et al., 2016). Addressing these constraints presents a significant opportunity to enhance agricultural productivity, strengthen food security, and improve livelihoods both regionally and globally.

Identifying, quantifying, and understanding the factors contributing to yield gaps is therefore critical for guiding interventions that optimize agricultural systems in a holistic and sustainable manner. This process involves examining biophysical, socioeconomic, and management-related constraints that limit crop productivity, such as suboptimal soil fertility, water limitations, pest and disease pressures, inadequate access to advanced agricultural technologies, and inefficiencies in farm management practices.

Beyond boosting food production, closing yield gaps also has critical environmental

benefits. Reducing the gap between actual and attainable yields minimizes the need for agricultural expansion into forests and other ecologically sensitive areas, thereby preserving biodiversity and mitigating climate change (Phalan et al., 2014). Conversely, failing to address these gaps could exacerbate global hunger and malnutrition while accelerating the depletion of natural resources, such as soil, water, and nutrients. Unchecked agricultural expansion could further intensify ecosystem pressures, leading to habitat destruction, biodiversity loss, and increased greenhouse gas emissions (Foley et al., 2005).

Therefore, closing yield gaps is not just a matter of increasing food production, it is a fundamental strategy for balancing human needs with environmental conservation. By adopting sustainable agricultural practices and investing in innovations that enhance productivity, the global community can work toward a future where food security is achieved without compromising the health of the planet.

1.1.2 Modelling Yield Gaps

Closing yield gaps begins with their identification and quantification (FAO & DWFI, 2015). Since a yield gap represents the difference between a reference and actual yield, its measurement requires a reliable estimate of a benchmark yield for comparison. Accurately modelling yield gaps, therefore, depends on establishing this benchmark yield, which serves as the reference point against which actual yields are assessed. Various benchmark yields have been defined in literature which include the potential yield, water-limited potential yield, and attainable yield. Potential yield refers to the maximum possible yield of a crop cultivar grown under ideal conditions, including a non-limiting supply of water and nutrients, absence of biotic stresses such as pests and diseases, and optimal agronomic management such as timely sowing, appropriate plant density, and effective weed control (van Ittersum & Rabbinge, 1997). However, in rainfed agricultural systems, crop growth is often constrained by limited water availability, reducing the potential yield to what is termed the water-limited potential yield. Although irrigation can theoretically overcome this constraint, it is frequently not a practical option for many farmers. Under such conditions, the attainable yield is defined as the largest yield that can be achieved with good agronomic practices such as proper fertilizer use and timely weeding despite environmental limitations (FAO & DWFI, 2015).

Benchmark yields can be estimated using experimental data, empirical models, or process-based simulation models (Figure 1.2). Experimental studies have been used to determine benchmark yields, enabling the assessment of yield gaps and their potential causes (Oteng-Darko et al., 2013). In these cases, this benchmark yield represents the attainable yield i.e. yield estimated from well-managed plots where, as far as possible, all limiting factors such as nutrient deficiencies and diseases are eliminated. While this

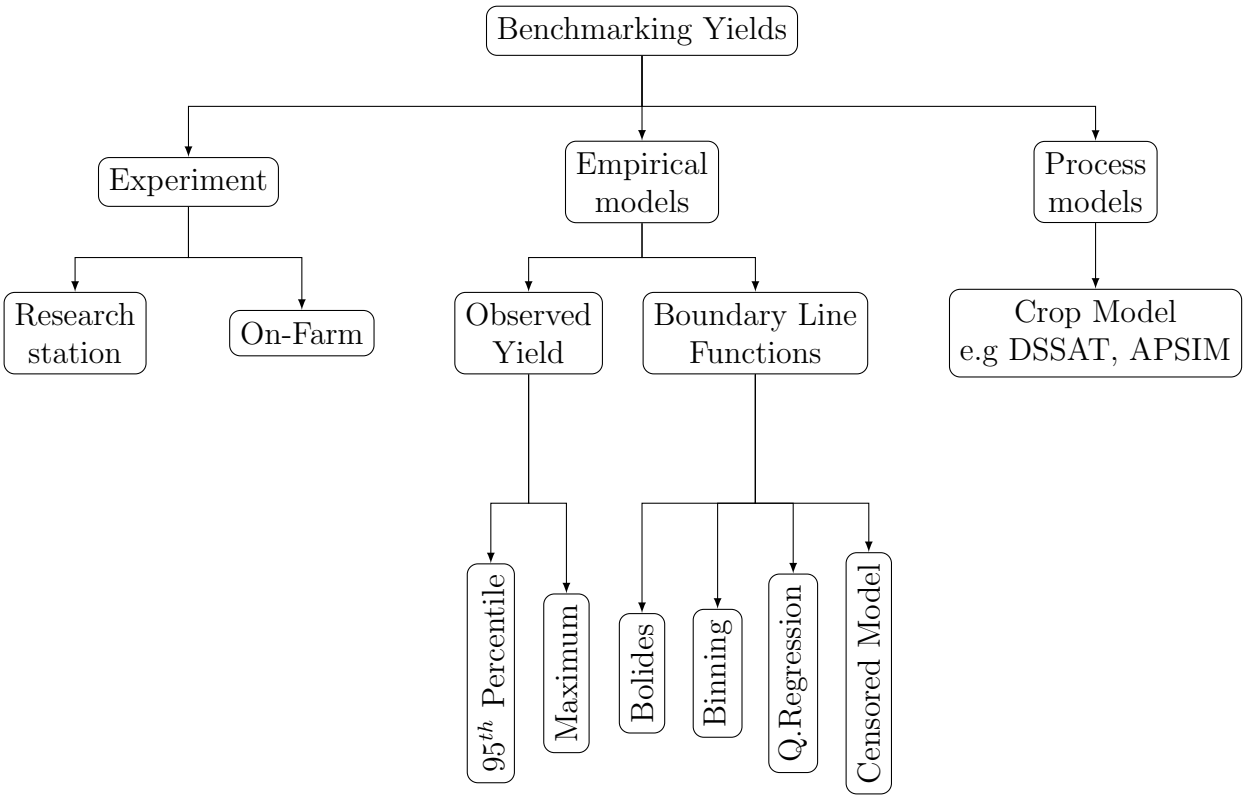


Figure 1.2: Methods for benchmarking yield for yield gap analysis

approach provides valuable insights, it has limitations in capturing the variability and constraints present in real-world farming systems. Moreover, it requires carefully managed field experiments in which yield-limiting and yield-reducing factors are controlled and must be replicated over multiple years to obtain a robust estimate of average attainable yield and its variation (Cassman et al., 2003).

Surveys that collect data on crop yield, soil conditions, farming systems, and economic factors from non-experimental settings have gained popularity as a valuable tool for agricultural research. Empirical approaches offer a valuable means to model benchmark yields and identify limitations to crop production using such data. An empirical model is a data-driven approach used to describe or predict relationships between variables based on observed data rather than a complete theoretical description of mechanisms (Plantinga & Irwin, 2017). These models rely on statistical techniques to identify patterns, correlations, and trends within datasets, making them particularly useful for analyzing complex systems where underlying processes are not fully understood or are difficult to quantify.

A common empirical approach to estimating yield gaps from survey data involves using the yields from high-performing farms within a region with similar climatic and soil conditions as a benchmark. By analyzing farm surveys, a benchmark attainable yield can be approximated as either the largest recorded yield in the region or as the average yield of farmers achieving yields above the 95$^{\text{th}}$ percentile. While farmer-reported yields

provide valuable insights into maximum achievable yields in a specific location or across a population of fields (i.e., the best genotype × environment × management interaction), it remains challenging to confirm whether all biotic and abiotic stresses were avoided. Therefore, yields derived from such data may not always provide robust estimates of yield potential representative of the dominant weather and soil conditions in a given cropping system or region (van Ittersum et al., 2013). In these cases, yield gaps are primarily attributed to differences in management.

The boundary line methodology is an alternative empirical approach that can be used to model benchmark yields and estimate yield gaps based on data from non-experimental settings. This concept, first introduced by Webb (1972), describes the relationship between a system's output (e.g., crop yield, y) and a potentially limiting factor (e.g., soil nutrient concentration, x). The fundamental principle is that the upper boundary of a scatter plot of these variables represents the maximum attainable yield for a given value of x, known as the x-limited response, while observed values below this boundary reflect the influence of additional limiting factors.

The boundary line methodology has been widely applied in yield gap analysis to assess the impact of various constraints, including soil nutrient deficiencies, weed infestations, pest and disease pressures, water limitations, and labour availability (Abravan et al., 2016; Baudron et al., 2019; Casanova et al., 1999; Chen et al., 2018; Fermont, van Asten, et al., 2009; Hajjarpoor et al., 2018; Nehbandani et al., 2020; Nezamzade et al., 2020). By identifying these constraints, this approach provides valuable insights into strategies for improving agricultural productivity and guiding policy interventions.

Boundary lines are commonly constructed using heuristic and statistical approaches. Heuristic methods rely on subjective decisions for their implementation and includes approaches like the visual assessment where 'boundary observations' are picked out by eye (Hajjarpoor et al., 2018), binning where the range of x-values is divided into some number of bins and some value (e.g. 95^{th} or 99^{th}) in each bin are selected as boundary point (Casanova et al., 1999), quantile regression where some percentile of the modelled distribution of y for x is picked out as the boundary value (Cade & Noon, 2003), and the boundary line determination technique (bolides) where the boundary points are selected as the points that completely enclose the whole data (Schnug et al., 1995). In these methods, boundary points in visual assessment, the choice of bin size in binning, and the selection of the percentile to define the boundary in both binning and quantile regression are all arbitrary, introducing subjectivity into the analysis. In the bolides method, subjectivity stems from the process of outlier identification, which relies on an arbitrarily defined separation distance from the nearest point. In contrast, statistical approaches involve specifying a model for the boundary process and the data, which can be fitted using objective criteria such as maximum likelihood. Examples include the censored bivariate

normal model (Milne et al., 2006) and Makowski quantile regression approach (Makowski et al., 2007), among others.

The boundary line approach offers several advantages in agricultural research and yield gap analysis. It effectively accounts for variability in agricultural systems, capturing the complexities of real-world farming conditions. By establishing the relationship between yield and specific limiting factors, it provides actionable insights for farmers, researchers and policymakers, facilitating targeted interventions to enhance productivity. This method is particularly valuable for identifying attainable yields across diverse environmental and management conditions, helping to prioritize critical constraints that limit agricultural output. Additionally, boundary line analysis integrates both environmental and management factors, enabling a comprehensive assessment of production systems and supporting strategies for sustainable intensification.

The third approach to modelling benchmark yield is the use of process-based models. A process-based model explicitly represents the mechanisms and processes within a system, providing a more detailed understanding of how changes in inputs affect outputs e.g. crop simulation models. Modelling can be defined as the use of equations or sets of equations to represent the behaviour of a system. Therefore, a crop model imitates the behavior of a real crop by predicting the growth of its components, such as leaves, roots, stems and grains. Thus, a crop growth simulation model not only predicts the final state of crop production or harvestable yield, but also contains quantitative information about major processes involved in the growth and development of the crop. These models simulate the growth and development of crops over time, taking into account the interactions between crop and its environment. They use biophysical principles to estimate potential yields under optimal management and environmental conditions using a wide range of factors, including weather data (e.g., temperature, precipitation, solar radiation), soil properties (e.g., texture, fertility, water-holding capacity), crop genetics (e.g., growth traits, resistance to pests and diseases) and management practices (e.g., planting dates, irrigation, fertilization). The simulated yield can be used as a benchmark and compared to the actual observed yields so as to determine the yield gap.

Examples of crop simulation models include AquaCrop (Steduto et al., 2009), Crop Estimation through Resource and Environment Synthesis (CERES) (Boote et al., 1996), Agricultural Production Systems Simulator(APSIM) (Keating et al., 2003), World Food Studies (WOFOST) (Van Diepen et al., 1989) and the Decision Support System for Agrotechnology Transfer (DSSAT) (Jones et al., 2003). Crop simulation models are notably applied in the Global Yield Gap Atlas (GYGA) (van Ittersum et al., 2025), a comprehensive initiative designed to assess and map yield gaps at regional and national levels worldwide.

Crop simulation models have the advantage of accounting for interactions between

weather, soil, and management practices, providing a more comprehensive understanding of crop growth and yield potential (Van Asten et al., 2003). However, the use of crop simulation models comes with its own challenges. Although specification of weather, soil, and management practices in cropping systems is essential for robust simulations of yield potential, these data are typically not available for most cropping systems with adequate geo-spatial detail, even in developed countries (Van Asten et al., 2003). In addition, models rely on complex assumptions and may need to be rigorously evaluated for their ability to reproduce measured yields of field crops that received near-optimal management practices, across a wide a range of environments and management practices (Van Asten et al., 2003). FAO & DWFI (2015) stated that the current crop simulation models are unsuitable to answer some questions of local relevance. For example, generic crop models often fail to account for the complex biophysical interactions between green manure and soils with contrasting chemical and physical properties (FAO & DWFI, 2015).

Accurately modelling yield gaps is essential for identifying constraints to agricultural productivity and development of targeted interventions to enhance food production. This section has outlined various methods for estimating benchmark yields, including experimental data, empirical models, and process-based simulation models. However, each of these approaches has inherent limitations. Experimental data, while valuable, often fails to fully capture real-world variability, as controlled field conditions may not reflect the complexities of diverse farming environments. Process-based simulation models, on the other hand, require extensive calibration and often rely on assumptions that may not hold across different agro-ecological settings, limiting their applicability. Among empirical approaches, the boundary line methodology has gained particular traction as a practical and statistically grounded means of estimating yield potential directly from observed data. Unlike experimental or simulation-based methods, it provides a way to assess upper-bound yield limitations using real-world datasets while avoiding some of the constraints of other modelling techniques (e.g. extensive calibration and reliance on complex assumption). Its ability to identify attainable yields under specific conditions makes it a valuable tool in yield gap analysis and agronomic decision-making.

1.1.3 Need for Standardized Analytical Procedures and Tools for Yield Gap Analysis

Yield gap studies have expanded significantly in response to the urgent need to enhance crop productivity and meet the demands of a growing global population. As a result, there is an increasing demand for reliable information on strategies to improve agricultural yields. Synthesizing findings from various studies worldwide can provide valuable insights into the most effective agronomic practices for enhancing productivity. However, achieving

this requires reliable and reproducible agronomic research for quantifying yield gaps across multiple scales, from individual fields to global assessments. Accurate modelling of yield gaps and their underlying causes is, therefore, essential for developing appropriate and consistent interventions aimed at sustainably intensifying agriculture (van Ittersum et al., 2013). One major challenge in current yield gap studies is their often localized relevance, meaning that results may only be applicable to specific geographic regions (van Bussel et al., 2015). Additionally, inconsistencies in the methodologies used to model yield gaps present a significant limitation. For instance, studies employing crop simulation models may use different input data, parameterization procedures, and assumptions, making it difficult to compare results across studies and potentially leading to varying agronomic recommendations for similar conditions. To address this issue, there is a pressing need to standardize analytical procedures to ensure consistency and reproducibility in yield gap assessments and improve the reliability of agronomic recommendations.

One example of efforts to standardized yield gap analysis procedures at global level is the GYGA which employs a structured approach to estimate yield gaps. It begins by defining climate zones based on long-term weather patterns, followed by selecting representative weather stations to provide high-quality climate data. Soil profile data, including texture and water-holding capacity, is then collected to support accurate crop modelling. Crop simulation tools such as APSIM or DSSAT are used to estimate potential and water-limited yields under optimal management conditions. Actual yields are gathered from agricultural statistics or field surveys, and yield gaps are determined as the difference between potential and actual yields. The results are then mapped to provide spatial insights, aiding policymakers, researchers, and farmers in identifying opportunities for yield improvement. Full details can be found in van Bussel et al. (2015).

By providing a standardized and globally consistent method for yield gap analysis, GYGA ensures comparability across different regions and cropping systems. This consistency is vital for tracking progress over time and for assessing the impact of agricultural interventions and climate change on global food production. As a result, the GYGA has become a common tool for determining yield gaps and has been widely used in many studies (e.g. Grassini et al., 2017; van Bussel et al., 2015; Ishikawa et al., 2020; Hochman et al., 2016; Guilpart et al., 2017; Schils et al., 2018; Fayazi et al., 2022). The boundary line methodology, on the other hand, lacks such a standardized procedure (Hajjarpoor et al., 2018), resulting in the use of varied approaches for fitting the boundary line model and interpreting results. This inconsistency makes it challenging to ensure uniformity in agronomic recommendations, potentially leading to variability in yield gap assessments and decision-making. There is a need to establish a conceptual framework for assessing the assumptions underlying the boundary line methodology, standardizing model fitting procedures, and developing more scientifically rigorous approaches for interpreting boundary

line models.

Standardizing analytical procedures alone is insufficient to ensure consistency, reproducibility, and effective dissemination of agronomic recommendations. Achieving these goals requires the development of accessible, robust, and user-friendly data analytical tools that facilitate the application of standardized methods. One example of such a tool is the GYGA, a publicly available platform that provides a structured approach for yield gap analysis. However, GYGA is one of the few open-source tools currently available, and its methodology does not cover all yield gap analysis approaches. There is a critical need to expand such analytical tools to encompass a broader range of yield gap methodologies, improving accessibility and ensuring that standardized procedures can be effectively applied across diverse agricultural systems.

Most yield gap analysis procedures are currently implemented using popular programming languages such as `R` and `Python` (e.g., Gerber et al., 2024). These languages provide a robust environment for data analysis and modelling, making them well-suited for yield gap studies. However, there remains a significant opportunity for the development of specialized libraries/packages on these platforms that are solely dedicated to yield gap analysis. By creating comprehensive, user-friendly, and well-documented libraries, researchers and practitioners can streamline their workflow and improve the accuracy and efficiency of their analyses. Moreover, there is still considerable potential to enhance the accessibility and usability of yield gap analysis tools, particularly through open-source platforms such as the Comprehensive `R` Archive Network (CRAN) and Bioconductor which are increasingly popular for open-source tools. These platforms provide a collaborative ecosystem that allows for continuous improvements, peer contributions, and wide dissemination of tools to a global audience. Expanding the availability of yield gap analysis tools in such repositories can facilitate broader adoption, promote standardization of methodologies, and encourage interdisciplinary collaboration across agricultural, environmental, and economic research domains.

By utilizing these open-source platforms, developers and researchers can contribute to building an inclusive and dynamic environment where yield gap analysis tools are readily available to stakeholders, ranging from policymakers to farmers, enabling data-driven decision-making and fostering sustainable agricultural development.

1.2 Problem Statement

The boundary line methodology has become an increasingly popular and effective tool for yield gap analysis from non-experimental data. It has been widely adopted by researchers to develop agronomic recommendations. The growing body of research employing this methodology highlights its significance in modelling yield potential and identifying limit-

ing factors for yield gap analysis. However, despite its broad adoption and critical role, there is no standardized procedure for fitting boundary lines to data, leading to inconsistencies in its application (Hajjarpoor et al., 2018).

Most studies utilizing the boundary line approach rely on heuristic algorithms, such as binning, bolides, and quantile regression, to fit boundary line models. These methods, however, have significant limitations as they do not provide statistical evidence supporting the biological interpretation of the boundary as a yield limiting model. In contrast, Milne et al. (2006), Lark & Milne (2016), and Lark et al. (2020) proposed a censored bivariate normal model as a theoretical foundation for boundary line analysis. This statistical approach enables rigorous hypothesis testing by comparing the boundary model to an alternative linear relationship with additive random effects. Additionally, it quantifies uncertainty in boundary model parameters, providing a more robust framework for yield gap analysis. However, despite its advantages, the censored bivariate normal model has a limitation in its current form as it does not incorporate categorical variables, which are often essential in agronomic research. Factors such as slope (flat or steep), soil depth (shallow or deep), or management practice categories (e.g wedding frequency, date of first wedding and so on) usually recorded in surveys influence yield responses and should be integrated into boundary models to improve their applicability across diverse contexts. Consequently, this statistical approach has seen limited adoption in the research community. Addressing this limitation will enhance the usage, accuracy and interpretability of boundary line analysis for yield gap studies. Moreover, there is limited information on how the outputs and interpretations of different boundary line fitting methods compare, making it difficult to assess their robustness and consistency and further complicating methodological choices in boundary line analysis.

The boundary line methodology assumes that the dataset exhibits a clear limiting boundary structure at the upper edge of the $\{x, y\}$-scatter and that it represents a plausible biological relationship of y as function of x. This is possible when the dataset covers a sufficiently wide range of conditions such that there is a significant number of cases where the limiting effect of x is expressed because no other factors are limiting. In such a case, the data should exhibit some form of clustering at the upper edges representing maximum biological limit of y given x. This is important as not all data is suitable for boundary line modelling. Besides, anyone can fit a boundary model to the margins of any data cloud even if no such boundary exists. It is therefore important that prior to fitting the boundary line model, one should consider (1) whether the boundary model defining y as a function of a potentially limiting factor x is biologically or agronomically plausible, (2) does the data exhibit a limiting boundary structure and (3) whether the sample size is sufficient for the analysis. Despite the importance of addressing these questions, there are currently no well-established exploratory tools to assess whether a dataset is suitable

for boundary line modelling. Without such tools, heuristic methods risk being applied incorrectly, leading to misleading conclusions about yield constraints. Developing robust exploratory data analysis methods will help ensure that boundary line models are applied only in appropriate contexts, improving the reliability of agronomic recommendations.

Additionally, there is a lack of accessible and standardized open-source software for conducting boundary line analysis. Most existing implementations rely on custom scripts, making replication and broad adoption difficult. The absence of dedicated analytical tools not only hinders the implementation of more advanced statistical models but also limits the transparency and reproducibility of yield gap research. Developing an open-source `R` package with interactive tools will facilitate broader adoption, allowing researchers to apply boundary line methods more effectively and consistently.

1.3 Aims and Objectives

The objectives of this thesis are as follows

1. To conduct a critical review of the use of the boundary line methodology in yield gap analysis and develop a conceptual framework for its agronomic application.

2. To develop exploratory data analysis methods for assessing data suitability prior to boundary line analysis.

3. To compare and evaluate different methods for fitting boundary line models to data and the interpretation of the models.

4. To extend the censored bivariate normal boundary line model to allow for better integration of categorical variables.

5. To develop an `R` statistical package for boundary line modelling that is user-friendly, incorporating interactive tools for exploratory data analysis, initial parameter selection, model fitting and solutions to problems (2) and (4).

By advancing the boundary line methodology, this research aims to contribute to the broader goal of achieving sustainable and equitable food systems. The findings are expected to inform efforts to close yield gaps, enhance resource efficiency, and support agricultural resilience in response to global challenges. Ultimately, this work seeks to address the dual challenges of feeding a growing population while protecting natural resources, aligning with global priorities such as the United Nations Sustainable Development Goals.

1.4 Thesis Structure

Chapter 1 gives the background to the study including the problem and the objectives. It provides the definition of some critical terminologies and general methods used in yield gap analysis at local and global level. **Chapter 2** is a critical review of the agronomic interpretations of the boundary line model, the currently available boundary line fitting methods, their strengths, weaknesses and application for yield gap analysis. It also gives an overview of the commonly used methods and the factors that influence the selection of boundary line methods to use. **Chapter 3** describes the development of a data exploratory procedure for boundary line analysis. It describes the process of determining evidence that supports the assumption for the presence of upper boundary structures in a dataset which is an important initial step when fitting boundary lines to data. **Chapter 4** is based on a study to compare the outcomes of the different boundary line methods and its implication on agronomic interpretations. In this study, four widely used methods that include the binning, bolides, quantile regression and the censored bivariate normal model are compared. **Chapter 5** addresses the extension of the censored normal model to allow for integration of categorical independent variables. **Chapter 6** presents `BLA`, a newly developed `R` package sorely dedicated for boundary line analysis. It describes the theory of the contained functions, their structures and illustrative examples of their application. **Chapter 7** is the thesis discussion which provides a synthesis of **Chapters 1 to 6**. It gives a consolidated message from all the chapters addressing the objectives of the study while **Chapter 8** gives the general conclusions and recommendations.

1.5 Ethical Statement

Chapter 4 of this PhD study is based on data collected through anonymous questionnaires distributed to participants of two workshops held in Zimbabwe and Kenya. These workshops focused on the use of the boundary line methodology for yield gap analysis. The questionnaires aimed to gather participants' perspectives on various boundary line fitting methods and yield gap analysis in general. Ethical approval for this study (SBREC202324010FEO-SB202223/35 PhD - Miti (Lark)) was granted by the School of Biosciences Research Ethics Committee at the University of Nottingham. This approval ensured that workshop participants provided informed consent, including consent for the subsequent use of their data in this PhD study.

Chapter 2

The Concepts and Quantification of Yield Gap Using Boundary Lines: A Review

The chapter has been published as:
Miti. C, Milne. A.E, Giller. K.E, Lark. RM (2024). The Concepts and Quantification of Yield Gaps using Boundary Lines. A Review. *Field Crop Research Journal*, 311, 109365. https://doi.org/10.1016/j.fcr.2024.109365.

2.1 Chapter 2 Overview

This chapter presents a comprehensive review of the boundary line methodology as a tool for yield gap analysis in agronomic research. It explores the conceptual foundations of yield gaps and how boundary lines can be used to diagnose limiting and reducing factors in crop production. The review systematically examines the various methods used to estimate boundary lines, contrasting heuristic approaches with more rigorous statistical techniques, and evaluates their interpretability, reproducibility, and utility for guiding agronomic decision-making. Particular attention is given to the challenges of model inference, the implications of different conceptual interpretations (e.g., the Law of the Minimum and the Law of the Optimum), and the need for methodological standardization. This work sets the foundation for later chapters that apply, test, and refine these methods using empirical data. This work has been published in ***Field Crops Research Journal***. I contributed to this publication by leading the conceptualization, designing the methodology (including literature search and analysis), conducting the literature review and analysis, preparing the original manuscript draft, and developing the visualizations.

Field Crops Research 311 (2024) 109365

Contents lists available at ScienceDirect

Field Crops Research

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

The concepts and quantification of yield gap using boundary lines. A review

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

[a] School of Biosciences, University of Nottingham, Sutton Bonington, 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, Wageningen 6700AK, The Netherlands

ARTICLE INFO

Keywords:
Potential yield
Attainable yield
Yield gap
Boundary line

ABSTRACT

Context: The potential yield of crops is not usually realised on farms creating yield gaps. Methods are needed to diagnose yield gaps and to select interventions. One method is the boundary line model in which the upper bound of a plot of yield against a potentially limiting factor is viewed as the most efficient response to that factor and anything below it has a yield gap caused by inefficiency of other factors. If many factors are studied, the cause of the yield gap can be identified (yield gap analysis, YGA). Though the boundary line is agronomically interpretable, its estimation and statistical inference are not straightforward and there is no standard method to fit it to data.

Objective: We review the different methods used to fit the boundary line, their strengths and weaknesses, interpretation, factors influencing the choice of method and its impact on YGA.

Methods: We searched for articles that used boundary lines for YGA, using the Boolean "Boundary*" AND "Yield gap*" in the Web of Science.

Results: Methods used to fit boundary lines include heuristic methods (visual, Binning, BOLIDES and quantile regression) and statistical methods (Makowski quantile regression, censored bivariate model and stochastic frontier analysis). In contrast to heuristic methods, which in practice require ad hoc decisions such as the quantile value in the quantile regression method, statistical methods are typically objective, repeatable and offer a consistent basis to quantify parameter uncertainty. Nonetheless, most studies utilise heuristic methods (87% of the articles reviewed) which are easier to use. The boundary line is usually interpreted in terms of the Law of the Minimum or the Law of Optimum to explain yield gaps. Although these models are useful, their interpretation holds only if the modelled upper limit represents a boundary and not just a particular realization of the upper tail of the distribution of yield. Therefore, exploratory and inferential analysis tools that inform boundary characteristics in data are required if the boundary line is to be useful for YGA.

Conclusions and implications: Statistical methods to fit boundary line models consistently and repeatably, with quantified uncertainty and evidence that there is a boundary limiting the observed yields, are required if boundary line methods are to be used for YGA. Practical and conceptual obstacles to the use of statistical methods are required. Bayesian methods should also be explored to extend further the capacity to interpret uncertainty of boundary line models.

1. Introduction

The development of high-yielding crop varieties is critical to the global food security because global population is expected to rise while the area of land available for agriculture is shrinking (Mueller and Binder, 2015). This increased potential yield, however, is not always realised in actual production. This is known as the yield gap. The yield gap has taken centre stage in discussions about global food security (Giller et al., 2021; Timsina et al., 2018; van Ittersum et al., 2016), and if crop improvement objectives are to be met, agronomists must understand this gap and identify interventions to close it. Therefore, methods for understanding crop yield variability and unraveling the underlying factors that cause the yield gap are urgently needed.

The yield gap is defined in reference to some benchmark yield and here we consider three such benchmarks, the potential yield, the water-limited potential yield and the attainable yield. The potential yield is defined as the genetically-possible yield of a cultivar grown in non-limiting biophysical and environmental conditions i.e. under an

* Corresponding author.
E-mail address: chawezi.miti@nottingham.ac.uk (C. Miti).

https://doi.org/10.1016/j.fcr.2024.109365
Received 15 December 2023; Received in revised form 26 March 2024; Accepted 1 April 2024

optimum supply of water and nutrients, in the absence of biotic stress and with optimal agronomic management i.e. sowing dates, sowing density, weeding and so on (van Ittersum and Rabbinge, 1997). However, under field conditions in rainfed cropping, crop growth may be limited by water supply, hence the potential yield is reduced to the water-limited yield potential. While management interventions like irrigation can, in principle, alleviate this limitation, this might not always be feasible for the farmer. The largest yield that can be achieved under these limiting conditions in well-managed fields (i.e. with correct and timely fertiliser application rates, timely weeding, and so on) is called the attainable yield (FAO and DWFI, 2015; Tittonell and Giller, 2013). Attainable yield is sometimes estimated as 80% of the potential yield or the water-limited potential yield for rain fed systems.

Attainable yields are often not realised by farmers due to suboptimal management of biotic and abiotic factors within their control. These include yield-limiting factors like nitrogen supply, which constrain the crop's net primary production, and yield-reducing factors, such as weeds which compete for available resources and, pests and diseases, which sequester some or all of the net primary production before this is harvested. Hence, there is a yield gap between the potential yield, water-limited yield potential, attainable yield and the farmer's actual yields (Fig. 1), which could, in principle, be closed by applying appropriate management strategies (Cossani et al., 2010; van Ittersum and Rabbinge, 1997; van Ittersum et al., 2013).

To close this yield gap, there is a need to identify and rectify the potentially manageable yield-limiting and yield-reducing factors. The process of quantifying and identification of the underlying cause of yield gaps, is called yield gap analysis (Fermont et al., 2009; van Ittersum et al., 2013). The yield gap can be identified and quantified by comparing the actual yield to a reference benchmark yield, which can be the potential or water-limited potential yield predicted by a process model e.g. using the Global Yield Gap Atlas (GYGA) (see https://www.yieldgap.org). The use of process models to determine the benchmark yield depends on the assumption that the model is not significantly biased and that all factors determining yield are properly understood and quantified for a particular setting (Shao et al., 2023). An alternative is the use of an empirical approach that observes actual yields and plots them against corresponding measurements of potentially limiting factors. An example data set consisting of hypothetical wheat yield and a soil property x_i is presented to illustrate this. In a setting where farmers practice a wide range of management practices, it is expected that there will be a wide range of actual yields depending on how individual farmers manage the yield-limiting or yield-reducing factors, x_i (Fig. 2). If one can assume that the sample is large enough to include cases in which the crop is managed as well as possible, then this can be taken as the attainable yield and can be used as a benchmark yield in that particular setting (e.g 12.5 t ha^{-1} in Fig. 2).

While the attainable yield might be identified empirically, further insight into the yield gap is required to guide interventions. The interpretation of a plot of some biological response against one of several potentially limiting factors was first proposed by Webb (1972) as the boundary line model. Unlike controlled experiments, where sources of variations other than that of the factor of interest are controlled as far as possible and modelled as additive random effects allowing the fitting of a best-fit median regression model through the data scatter, non-experimental data (e.g. real farm data) have many uncontrolled sources of variation. This results in a scatter of the response variable, y, against a factor, x_i, for which there is a range of responses for a single level of the factor. As proposed by Webb (1972), this kind of plot for biological response may be interpreted in terms of an upper limit for the response which appears to depend on x_i (see Box 1). This upper limit can be modelled as a function of x_i, specifying the largest response for some value of x_i as shown in Fig. 2. The observed responses are assumed to be limited by factors other than x_i in some way and the upper limit of the response as a function of x_i represents the most efficient response (i.e. when all other factor are not limiting) for a given level of x_i, which Webb (1972) called a boundary line, $f_i(x_i)$. This gives a better representation of the relationship of y as a function of x_i than the median best fit line in this case. Points below the boundary line have a response gap equal to the difference between the actual responses and the maximum response at that level of factor x_i (Fig. 2) (Webb, 1972). A boundary line model can, therefore, be used to model the largest yield as a function of

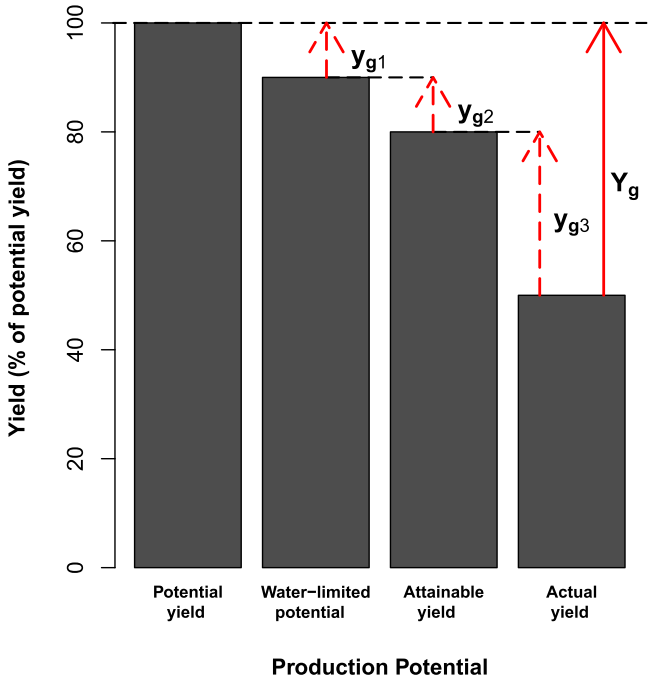


Fig. 1. Yield gap estimation using different production potentials as benchmarks. The yield difference between potential yield and water-limited potential yield (y_{g1}), water-limited potential yield and attainable yield (y_{g2}), and attainable yield and actual yield (y_{g3}) add up to the total yield gap (Yg).

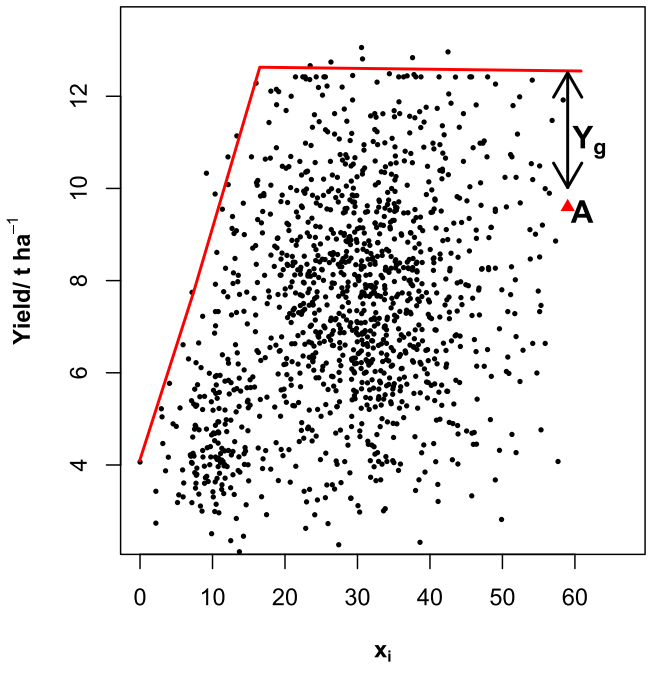


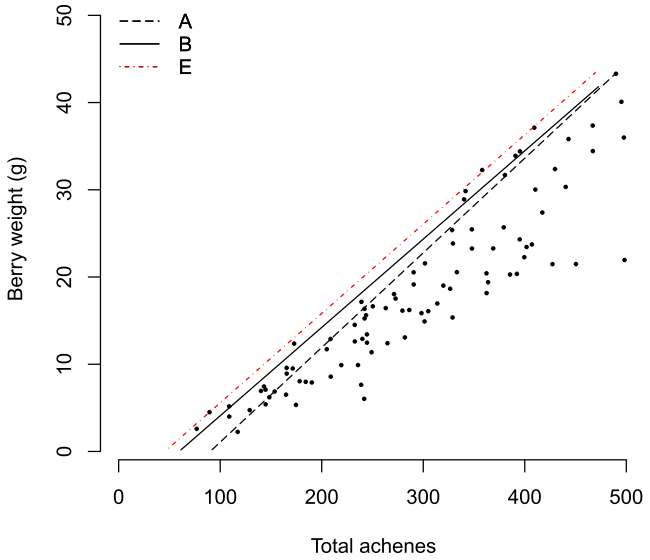
Fig. 2. A hypothetical scatter plot of maize yield against yield-limiting or yield-reducing factor, x_i, with a model (the red solid line) that predicts the maximum yield response that can be achieved at a given level of factor, x_i. Yield gap, Y_g, is the difference between the attainable yield (12.5 ha^{-1}) and the actual yield,A.

yield-limiting or yield-reducing factors and therefore, the yield gap. The largest predicted yield in a given setting represents its attainable yield (Dehkordi et al., 2020).

to what extent the different approaches provide an objective way of fitting the boundary line, their weaknesses and strengths, trends in the usage and factors influencing the choice of boundary line fitting

Box 1: Existence of a boundary in biological response data (Webb, 1972)

A scatterplot for achene number against berry (pseudocarp) weight of 94 well-shaped red gauntlet strawberries from a study by Abbott et al. (1970) showed presence of a boundary beyond which response did not exceed (line E). Line A represents the boundary when only the best-developed berries were considered (six achenes per cm^2 of the surface) while line B is boundary when strawberries were grouped at intervals of 20 achenes with the heaviest berry and its associated number of achenes representing each group. This study illustrates how such an upper limit in a dataset can be treated as a standard against which average performance can be assessed, so as to arrive at an estimate of possible increase in yield Box-Fig. 1.



Box-Fig 1. Boundary line relationship between the total number of achenes and berry weight re-plotted from Webb (1972).

Additional useful agronomic information can be extracted from the boundary line model for yield gap analysis such as the contribution of different factors to the identified yield gap as has been done by many agronomic studies (e.g. Casanova et al. 1999; Cao et al. 2019; Fermont et al. 2009; Hajjarpoor et al. 2018; Kintché et al. 2017; van Vugt and Franke 2018). However, the information obtained depends on how the boundary line is interpreted, which must reflect the natural process of crop growth due to the effect of various factors. Various laws which govern how factors affect plant growth in nature have been proposed which can be used to interpret the boundary line model.

Though boundary line analysis is empirically and theoretically plausible as a basis for yield gap analysis, in practice, the estimation of the boundary line, statistical inference as to whether it can be meaningfully interpreted as an upper limit, and an account of uncertainty in parameters of the line are not straightforward (Webb, 1972). There is, therefore, a need for quantitative methods for both estimation and inference if useful information is to be obtained from the boundary line model. FAO and DWFI (2015) gives an overview of methods (that include the boundary line methodology) used to benchmark yield, giving examples, but do not give full detail on the procedures to fit the boundary line function to a dataset. Various methods are available in the literature for fitting a boundary line to a dataset. However, consolidated information on their implementation, strengths, limitations and agronomic interpretations is lacking. In this review, we give an overview of (i) the different agronomic interpretations of the boundary line that have been used in yield gap analysis and (ii) the different approaches to fitting boundary line that are available for yield gap analysis. We analyse approach, and the impact of using the different approaches on yield gap analysis.

2. Methods for review

Peer-reviewed publications were searched using the Boolean term "Yield gap*" AND "Boundary*" applied to All fields in the Web of Science database on 1st June 2023. A total of 70 studies were obtained with this search. These studies were further screened and only publications that applied the boundary line methodology to access yield gaps were selected. Publications that applied other methods for accessing yield gaps as well as review papers were excluded (n = 17) resulting in 53 publications being retained. Using Google Scholar, we also searched for publications cited by the retained articles which used boundary lines for yield gap analysis but did not appear in the initial search. Eleven publications were added from this. This produced a total of 64 publications in which boundary line methods were used for yield gap analysis. The list of selected publications are listed in Table S.2 of the supplementary material. Key information extracted from these articles was stored in a database, specifically the year of publishing, crop studied, domain of study, the agronomic interpretation of the boundary line, criteria used to identify outliers, the method used to estimate the boundary line i.e whether this was based on a statistical model or heuristic method ('non-statistical') and the assumptions made for the boundary line method used (e.g. the percentile value assumed to be the boundary for binning methods). By a heuristic method we mean one that is intuitively reasonable, but generally entails arbitrary decisions and, because it does

not invoke an explicit statistical model, does not provide a natural basis for inference or the characterization of uncertainty. On the other hand, a statistical approach follows strict statistical principles and makes sound assumptions in fitting the boundary line.

3. Boundary line interpretations

In this section, we discuss how the boundary line methodology is interpreted for yield gap analysis. We focus on the principle, strength and limitations of these interpretations.

Plant growth is governed by a series of natural processes (Poorter et al., 2013). The empirical observations of the boundary line must, therefore, be interpreted in light of a conceptual model of these processes if useful information is to be obtained from them. One such conceptual model is the law of the minimum which states that a biological response can only be as large as the factor in least supply can permit (Liebig, 1840; Sprengel, 1826). This can be modelled by expressing the response, y, as a function of factors $x_1, x_2 \ldots, x_n$ as follows

$$y = k\min\{f_1(x_1), f_2(x_2), \ldots, f_n(x_n)\} \quad (1)$$

where $f_i(x_i)$ represents the boundary line response to variable x_i and for this interpretation is scaled so that its maximum value is 1, and k is a constant which represents the attainable yield in this case. If the law of the minimum holds, and the potentially limiting factors $x_1, x_2 \ldots, x_n$ take a wide range of values in the dataset, then the upper boundary on a plot of y against x_i should estimate the attainable yield scaled by the boundary line function, $kf_i(x_i)$, the largest biological response that can be attained by a given level of factor x_i given that no other factors are limiting (Lark et al., 2020). Boundary line functions, $f_i(x_i)$, for the responses of crop yield to several growth-defining, limiting, and reducing factors, $i = 1, 2, \ldots, n$, can be determined and the factor that predicts the minimum yield is viewed as the most limiting factor and therefore, the predictor of yield in that instance (Fermont et al., 2009). The identification of the most limiting factors allows agronomists to prioritize which factors need more attention to increase the yield in a given location (Wairegi et al., 2010).

Though the law of the minimum provides a simple and straightforward conceptual framework for boundary line analysis and has been used by most authors to interpret the boundary line (Cossani and Sadras, 2018; Fermont et al., 2009; Nehbandani et al., 2020; Shatar and McBratney, 2004; Wairegi et al., 2010), it does not account for possible interactions of factors associated with crop growth. That is to say, it assumes that all factors independently determine a possible yield, of which the smallest is actually attained, while in reality two or more independent variables can have a simultaneous effect on the dependent variable such that their joint effect significantly differs (greater or less) from the sum or minimum of the individual effects. One conceptual model that entails such interactions is the law of the optimum which states that the response of a dependent variable to an independent factor that is in minimum supply is largest when the other production factors are close to the optimum (Liebscher, 1895). This means that the growth response of a plant to a given factor of interest depends on a subset of other factors that interact with the factor of interest and that the response is largest when this subset is close to some optimum requirement. As an example, an intermediate process of uptake affects the relationship between the rate of nutrient application and yield. The nutrient uptake itself is also affected by other factors and hence there will be a relative change in position of the yield function in response to a rate of nutrient application due to changing conditions of factors that affect uptake. This intermediate effect of nutrient uptake considered in the law of optimum is not considered in the law of the minimum (de Wit, 1992).

The boundary line can be interpreted in light of the law of the optimum if it is thought of as a rate-limiting function, as proposed by Elliott and de Jong (1993) for the interpretation of boundary responses to soil properties of nitrous oxide emission from soil cores. The response variable, y, is influenced by the limiting factor, x_i, on the abscissa of the boundary plot, a set of other potentially limiting factors which interact with x_i, and a set of other factors which are not limiting. Therefore, the response, y, is given as a product of the function of the factor of interest and functions of factors that interact with the variable x_i.

$$y = k\prod_{i=1}^{n} f_i(x_i) \quad (2)$$

As with Equation (1), the boundary function, $f_i(x_i)$, is scaled to 1. When all factors other than the i^{th} factor on our boundary plot which potentially limits the response, y, are optimal, i.e. for any $j \neq i, f_j(x_j) = 1$, the response, y, will be determined by $f_i(x_i)$, and will be on the boundary of the plot of y against the i^{th} factor. If the data for the potentially limiting factors is large enough and covers a wide range of values, the function, $f_i(x_i)$, will fit the boundary of the plot of y against the i^{th} factor as illustrated in Fig. 3. The solid line represents the boundary line when all factors are optimum as the law of the optimum. If some factor(s) other than the i^{th} factor are not optimal, the boundary line will drop below the solid line as represented by a dashed line in Fig. 3. In this sense, the form of the response to the i^{th} factor is determined by the other factor(s), which is what we mean by an interaction. However, this type of interaction is a simplistic formulation as it is a product of the boundary functions and is, therefore, restrictive. In practice, more complex interaction models could be used to underpin boundary line modelling. In the form presented in Equation (2), the law of the optimum implies that observations on the boundary, below the attainable yield are subject to limitation by x_i while points below the boundary are limited by other factors. The boundary line, therefore, still allows us to identify a yield gap.

The law of the optimum which accounts for the simultaneous effect of different factors of crop growth, is more biologically plausible than the law of the minimum (de Wit, 1992). However, the rate-limiting interpretation of the boundary line has not been widely considered particularly in respect of plant growth. Out of the 64 reviewed articles, none interpreted the boundary line using the rate-limiting interpretation. This may be because the law of the optimum, on which the rate-limiting interpretation is based, has not been widely investigated since it was suggested by Liebscher (1895). A possible approach to account for interaction is the use of boundary surface models from

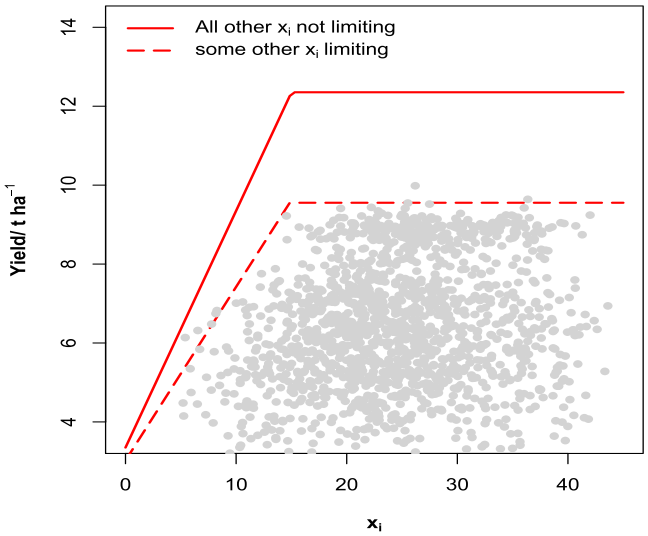


Fig. 3. The illustration of the law of the optimum for boundary line interpretation. The solid line represents the most efficient response of yield to factor x_i when all other factors are optimum and hence represents the boundary line response to x_i, while the dashed line represents the function of the response of yield to factor x_i when some other factors are not optimum.

multi-dimension plots similar to the response surface methodology (Myers et al., 2016). A boundary surface model would represent the maximum attainable yield for combinations of values of potential limiting factors, and a range of functional forms would be possible, including both additive and interactive effects. This approach has been used in plant nutrition studies to optimize nutrient application rates (Jahan and Amiri, 2018; Salawu et al., 2007). However, we are currently not aware of any study that has attempted to use response surfaces by fitting multi-dimension boundary lines to account for interaction.

Interaction has been incorporated in boundary line analysis in other ways including the use of nutrient ratios (nutrient balance indices) rather actual individual nutrients to plot the data. The importance of balanced nutrition (e.g. N:P, N:S, C:N, K:Mg) has been emphasised in agronomic literature (Duncan et al., 2018) as the deficiency of one nutrient might restrict the efficient use of another (Aulakh and Malhi, 2005). Use of these ratios reflects the combined effect of two factors on yield and hence interaction. However, ratio models are based on prior knowledge, and are not a general method to represent interaction of factors. The co-limitation framework (Cossani et al., 2010; Cossani and Sadras, 2018) has also been used to account for interaction in boundary line analysis (Carciochi et al., 2020). Co-limitation occurs when two factors simultaneously limiting plant growth (Riar et al., 2016). This interactive effect has been determined through the use of resource stress indices. A resource stress index (e.g. nitrogen stress index) is a measure of the level of stress experienced by the crop due to limitations in essential resources and is determined by subtracting from one, the ratio of the resource uptake at actual yield to the uptake at potential yield (Cossani et al., 2010). Individual stress indices are used to calculate various degrees of co-limitation which have then been used to evaluate their effect on observed yield gaps from boundary lines (Carciochi et al., 2020).

The boundary line interpretation using the law of the minimum and the law of the optimum provides useful agronomic information that helps to identify the causes of the yield gap and how they can be addressed. Silva et al. (2017b) introduce in addition to the potential yield (Y_p), the attainable yield (Y_a) and the actual yield (Y_r), the technically efficient yield (Y_t), which is the maximum yield attainable given x_i and here equal to the original unscaled $f_i(x_i)$-the boundary value. Note that Y_t is defined only when $f_i(x_i) < Y_a$, i.e. when $f_i(x_i) < f_i(\tilde{x}_i)$ where $f_i(\tilde{x}_i) = Y_a$.

The efficiency yield gap, g_e is

$$g_e = f_i(x_i) - Y_r, \qquad f_i(x_i) > Y_r \tag{3}$$

$$g_e = 0, \quad f_i(x_i) = Y_r \tag{4}$$

This gap indicates that some factor(s) other than x_i limit crop yield and so action is required to remove this limitation. If these limitations are removed, then the expected yield is $Y_t = f_i(x_i)$. The efficiency gap, g_e implies that any resource used to sustain factor x_i at its observed level e.g fertilizer or labour, is inefficiently used as the actual yield could be sustained at some $\breve{x}_i$ where:

$$f_i(\breve{x}_i) = Y_r < f_i(x_i) \tag{5}$$

if g_e were zero, but $f_i(x_i) < Y_a$ then there is a resource gap.

$$g_r = Y_a - Y_t \tag{6}$$

The resource yield gap is attributed to the fact that $x_i < \tilde{x}_i$, so if, for example, x_i is an available nutrient, then some resource (e.g. fertilizer) must be used to close the gap. This is because it is assumed that all other factors are optimum and hence there is a maximum response to factor x_i. If the law of the minimum is considered, however, some other factors may become limiting as you increase the factor x_i and hence the yield may fall below the technically efficient yields (Fig. 4).

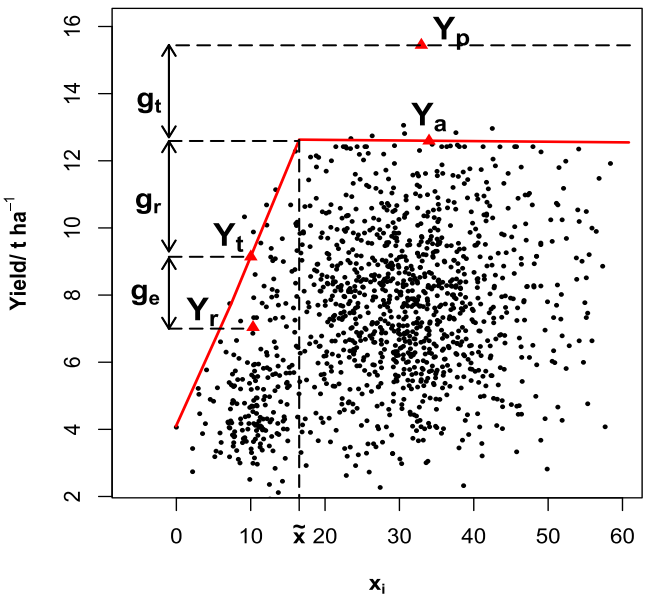


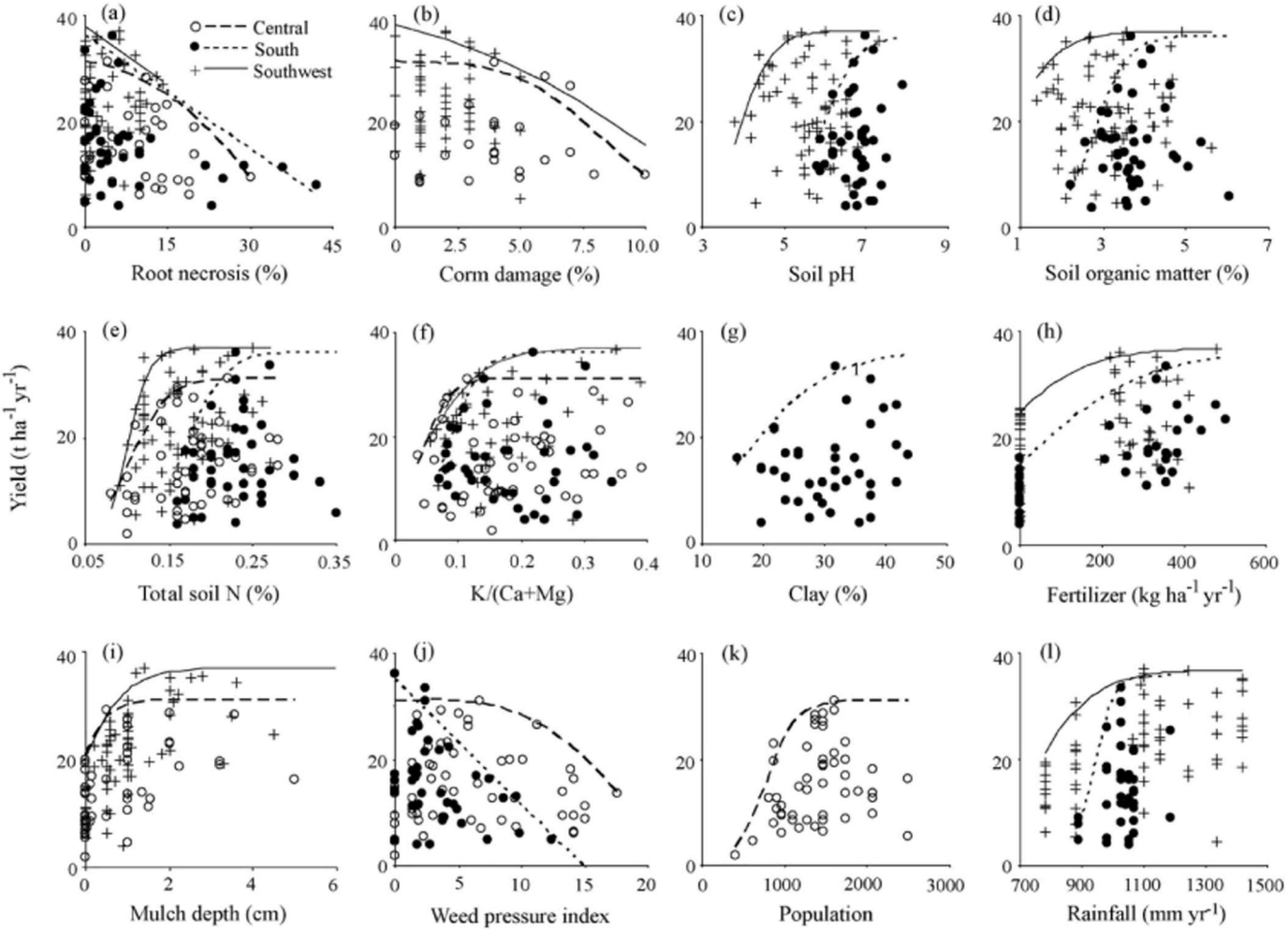
Fig. 4. Decomposition of yield gaps into efficiency (g_e), resource (g_r) and technological (g_t) yield gaps. Y_p is the potential yield, Y_r is the actual yield and Y_t is the technically efficient yield given the limitation of x_i. The level of factor at which the yield reaches attainable yield is given by $\tilde{x}_i$.

The technology yield gap (g_t) is the difference between the potential yield (Y_p) and the attainable yield (Y_a). This gives a measure of how much room there is to increase the yield if improved technologies are applied like the addition of an irrigation system or the growing of crops in a controlled environment like a greenhouse.

When the boundary line is applied to several factors that affect yield in a given location, one is able to check if one of the factors considered can explain the identified yield gap (Box 2). The explained yield gap, as illustrated in Fig. 5, can be quantified as the difference between the attainable yield, Y_a, and the largest yield predicted by the most limiting factor while unexplained yield gap can be quantified as the difference between the largest yield as predicted by the most limiting factor and the actual yield (Fermont et al., 2009; Wairegi et al., 2010) as shown in Fig. 5(b). Given a set of actual yields measured simultaneously with two soil factors A and B as shown in Fig. 5, we can select one particular farm with actual yield, Y_r, and show it on the plot of the two soil factors. It can be seen that the total yield gap is approximately 9.5 t ha^{-1}. Factor B is the most limiting factor as it predicts a smaller yield, Y_{lim} (approximately 8.5 t ha^{-1}), as compared to factor A (12 t ha^{-1}). The unexplained yield gap will, therefore, be the difference between Y_{lim} and Y_r as there are unknown factors that need to be addressed for Y_r to be increased to Y_{lim}. Once these unknown factor are addressed, increasing the yield from Y_{lim} to Y_a can be achieved by increasing the level of factor B and therefore, this is referred to as explained yield gap.

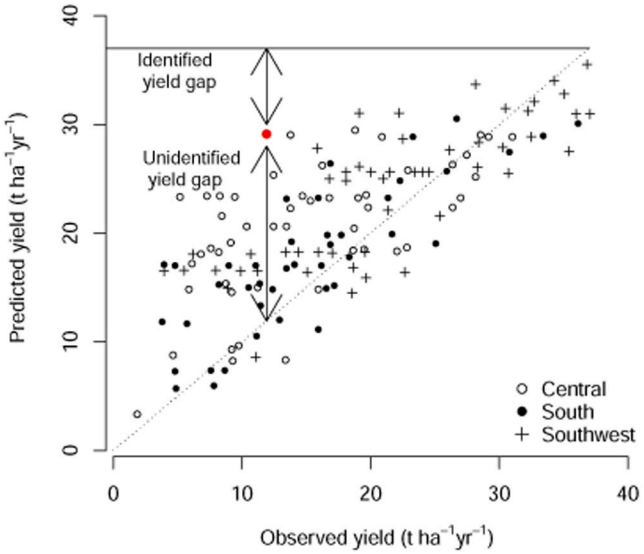
Box 2: Explaining yield gaps in East African highland banana systems (Wairegi et al., 2010)

A study was carried out to identify the factors limiting banana production in Uganda. Banana yield data for different farms were collected in central, south and south-west regions of Uganda and a variety of limiting factors including soil pH, SOM, Total-N, K, Ca, Mg, nematodes, weevils, weeds, plant population and rainfall were also measured. Boundary lines were estimated for each factor (Box-Fig. 2).



Box-Fig 2. Boundary relationship between banana yield and biophysical factors in Uganda (Wairegi et al., 2010).

The expected yield for each farm was then estimated using the boundary lines and the factor that predicted the lowest yield was taken as the most limiting factor. Yield gap estimation and decomposition was conducted by plotting the predicted yield using boundary lines against the actual yields as shown in Box-Fig. 3. The continuous line at 37 t ha^{-1} yr^{-1} represents the highest yield observed in the study area and it is taken as the attainable yield. The 1:1 dotted line represents the situation when the actual yield is equal to the predicted yield. The points that fall above the 1:1 line means the predicted yield by the most limiting factor is larger than the actual yield and therefore, its yield gap cannot be explained by any of the factors investigated in the study. The yield difference between the predicted and attainable yield can be explained by the most limiting factor. For example, taking the yield represented by the red dot above the 1:1 line, the unidentified yield gap of 18 t ha^{-1} yr^{-1} while the explained yield gap is 7 t ha^{-1} yr^{-1}. In this way, they were able to identify which factors were mostly limiting yield in the area and which one could be prioritised to increase banana yields.



Box-Fig 3. Observed and predicted yield from boundary line approach replotted from Wairegi et al. (2010).

4. Estimation of boundary line function and its statistical inference

In this section, we discuss the process of fitting a boundary line function, $f_i(x_i)$, as described in Section 3 to an $\{x, y\}$ scatter plot. We first discuss the importance of statistical inference about the boundary line model for yield gap analysis and then we explore the different methods currently available for fitting the boundary line function to data.

4.1. The need for statistical inference about boundary line functions for yield gap analysis

We have shown that the boundary line model can be interpreted consistently with two different conceptual models. Under either model, the boundary function may be useful for developing practices and interventions to close the yield gap as illustrated by our cited examples. However, we must also consider whether the boundary line interpretation is supported fully by the data (FAO and DWFI, 2015). After all, a boundary could be drawn by hand on any plot of some y and x variable, including the familiar model where deviations of y from some $f(x)$ are independent additive random effects. Even if one of the conceptual models holds, some of the variation in y is likely to behave as additive random variation, specifically the measurement error for the crop yield, or other response.

Milne et al. (2006b) developed an exploratory approach to this hypothesis in which the density of points in the upper section of successive convex hulls (peels) of data in $\{y, x_i\}$ space is compared with the same statistic for a null model with additive random effects only. The convex hull of the data in a 2-D scatter plot is the smallest subset of the data which constitute a convex subset containing all the observations (Skiena, 2008). All the points in a convex hull of a data set comprise the first "convex hull peel". The second peel is the convex hull of the remaining data after the first peel is removed. If a boundary exists for which responses cannot exceed, it is expected that the data will have a distribution that has a denser concentration of points than normal near the boundary (upper convex hulls). The distribution will take the form of a censored bivariate distribution. Otherwise, if a boundary does not exist, the data is likely to follow a bivariate normal distribution with randomly distributed points some of which lie at the extreme fringes. The data points at the extreme fringes arise due to random additive error associated with the effect of other factors affecting response other than the variable of interest as in a general linear model with additive random effects. This can be taken as a null model against which a boundary line model can be tested. This method is, however, less powerful as it is based on just the number of vertices in the peels ignoring the distribution of the peels at the upper bounds of the data.

This work by Milne et al. (2006b) is the first of which we are aware that attempted quantitatively to test the plausibility of a boundary model. We suggest that further work is needed on exploratory and inferential analysis to test the plausibility of boundary models if they are to be used to interpret data in terms of yield gaps and provide recommendations. It should also be noted that the interpretations from boundary line analysis are conditional on the support of the data (i.e. if the yields and environmental variables are field means, farm means, or means for small plots within fields) (FAO and DWFI, 2015). Results at different scales may differ as additional constraints may apply more at one scale compared to another.

4.2. Methods for fitting boundary line functions

The reliability of the information obtained from a boundary line model is dependent on method used to fit the boundary line function. A good fitting method must possess three important qualities. Firstly, it should be objective (Schnug et al., 1995; Shatar and McBratney, 2004)

as this ensures that the results are consistent, reproducible and can easily be compared. Secondly, the method should be able to account for measurement error in response variables and allow quantification of uncertainty of the boundary line (Lark et al., 2020; Makowski et al., 2007; Milne et al., 2006a). Finally, as noted in the previous section, the method should allow some objective test against a null (non-boundary) alternative to check if a boundary exists within a dataset (Lark and Milne, 2016; Lark et al., 2020; Milne et al., 2006a).

Despite the increasing use of the boundary line approaches in the agronomic literature (see Fig. 8(a)), there is no standard protocol to estimate a boundary line for yield gap analysis (Hajjarpoor et al., 2018; Shatar and McBratney, 2004). Many methods currently in use to derive the boundary line follow a similar process of (i) Plotting a scatter for dependent against independent variables, (ii) removal of outliers (iii) selection of the boundary data points, and (iv) fitting a boundary line to the selected points which may take the form of a linear model, broken-stick model or non-linear model. Important difference amongst the methods are the criteria used to identify outliers, the procedure used to select the boundary line points and the method of fitting the boundary line to the selected boundary points.

Outliers are extreme data values which appear to arise from a different process to most of the data. As a result they can cause bias and influence estimates of a statistical model. Outliers are of particular concern in boundary line analysis where the hypothesis is that the interesting biological relationship is expressed by the bounding observations of a response variable and a potentially-limiting factor. Despite the likely sensitivity of boundary line models to outliers, most of the reviewed articles did not indicate whether they removed outliers from datasets in their analysis. Of the 64 articles reviewed, only 14 indicated that they identified and removed outliers from their data. Of these 14, four articles did not indicate the criteria used to identify the outliers. A *neighbourhood density* procedure was used to identify outliers in two articles. In the neighbourhood density procedure, an observation is regarded as an outlier if it does not have at least some threshold number of neighbouring observations within a specified radius (Schnug et al., 1995). In one article the authors examined the scatter plot of y against x and identified points that looked unusual on the upper bound of dataset by '*inspection and judgement*'. One article used a bag plot to identify outliers. A *bagplot* is a two-dimension boxplot based on the measure of half-space depth (Rousseeuw et al., 1999). In five of the reviewed articles, outliers were identified on the *boxplot* of the response variable. However, it was not indicated the criteria used to define an outlier i.e. values of the upper and lower fences. A standardized objective method is needed to deal with outliers if boundary line analysis workflows are to be repeatable. Since boundary line models are applied to two variables with a joint bivariate distribution, the *bagplot* may provide a better tool for identifying outliers as it allows for joint distribution outliers to be identified in a dataset. Although the boxplot is commonly used, it is most suitable for univariate analysis as it takes no account of the value of the potentially limiting variable. The visual and neighbouring density methods are subjective as they involve arbitrary judgements or decisions e.g. the number of points and the size of neighbourhood is subjectively chosen, and so are not repeatable.

There are several approaches commonly used to select the boundary points and fit the boundary line in literature (Table S.2, Supplementary material). We have categorised them into two broad categories namely heuristic and statistical approaches. Heuristic approaches include the visual, binning, boundary line determination method (BOLIDES) and general quantile regression while the statistical approaches include Makowski quantile regression, stochastic frontier analysis, and the censored bivariate normal model methods. We describe these methods in detail in the following subsections. Another important considerations in boundary line analysis is the functional form of the boundary model. Various functional forms including linear (French and Schultz, 1984b), linear plateau (Andrade et al., 2023), trapezium (Nezamzade et al., 2020) and logistic regression (Fermont et al., 2009) models among others have been used in previous agronomic studies. Though the functional form is influenced by and should follow the upper boundary structure, it is recommended that the shape and parameters of the boundary line are assessed on the basis of their biophysical meaning (agronomic or physiological) (FAO and DWFI, 2015). If a boundary surface does not reflect that, it may raise concern as to whether the upper edges of the scatter actually represent the region of most efficient response.

4.2.1. *Visual approach*

This is a heuristic approach of boundary line fitting and is one of the simplest methods initially proposed by Webb (1972). A boundary line can be drawn along the largest data points which are simply identified by eye and the parameters of the fitted line can be obtained by the least square methods. Webb (1972) recognized the statistical difficulties that this method may bring as it has no means of accounting for measurement errors, measures of uncertainty, nor enabling reproducibility. Despite these weaknesses, this method has been applied in many studies involving yield gap analysis due to its simplicity (Abravan et al., 2016; Asten et al., 2003; Baral et al., 2022b,a; Dehkordi et al., 2020; French and Schultz, 1984b,a; Gorjizad et al., 2019; Haefele et al., 2003; Hajjarpoor et al., 2018; Mohammadi-Kashka et al., 2023; Nehbandani et al., 2020; Tagliapietra et al., 2018; Yousefian et al., 2021).

4.2.2. *Binning approach*

The binning approach is a heuristic approach that includes all variants of approaches that categorise the independent variable into ranges of values from which a single value of response variable is determined for each range and used to fit the boundary line (Fig. 6). For a scatter plot of yield against a factor x_i, x_i is divided into n number of sections (bins) and in each section, a boundary value corresponding to a set criteria, which can be the 90th, 95th, 99th or 99.7th percentile, is selected (Casanova et al., 1999; Kintché et al., 2017; Schmidt et al., 2000; van Vugt and Franke, 2018). An appropriate percentile value, is chosen so that it is not too high (e.g. 100th percentile) so as to remove the influence of outlier boundary points but also not too low to avoid cutting off influence larger response values (Schmidt et al., 2000). The selected boundary points (red solid symbols on Fig. 6) are used to fit the boundary line model.

Many authors have adopted this approach for yield gap analysis (Affholder et al., 2013; Casanova et al., 1999; Hoogmoed et al., 2018; Huang et al., 2008; Kintché et al., 2017; Li et al., 2017; Luo et al., 2020; Patrignani et al., 2014; Scarlato et al., 2017; Schmidt et al., 2000; Tasistro, 2012; van Vugt and Franke, 2018; Walworth et al., 1986). Similar to the visual method, this approach includes elements of subjectivity and does not account for measurement error, has no measure of uncertainty in the position of the boundary line and the model has no basis for checking for evidence of a boundary. Subjectivity arises in the selection of bin size and the selection of which quantile value to treat as a boundary point. There is currently no standardized method for bin size selection (Makowski et al., 2007; Milne et al., 2006a). The selection of different sizes of bins affects the number of points selected for fitting the boundary line and in turn, the position and parameters of the boundary line (Makowski et al., 2007).

Shatar and McBratney (2004) suggested a procedure to reduce the effect of arbitrary bin selection. They fix a moving window on the x-axis of the boundary plot of width 1/10 the range of x, then select a boundary value from successive increments of this window over the range. These selected values, a larger subset of the data than for other procedures, are then used to fit the model. Shatar and McBratney (2004) further suggested that measurement error could be accounted for through use of bootstrapping. Several boundary lines (say about 1000) are drawn by resampling and replacement of the data points in the dataset. The range of boundary lines drawn becomes the confidence interval of the boundary line. However, this method of determining the confidence interval is biased because the repeated replacement sampling from the

same data results in a confidence interval always below the upper bounds of the dataset (Milne et al., 2006a).

4.2.3. BOLIDES approach

A third heuristic approach that has been commonly used in yield gap analysis (Bhattarai et al., 2017; Bucagu et al., 2014; Berrueta et al., 2020; Cao et al., 2019; Chen et al., 2018,2019; Fermont et al., 2009; Fu et al., 2021; Guo et al., 2021; Ndabamenye et al., 2013; Rhebergen et al., 2018; Silva et al., 2019; Wairegi et al., 2010; Wang et al., 2015; Zhang et al., 2019b, 2019, 2020) is the BOLIDES algorithm (Schnug et al., 1995). The dataset is first cleaned of outliers using a *neighbourhood density* procedure and then points are selected to which a boundary is fitted. The limitation of this method is that the minimum number of data points around a given point and the size of the neighbourhood is subjectively chosen. This affects the number of outliers in the data and ultimately, the position of the boundary line. There is currently no standard procedure for selecting the neighbourhood and the minimum number of neighbouring points.

The selection of the boundary points is done in a stepwise manner (Fig. 7a). Firstly, the minimum (x_{min}), maximum (x_{max}), and the maximum response point (x_{ymax}) values for the factor, x_i, in the dataset are identified. Starting with the largest response value at x_{min}, the next boundary data point is the data point along the x-axis which encloses the dataset. Further boundary data points along the x-axis are identified until the boundary point equal to x_{ymax} is reached. The boundary points between the x_{max} and x_{ymax} are identified in a similar way starting from x_{max} but moving in the opposite direction. The selected points are then used to fit the boundary line and the parameters are extracted from it (Fig. 7b).

This approach, unlike the visual and binning approach, provides reproducible results if the same criteria are used to define outliers. However, it does not account for measurement error or uncertainty of the position of boundary line and does not provide any means of checking for evidence of the existence of a boundary in the dataset.

4.2.4. Quantile regression approach

Quantile regression models the τ^{th} quantile ($0 < \tau < 1$) of the prediction distribution. The τ^{th} quantile of a random variable Y can be defined as the value y for which the probability of obtaining values less than Y is greater or equal to τ. This concept can similarly be extended to linear regression such that the regression quantiles are estimated as conditional regression for different values of τ (Davino et al., 2014) as:

$$Q_Y(\tau|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_\tau \quad (7)$$

where $\mathbf{Q_Y}(\tau|\mathbf{X})$ is a n x 1 vector of dependent variable Y conditional on τ, β is a $p \times 1$ vector of regression parameters, $\mathbf{X}$ is an $n \times p$ matrix of predictors. Estimates of the coefficients, $\mathbf{b}_\tau$, of β_τ are obtained by minimizing the weighted absolute residual values. The negative residuals are given weight equal to 1 - τ while the positive residuals are given weights equal to τ (Davino et al., 2014).

Quantile regression has been proposed as model for data in which the dependent variable responds in a complex way to interacting covariates. This is the case, for example, with the rate-limiting form of the law of the optimum presented in Equation (2). Under the model, the form of the response of the dependent variable to one covariate can vary, expressed in different functional forms for quantiles of the prediction distribution. Therefore, some upper regression quantile can be regarded as a boundary line and modelled as the quantile regression for some τ (e.g. Baudron et al. 2019 used the 90th percentile (See Box 3) while Wang et al. 2022 used the 95th percentile). The percentile value used to represent the boundary line is typically chosen so that it is not too low to include points that are affected by other unmeasured limiting factors but also not too high to be affected by measurement error (Schmidt et al., 2000). There is however, no inference that can be made to ascertain whether the selected quantile function can be regarded as a boundary. The quantile regression method has been applied in yield gap analysis studies using boundary lines (Edreira et al., 2017; Fink et al., 2022; Grassini et al., 2009; Lollato et al., 2017; Neuhaus and Sadras, 2018; Rizzo et al., 2021; Sadras and Angus, 2006).

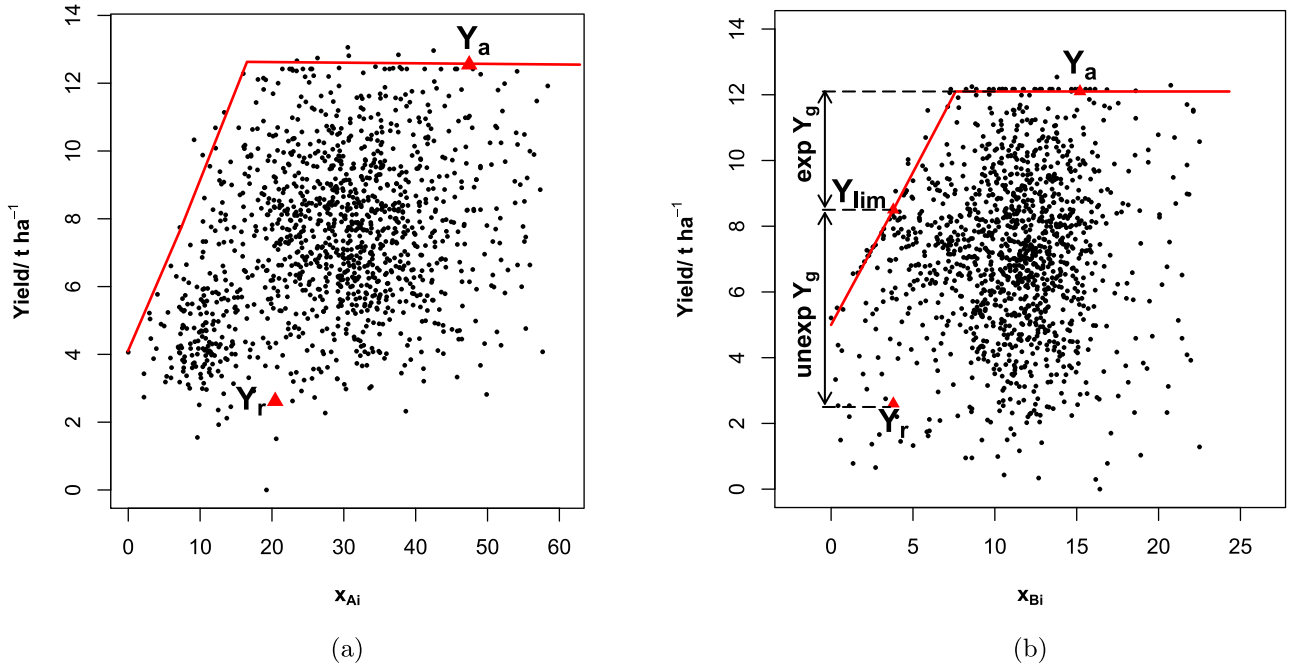
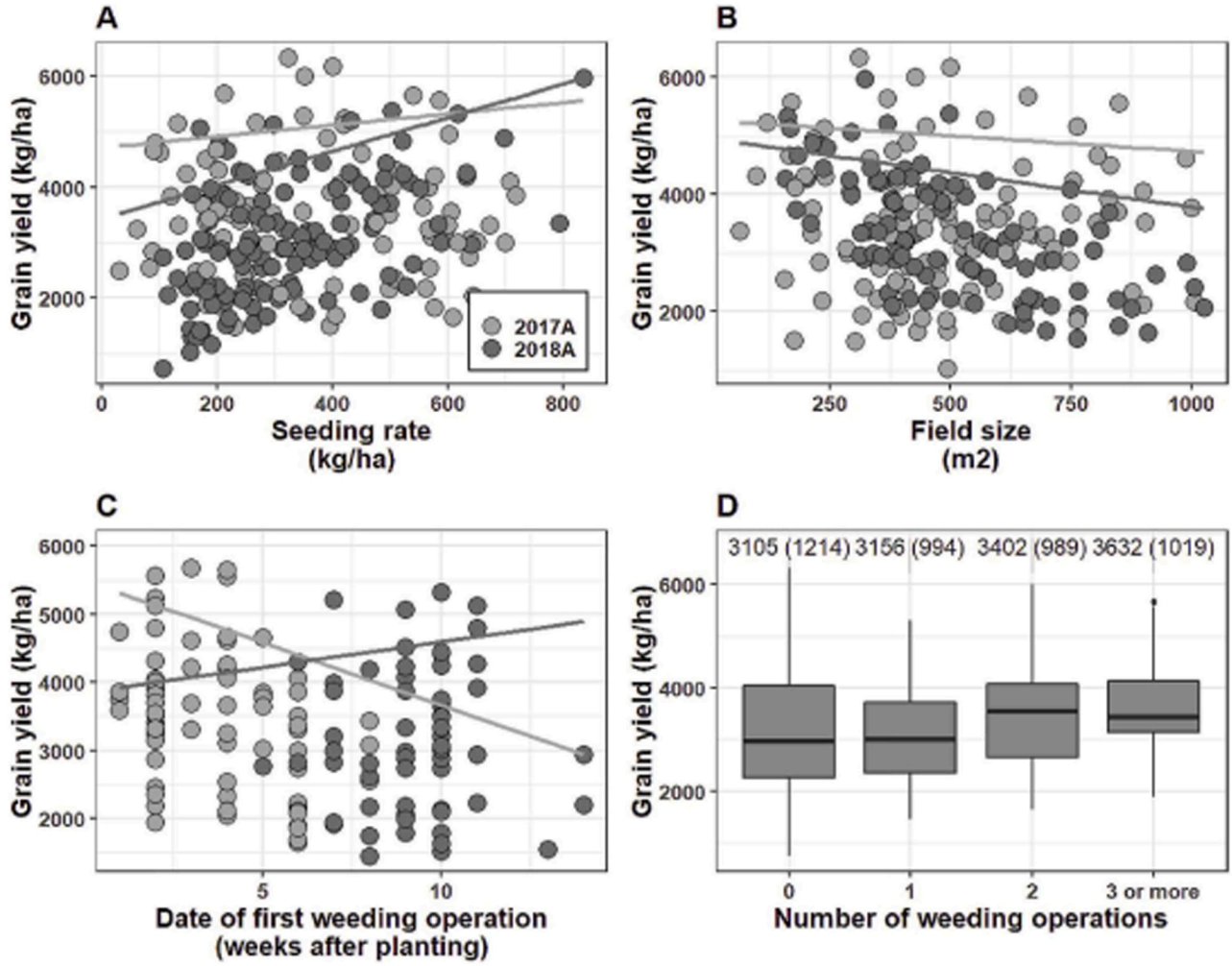


Fig. 5. Decomposition of yield gaps into explained (exp Y_g) and unexplained (unexp Y_g) yield gaps when only two factors, A and B, are considered. Y*r* is actual yields of farm, Y*a* is the attainable yield and Y*lim* is the yield predicted by most limiting factor.

Box 3: How to increase the productivity and profitability of smallholder rainfed wheat in the Eastern African highlands? Northern Rwanda as a case study (Baudron et al., 2019).

Baudron et al. (2019) studied how different factors affect wheat yield in Rwanda with an aim of finding ways of increasing productivity. They used multivariate stochastic frontier analysis in combination with univariate boundary line analysis for this purpose. The factors which were found to significantly affect yield gap were identified by the stochastic frontier analysis and were further evaluated using the boundary line analysis to gain further insight into how they affect yield. Factors including seeding rate, farm size and date of first weed were subjected to boundary line analysis using quantile regression method with the 90th percentile used as the boundary line as shown in Box-Fig. 4.



Box-Fig 4. Wheat grain yield as a function of (A) seeding rate; (B) field size; and (C) date of the first weeding operation; and (D) wheat grain yield as a function of the number of weeding operations. In (A), (B) and (C), the light grey lines and the dark grey lines represent the linear regressions fitted through the 90th percentile of the season 2017A data and the 2018A data, respectively. In (D), means are given for each category, followed by the standard deviation in parentheses.

The positive effect of seeding rate on wheat grain yield during the season 2018A (obtained From SFA) was confirmed by the increasing boundary lines in the univariate boundary line analysis. Similarly, the negative effect of field size on wheat grain yield was confirmed by the decreasing boundary lines. The boundary line was also found to decrease when considering time between planting and the first weeding operation as the independent variable during the season 2017A. Boundary lines in this case provide a tool for studying how the different factors affect the yield and how they can be improved to increase productivity.

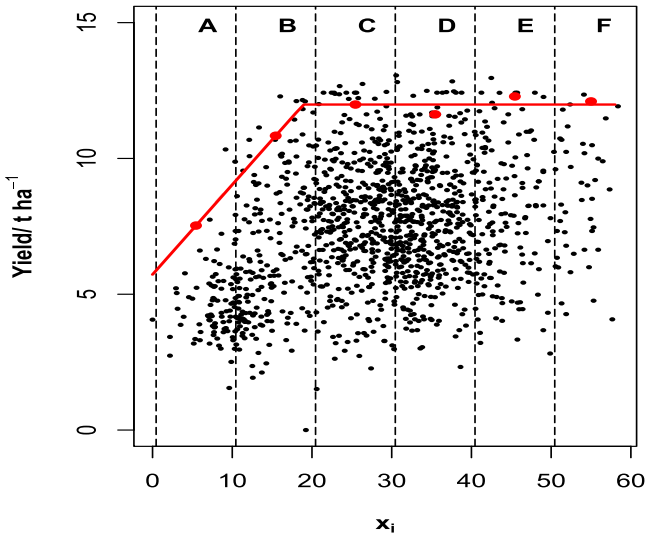


Fig. 6. The process of fitting the boundary line using the binning approach with fixed, non-overlapping bins. Factor x_i is divided into sections A, B, C, D, E and F separated by dashed vertical lines. The red solid symbols, representing the boundary point of each section, are selected as the 95th percentile of the data in each section. The red line is the fitted boundary line model to the selected points.

When the targeted quantile value which corresponds to the boundary line function is known, quantile regression provides a good basis for deriving the boundary line as it uses all the data points in a dataset without the bias of removing or selecting a subset of data points or making arbitrary bins (Makowski et al., 2007). However, the suitable quantile value is usually not chosen objectively and therefore, this method is classified as heuristic. Without an objective method of choosing the quantile value, reproducibility is a challenge. In a recent study, Andrade et al. (2023) used a Bayesian approach called Bayesian segmented quantile regression (BSQR) to fit the boundary lines and determine various critical soil nutrient values in vineyard soils for fertilizer recommendations. They found that the BSQR model was highly sensitive to quantile selection which can directly affect parameter estimates of the boundary line and subsequent critical levels, and sufficiency ranges for nutrients in soils. This sensitivity may lead to variations in the estimated nutrient critical values, potentially affecting the accuracy of fertilization recommendations. This highlights the need for a more objective method for quantile selection. Furthermore, this method does not have any mechanism to test for the presence of an upper boundary.

4.2.5. *Makowski quantile regression method*

The Makowski quantile regression approach is a development of the general quantile regression approach described in Section 4.2.4. The lack of an objective method to decide on the quantile value (τ) for the quantile regression approach is the main challenge of the general quantile regression. Makowski et al. (2007) proposed a method to determine of the appropriate quantile value (τ) statistically from the distributions of measurement error of yield and the limiting factor, using some expert assumptions and judgments. Makowski et al. (2007) propose that the distribution of the measurement error and the limiting factor can be assumed by agronomists and statisticians based on previous studies or expert knowledge. The challenge for using this model is that in most cases where the boundary line method is used for yield gap analysis, data are gathered from surveys or different non-experimental plots without replicate measurements making it difficult to eliminate the measurement errors. In addition to this, the distribution of the limiting factor(s) is also difficult to estimate as it is unknown in reality

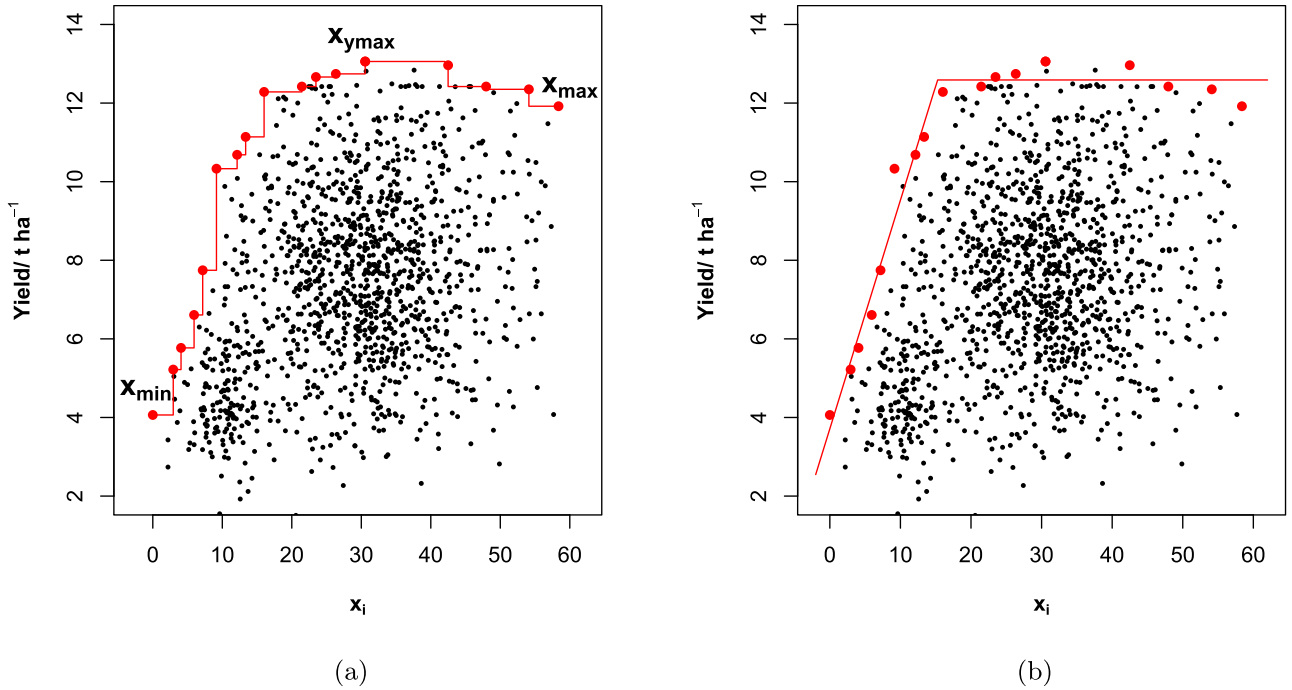


Fig. 7. (a) Process of selecting boundary points using BOLIDES. x_{ymax} is the value of the factor corresponding to the maximum yield, x_{min} corresponds to the largest yield at the smallest value of factor in dataset and x_{max} corresponds to the largest yield at the largest value of factor in dataset. The red dots represents the selected boundary points by implementing BOLIDES (b) The boundary line fitted to the selected points using a broken stick model.

thereby making the estimation of the quantile value very difficult to achieve practically as a lot of information is needed to make the assumptions of the error distribution and the limiting factor distribution. Therefore, significant further work is needed before this can be considered a working method.

4.2.6. Censored bivariate normal model

The censored model approach for setting out the boundary line developed by Milne et al. (2006a), is categorised as a statistical approach as it uses explicit statistical assumptions to set out the boundary line and estimate its parameters. It is based on the principle of a censored bivariate distribution. Parameters comprise the parameters of the bivariate distribution (the mean of the response variable (μ_y), mean of the independent variable (μ_x), the variance of the response variable (σ_y), the variance of the independent variable (σ_x), and correlation (ρ)), the parameters of the boundary lines which acts as the censor of the bivariate model and the measurement error of the response variable. The parameters of the boundary line depend on the chosen model (i.e. linear, polynomial, etc.) as well as the parameter that describes the error of the boundary line (see Box 4 and 5). The model parameters are estimated using the maximum likelihood approach given the available data (Milne et al., 2006a). The likelihood in simple terms will be the best combination of unknown parameters that could have likely produced the available dataset. Therefore, the combination of parameters with the highest likelihood is most likely to have produced the available data. The suitability of the derived model can be tested by comparing it with a simple uncensored bivariate model (null model) using Akaike's information criterion (AIC). The AIC imposes a penalty for additional parameters when comparing with a simple model. If an additional parameter does not improve the model significantly, it is better to use the simple model (null model).

Box 4: Theory of the censored bivariate normal model (Lark and Milne, 2016)

A censored bivariate normal distribution for which the boundary line acts as the upper censor can be described by the function:

$$\mathrm{f}(\mathrm{y},\mathrm{x}) = \phi(\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{C}) \tag{8}$$

where ϕ represents the bivariate normal density function with means $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$, and the vector $\boldsymbol{Z}$ represents the censor function with parameters, $\boldsymbol{\beta}$. Given a boundary model, $b(x) = \overline{y}$, is the censor, a variate {y, x} for which $\mathrm{y} > \overline{y}$ is replaced by variate $\{\overline{y},x\}$. As the variable y is measured with error, $N(\overline{y},\sigma_e)$, and so observations above the censor are only due to measurement error, the variate $\{y, x\}$ can be written as $\{\widetilde{y}, x\}$ to show influence of measurement error. The censored bivariate normal model can be written as a function of three sets of parameters, the censoring parameter, $\boldsymbol{\beta}$ (which represents the boundary line), the parameters of the bivariate random normal distribution (means, $\boldsymbol{\mu}$, and (co)-variance, $\boldsymbol{C}$) and the measurement error, σ_e, of response variable y.

$$\mathrm{f}(\widetilde{\mathrm{y}},\mathrm{x}|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{C},\sigma_\mathrm{e}) \tag{9}$$

To keep things brief, the parameters from the density functions can be dropped. Following the properties of conditional densities, the function $f(\overline{y},\boldsymbol{x})$ can be written as

$$\mathrm{f}(\widetilde{\mathrm{y}},\mathrm{x}) = \mathrm{f}(\widetilde{\mathrm{y}}|\mathrm{x})\mathrm{f}(\mathrm{x}) \tag{10}$$

where $f(x)$ is the probability density function of x. Assuming x is measured without error, as in the general linear model, the conditional density in Equation (10), can be written as a convolution (f^*g) of the two functions as

$$\mathrm{f}(\widetilde{\mathrm{y}},\mathrm{x}) = \mathrm{f}(\overline{\mathrm{y}}|\mathrm{x}) * \mathrm{f}_\mathrm{N}(\mathrm{v}|0,\sigma_\mathrm{e}) \tag{11}$$

The conditional density $f(\overline{y}|x)$ in Equation (11) is the censoring of the conditional density $f(y|x)$ and can, therefore, be written as

$$\mathrm{f}_\mathrm{N}(\mathrm{y}|\mu_{\mathrm{y}|\mathrm{x}},\sigma_{\mathrm{y}|\mathrm{x}}) \tag{12}$$

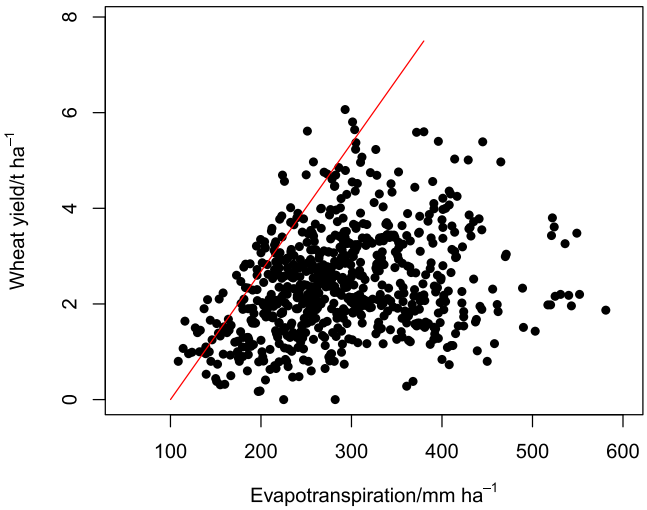
where $\mu_{y|x}$ and $\sigma_{y|x}$ are the conditional mean and standard deviation, respectively, of y. The censored conditional density as a right-censored distribution (upper boundary), can be therefore, be broken down as

$$\mathrm{f}(\overline{\mathrm{y}}|\mathrm{x}) = \begin{cases} \mathrm{f}_\mathrm{N}(\mathrm{y}|\mu_{\mathrm{y}|\mathrm{x}},\sigma_{\mathrm{y}|\mathrm{x}}) & \text{if } \mathrm{y} < \mathrm{b}(\mathrm{x}), \\ \displaystyle\int_{\mathrm{b}(\mathrm{x})}^{\infty} \mathrm{f}_\mathrm{N}(\mathrm{y}|\mu_{\mathrm{y}|\mathrm{x}},\sigma_{\mathrm{y}|\mathrm{x}})\mathrm{dy} & \text{if } \mathrm{y} = \mathrm{b}(\mathrm{x}), \\ 0 & \text{if } \mathrm{y} < \mathrm{b}(\mathrm{x}) \end{cases} \tag{13}$$

For some proposed set of parameters, $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{C}$ and σ_e and a pair of observed values, $\widetilde{y}$ and x, the density can be computed using Equation (13) and therefore, the likelihood given the observations. It can be seen that if there are more observations at the boundary, given the measurement error, the likelihood will be larger. The observations above the boundary line have zero density values hence do not contribute to the likelihood value.

Box 5: Yield gap analysis using the bivariate censored method (FAO and DWFI, 2015)

Data for the actual wheat yield and estimated evapotranspiration in China, Mediterranean Europe, North America, and Australia was collected. The boundary line was fitted using the bivariate censored model of Milne et al. (2006a). In this particular case a linear boundary line model, $y = ax + b$ was assumed and its parameters were estimated using the maximum likelihood approach. The estimated parameter values were $a = 0.025$ with 95 percent confidence interval (0.020, 0.030) and $b = -2.458$ with 95 percent confidence interval (-3.412, -1.504). The variation around the boundary, $\sigma_B = 0.971$, with confidence interval (0.791, 1.151). The suitability of the model was checked using AIC and the boundary line model was found to describe the data better than the null model which was assumed to be the bivariate normal model Box-Fig. 5.



Box-Fig 5. Relationship between estimated evapotranspiration and wheat yield (FAO and DWFI, 2015).

The strength of this approach is that it has explicit statistical assumptions and eliminates the subjectivity. Unlike the heuristic approach, all the data points are used in this approach which is consistent with statistical principles that guard against deliberate or undeliberate removal of data in a datasets. This approach ensures that reproducibility of the boundary line is achieved, accounts for observational errors and the uncertainty of the boundary line position can be derived. Unlike the methods that use the bootstrapping approach, the confidence interval is not limited by the upper bound of the data points. The actual value of the measurement error parameter may not always be available. In such cases, other methods to estimate the measurement error parameter have been used including the use of the nugget variance for a variogram of the response variable (Lark et al., 2020), use of profile likelihood of measurement error (Lark and Milne, 2016) or finding the maximum likelihood estimate along with the other model parameters (Kindred et al., 2015).

4.2.7. *Stochastic frontier approach*

Stochastic frontier analysis (SFA) is a statistical method of economic modelling which has been used in production ecology to fit the upper limit of a response variable given an independent variable(s). It has been applied in yield gap analysis studies in recent studies (Baudron et al., 2019; Dossou-Yovo et al., 2020; Silva et al., 2017b, 2019, 2021). The frontier model describes the maximum output that can be achieved given a level of input(s). Any point below this frontier is a result of a composite error that is made up of the sum inefficiency and random error of measurement as shown in the stochastic frontier equation (14). This is normally expressed in log form

$$\log(Y_r) = \log(f(x_i|\beta) + \epsilon_i) - u_i \tag{14}$$

where $f(x_i|\beta)$ is the frontier model, ϵ is a random error term and u_i, the inefficiency is a positively valued random variable which expresses the effect of factors which reduce the output below the frontier response-equivalent to the efficiency gap described above. Technical efficiency in SFA is defined as the ratio of the observed output, Y_r here, to the term $f(x_i|\beta) + \epsilon$ (Silva et al., 2017a). From equation (14) it follows that

$$\begin{aligned} -u_i &= \log\left(\frac{Y_r}{f(x_i|\beta) + \epsilon}\right) \\ &= \log(E_T) \end{aligned} \tag{15}$$

and so

$$E_T = e^{-u_i} \tag{16}$$

It is assumed that the error term ϵ is normal with zero mean and that u_i is a non-negative number with an exponential or gamma distribution or some truncated distribution such as the half normal with mean $z_i\nu$ and variance σ^2. The latter distribution is commonly preferred (Aigner et al., 1977). The factors that affect inefficiency can further be modelled as

$$u_i = \mathbf{Z}_i^T\mathbf{V} + w_i \tag{17}$$

where z_i is a vector of explanatory variables associated with inefficiency and ν is a vector of unknown coefficients and w_i is a random variable. The unknown parameters in equations (14) and (17) can be estimated simultaneously by maximum likelihood. Besides the maximum likelihood approach, other methods like the corrected ordinary least square, generalized method of moments and Bayesian methods have been used in econometrics. The Bayesian methodology has potential to improve the robustness of modelling, particularly from small datasets, if robust and important prior distributions are available for parameters.

The advantages of SFA are that it is objective and reproducible, uses all the data points, and can account for measurement errors. In addition, this method allows the input of multiple factors to model their simultaneous effect on a response variable. This provides the additional capability of studying yield gaps at farm level rather than just at crop

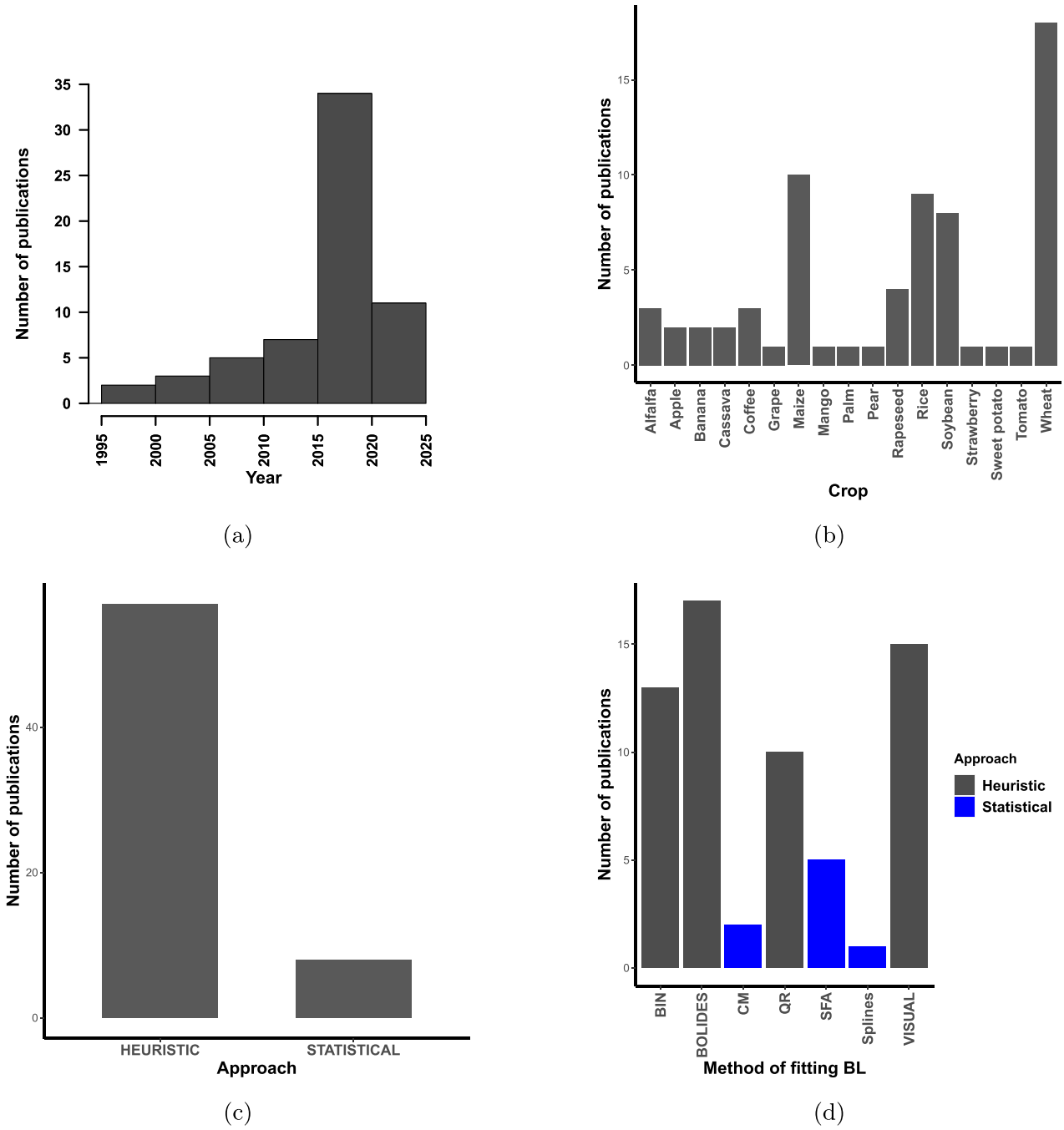


Fig. 8. Number of articles published during the last 28 years (1995–2023) that have utilised the boundary line (BL) approach for yield gap analysis (a), Crops that have been studied (b) and the number of publications that used the different approaches (c) and boundary line fitting methods (d). CM represents the censored model, QR is quantile regression and SFA is the stochastic frontier analysis.

level (Silva et al., 2017b). For example, labour at farm level is distributed amongst different enterprises, therefore, labour can be inputed as an input in the SFA for a crop of interest to determine whether it is the cause of the observed inefficiency (Silva et al., 2017b, 2019). This can help to restructure labour provisions to achieve efficiency at farm level.

5. Trends in the usage of various boundary line fitting approach

There has been an increase in the use of boundary line methodology for yield gap analysis over the past two decades (Fig. 8a). This indicates its importance as a method for yield gap analysis and therefore, the need to standardise the methodology for better interpretation and comparison amongst studies. A variety of crops (alfalfa, apple, banana,cassava,coffee, grape, maize, mango, palm, pear, rapeseed, rice, soybean, sweet potato, tomato and wheat) have been studied using this methodology in different regions of the world (Casanova et al., 1999; Cossani and Sadras, 2018; Fermont et al., 2009; Hajjarpoor et al., 2018; Kintché et al., 2017; Lark et al., 2020; Shatar and McBratney, 2004; Silva et al., 2019; van Vugt and Franke, 2018) with cereals (maize, wheat and rice) accounting for over 50% of the studies (Fig. 8b), Soybean was studied in about 12% of the studies while the rest of the crops accounted for less than 7% each. Different boundary line fitting methods are available and have been used in these studies. The question arises, which approach of fitting the boundary line should one use? Although the statistical

methods provide a more reproducible analysis and allow quantification of uncertainty of the boundary line position, few studies have used them in general (e.g. Kindred et al., 2015; Lark and Milne, 2016; Lark et al., 2020). More recent studies for yield gap analysis have continued to use the heuristic approaches because of their simplicity to execute (e.g. Gorjizad et al., 2019; Hajjarpoor et al., 2018; Nezamzade et al., 2020).

Of the peer-reviewed articles that were selected during the search for this review (n = 64), 58 used a heuristic approach for fitting boundary lines while only 8 used the statistical approach (two studies combined two methods) (Fig. 8c). BOLIDES was the most commonly used method, 18 studies, while the binning, visual and the quantile regression methods were used in 13, 15 and 10 studies respectively (Fig. 8d). The statistical methods, censored model and SFA were used in two and five studies respectively. Two studies (Duan et al., 2022; Wairegi et al., 2018) did not clearly indicate which method they used. One study fitted the boundary line using splines (Niang et al., 2017). It is not known whether there are systematic and important differences in the outcome of the boundary line analysis if statistical rather than heuristic approaches are used or when different statistical approaches are used. There is, therefore, a need for comparative studies to examine the differences and to assess their practical importance. The choice of a method to set out the boundary line for use in yield gap analysis may depend on various factors. These may include the objective of the study, usability (simplicity/complexity) and availability of the method, and the amount of data available.

The researcher's objective is an important determinant of the approach which is used. Many authors have concluded that the boundary line is most useful for checking the relative importance of factors in yield gap analysis (Shatar and McBratney, 2004; Silva et al., 2019). However, boundary line analysis allows us to draw stronger quantitative conclusions. An example is the use of the boundary line to verify some standard nutrient guidelines. Evanylo et al. (1987) used the boundary line approach to determine the soil critical values of Ca, P, K and Mg for soyabean on fine and coarse textured soils. Similarly, Lark et al. (2020) used the statistical censored model method for fitting boundary line to check if the model outcomes were consistent with RB209 index values for P, K and Mg requirements for wheat production in the soils of the United Kingdom. Furthermore, they compared boundary lines for available phosphorus for different subsets defined on pH to check its effect on the boundary line parameters. A similar study by Andrade et al. (2023) used boundary line methodology to obtaining reference values for nutrients in vineyard soils. These kind of studies may require precise parameter estimation with confidence intervals so statistical rather than heuristic approaches may be advantageous. The use of Bayesian statistical methods such as the BSQR proposed by Andrade et al. (2023) as well as the censored bivariate normal model by (Milne et al., 2006b) provide ability to attach uncertainty to the boundary line parameters if all other bottlenecks to their use are resolved. The censored bivariate normal model can surely benefit from application of Bayesian approaches to estimate the measurement error value which is an input but rarely available in most datasets.

The usability of a method is another determining factor of which approach to use. Though statistical approaches provide more robustness than heuristic approaches, they may be more complex to use which may make it difficult for researchers with a limited statistical background (Harris and Smith, 2009). As an example, the censored model method of Milne et al. (2006a) requires initial parameter values which may provide a challenge for agronomists with limited statistical experience especially when the boundary line model is non-linear. Another statistical approach, the quantile regression of Makowski et al. (2007), has not been developed into a full working method as it requires some strong assumptions about the distribution of the errors and limiting factor(s). It is, therefore, vital to come up with interactive tools that help the end-users (researchers or agronomists in this case) utilise and make good interpretations if statistical methods are to be fully utilised. For instance, many studies (Gorjizad et al., 2019; Hajjarpoor et al., 2018; Nehbandani et al., 2020; Nezamzade et al., 2020) have used the visual method, a heuristic approach, for fitting boundary line despite recognising the availability of statistical approaches like the censored model of Milne et al. (2006a) and quantile regression of Makowski et al. (2007) citing simplicity as the reason for selecting the visual method. The BOLIDES has also been widely used maybe because it is easier to follow and probably because of the availability of a software that executes the process.

The amount of data available is another factor that is vital in the choice of method to use for boundary line fitting and may render some methods suitable or unsuitable. While the heuristic methods can be easily executed with relatively fewer data available, some statistical approaches may result in the poor fitting of a boundary line. Approaches that utilise the maximum likelihood method in parameter estimation (e.g. censored model and stochastic frontier analysis) require that there is sufficient amounts of data available as its outcome is dependent on the observations in the dataset (Myung, 2003). Without sufficient data these methods will not converge on a suitable boundary. Although the heuristic methods can easily be implemented using a smaller datasets, it does not make them better than statistical methods in these cases. The use of Bayesian methods can help overcome the challenge of less data availability as it uses some prior knowledge and this is vital in cases where there is limited data. Bayesian approaches have not been fully utilised for yield gap analysis despite their ability to resolve the challenge of lack of sufficient data. Though the addition of the expert opinion and knowledge may enhance the results of the model, there is a need to be cautious when including expert opinion in models as this may lead to unsatisfactory results if poorly implemented. Expert information may contain personal biases and preferences that reduce its objectivity and reliability. Data on qualitative factors (e.g. crop variety and manure type among others) have not been incorporated into yield gap analysis using boundary lines (see Table S.2) except when SFA has been applied (Baudron et al., 2019; Silva et al., 2017b,a). However, these are important determinants of yield gaps along with quantitative factors. Further research is required to enable the addition of qualitative factors in the yield gap process using boundary lines. This may be a challenge for statistical approaches like the censored model of Milne et al. (2006a) which works on the assumption the data follows a bivariate normal distribution.

It is also important to note that the choice between the statistical and heuristic approaches may affect the conclusion of yield gap analysis using boundary lines when interpreted as the law of the minimum. When statistical methods, which allow for the computation of confidence interval of the boundary line, are used, boundary lines with confidence intervals that do not overlap can be concluded to be different and the boundary line that predicts the minimum response can easily be identified. However, if the confidence intervals of boundary lines are overlapping, further statistics are needed to check which boundary line predicts the minimum response. This is not possible when heuristic methods are used. Studies that interpreted the boundary line as the law of the minimum while utilising the heuristic approaches in setting out the boundary line (e.g. Casanova et al., 1999; Wairegi et al., 2010), may have different outcomes if statistical approach which account for confidence interval of boundary lines were used.

6. Conclusion

The methods for fitting boundary lines for yield gap analysis have been identified with the heusteric methods being commonly used (89%). The selection of boundary line fitting method has been identified to be affected by different factors that include the objective of the study, the usability of the method (simplicity/complexity) and data availability. The heuristic methods are often simpler to use and this is likely to be the reason they are more commonly adopted. However, the statistical methods provide a more robust approach despite being more complex. It is, therefore, important that interactive tools are developed that can help

facilitate the use of statistical methods by researchers and agronomists. A second issue with statistical methods, especially those based on the maximum likelihood, is that they require large amounts of data to converge. The use of Bayesian statistics provide a solution for this data availability challenge because it incorporates prior knowledge of distributions in its implementation. Nonetheless, there is a gap in knowledge on how the statistical and heuristic methods of setting out the boundary line compare (as well as how different statistical methods compare), if there are important systematic differences, and how they affect outcome of the yield gap analysis and its interpretation. We therefore, recommend a comparative study using various data sets is needed to explore the strengths and weaknesses of statistical and heuristic methods. Further, there is a lack of exploratory tools in the boundary line framework that evaluate evidence of bounding effects in a plot despite it being a vital initial step for boundary line modelling.

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

No data was used for the research described in the article.

Acknowledgements

This work was funded by the Nottingham-Rothamsted Future Food Beacon Studentships in International Agricultural Development. Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom (BBSRC). We acknowledge support from the Growing Health Institute Strategic Programme [BB/X010953/1;BBS/E/RH/230003C].

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fcr.2024.109365.

References

Abravan, P., Soltani, A., Majidian, M., Mohsenabadi, G., 2016. Factors limiting canola yield and determining their optimum range by boundary line analysis. Iioab J. 7 (8), 161–167.

Affholder, F., Poeydebat, C., Corbeels, M., Scopel, E., Tittonell, P., 2013. The yield gap of major food crops in family agriculture in the tropics: assessment and analysis through field surveys and modelling. Field Crops Res. 143, 106–118. https://doi.org/10.1016/j.fcr.2012.10.021.

Aigner, D.L., Lovell, C.A., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. J. Econ. 6 (1), 21–37. https://doi.org/10.1016/0304-4076(77)90052-5.

Andrade, C.B., Comin, J.J., Moura-Bueno, J.M., Brunetto, G., 2023. Obtaining reference values for nutrients in vineyard soils through boundary line approach using Bayesian segmented quantile regression on commercial farm data. Eur. J. Agron. 150, 126928.

Asten, P.J.V., Wopereis, M.C., Haefele, S., Isselmou, M.O., Kropff, M.J., 2003. Explaining yield gaps on farmer-identified degraded and non-degraded soils in a Sahelian irrigated rice scheme. Neth. J. Agric. Sci. 50, 277–296. https://doi.org/10.1016/s1573-5214(03)80013-1.

Aulakh, M.S., Malhi, S.S., 2005. Interactions of nitrogen with other nutrients and water: Effect on crop yield and quality, nutrient use efficiency, carbon sequestration, and environmental pollution. Vol. 86. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, Cambridge, MA, USA, pp. 341–409. https://doi.org/10.1016/S0065-2113(05)86007-9. Vol. 86.

Baral, R., Bhandari, K., Kumar, R., Min, D., 2022a. Yield gap analysis of alfalfa grown under rainfed condition in Kansas. Agronomy 12 (9), 2190.

Baral, R., Lollato, R.P., Bhandari, K., Min, D., 2022b. Yield gap analysis of rainfed alfalfa in the United States. Front. Plant Sci. 2492.

Baudron, F., Ndoli, A., Habarurema, I., Silva, J.V., 2019. How to increase the productivity and profitability of smallholder rainfed wheat in the Eastern African highlands? Northern Rwanda as a case study. Field Crops Res. 236, 121–131. https://doi.org/10.1016/j.fcr.2019.03.023.

Berrueta, C., Heuvelink, E., Giménez, G., Dogliotti, S., 2020. Estimation of tomato yield gaps for greenhouse in Uruguay. Sci. Hortic. 265, 109250.

Bhattarai, S., Alvarez, S., Gary, C., Rossing, W., Tittonell, P., Rapidel, B., 2017. Combining farm typology and yield gap analysis to identify major variables limiting yields in the highland coffee systems of Llano Bonito, Costa Rica. Agric., Ecosyst. Environ. 243, 132–142. https://doi.org/10.1016/j.agee.2017.04.016.

Bucagu, C., Vanlauwe, B., Van Wijk, M.T., Giller, K.E., 2014. Resource use and food self-sufficiency at farm scale within two agro-ecological zones of Rwanda. Food Secur. 6 (5), 609–628. https://doi.org/10.1007/s12571-014-0382-0.

Cao, H., Li, Y., Chen, G., Chen, D., Qu, H., Ma, W., 2019. Identifying the limiting factors driving the winter wheat yield gap on smallholder farms by agronomic diagnosis in North China Plain. J. Integr. Agric. 18, 1701–1713. https://doi.org/10.1016/S2095-3119(19)62574-8.

Carciochi, W.D., Sadras, V.O., Pagani, A., Ciampitti, I.A., 2020. Co-limitation and stoichiometry capture the interacting effects of nitrogen and sulfur on maize yield and nutrient use efficiency. Eur. J. Agron. 113, 125973.

Casanova, D., Goudriaan, J., Bouma, J., Epema, G., 1999. Yield gap analysis in relation to soil properties in direct-seeded flooded rice. Geoderma 91 (3-4), 191–216. https://doi.org/10.1016/S0016-7061(99)00005-1.

Chen, G., Cao, H., Chen, D., Zhang, L., Zhao, W., Zhang, Y., Zhang, F., 2019. Developing sustainable summer maize production for smallholder farmers in the North China Plain: An agronomic diagnosis method. J. Integr. Agric. 18, 1667–1679. https://doi.org/10.1016/S2095-3119(18)62151-3.

Chen, G., Cao, H., Liang, J., Ma, W., Guo, L., Zhang, S., Zhang, F., 2018. Factors affecting nitrogen use efficiency and grain yield of summer maize on smallholder farms in the North China Plain. Sustainability 10 (2), 363.

Cossani, C.M., Sadras, V.O., 2018. Chapter six - water-nitrogen colimitation in grain crops. Vol. 150. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press Inc, Cambridge, MA, USA, pp. 231–274. https://doi.org/10.1016/bs.agron.2018.02.004. Vol. 150.

Cossani, C.M., Slafer, G.A., Savin, R., 2010. Co-limitation of nitrogen and water, and yield and resource-use efficiencies of wheat and barley. Crop Pasture Sci. 61, 844–851. https://doi.org/10.1071/CP10018.

Davino, C., Furno, M., Vistocco, D., 2014. Quantile regression: Theory and applications. John Wiley and Sons.

Dehkordi, P.A., Nehbandani, A., Hassanpour-bourkheili, S., Kamkar, B., 2020. Yield gap analysis using remote sensing and modelling approaches: Wheat in the Northwest of Iran. Int. J. Plant Prod. 14, 443–452. https://doi.org/10.1007/s42106-020-00095-4.

de Wit, C.T., 1992. Resource use efficiency in agriculture. Agric. Syst. 40, 125–151. https://doi.org/10.1016/0308-521X(92)90018-J.

Dossou-Yovo, E.R., Vandamme, E., Dieng, I., Johnson, J.-M., Saito, K., 2020. Decomposing rice yield gaps into efficiency, resource and technology yield gaps in sub-Saharan Africa. Field Crops Res. 258, 107963.

Duan, Z., Zheng, C., Zhao, S., Feyissa, T., Merga, T., Jiang, Y., Zhang, W., 2022. Cold climate during bud break and flowering and excessive nutrient inputs limit apple yields in Hebei Province, China. Horticulturae 8 (12), 1131.

Duncan, E.G., O'Sullivan, C.A., Roper, M.M., Biggs, J.S., Peoples, M.B., 2018. Influence of co-application of nitrogen with phosphorus, potassium and sulphur on the apparent efficiency of nitrogen fertiliser use, grain yield and protein content of wheat: review. Field Crops Res. 226, 56–65. https://doi.org/10.1016/j.fcr.2018.07.010.

Edreira, J.I.R., Mourtzinis, S., Conley, S.P., Roth, A.C., Ciampitti, I.A., Licht, M.A., Kandel, H., Kyveryga, P.M., Lindsey, L.E., Mueller, D.S., et al., 2017. Assessing causes of yield gaps in agricultural areas with diversity in climate and soils. Agric. For. Meteorol. 247, 170–180.

Elliott, J.A., de Jong, E., 1993. Prediction of field denitrification rates: a boundary-line approach. Soil Sci. Soc. Am. J. 57, 82–87. https://doi.org/10.2136/sssaj1993.03615995005700010016x.

Evanylo, G.K., Sumner, M.E., Francis, T., 1987. Utilization of the boundary line approach in the development of soil nutrient norms for soybean production. Commun. Soil Sci. Plant Anal. 18, 1397–1401. https://doi.org/10.1080/00103628709367906.

FAO and 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 Report, 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. https://doi.org/10.1016/j.fcr.2009.01.009.

Fink, K.P., Grassini, P., Rocateli, A., Bastos, L.M., Kastens, J., Ryan, L.P., Lin, X., Patrignani, A., Lollato, R.P., 2022. Alfalfa water productivity and yield gaps in the US Central Great Plains. Field Crops Res. 289, 108728.

French, R.J., Schultz, J.E., 1984a. Water Use Efficiency of Wheat in a Mediterranean-type Environment 1. The Relation between Yield. Water Use Clim. 35, 743–764.

French, R.J., Schultz, J.E., 1984b. Water Use Efficiency of Wheat in a Mediterranean-type Environment 2. Some Limitations to Efficiency. Aust. J. Agric. Res. 35, 765–775.

Fu, H., Ma, Q., Ma, Z., Hu, Y., Liu, F., Chen, K., Pan, W., Tang, S., Zhang, X., Wu, L., 2021. Quantifying key internal and external yield-limiting factors for Chinese pear in smallholder dominant areas. HortScience 56, 1395–1401. https://doi.org/10.21273/HORTSCI16115-21.

Giller, K.E., Delaune, T., Silva, J.V., Descheemaeker, K., van de Ven, G., Schut, A.G., van Ittersum, M.K., 2021. The future of farming: Who will produce our food? Food Secur. 13, 1073–1099. https://doi.org/10.1007/s12571-021-01184-6.

Gorjizad, A., Dastan, S., Soltani, A., Norouzi, H.A., 2019. Large scale assessment of the production process and rice yield gap analysis by comparative performance analysis and boundary-line analysis methods. Ital. J. Agron. 14, 123–131. https://doi.org/10.4081/ija.2019.1174.

Grassini, P., Yang, H., Cassman, K.G., 2009. Limits to maize productivity in Western Corn-Belt: a simulation analysis for fully irrigated and rainfed conditions. Agric. For. Meteorol. 149, 1254–1265. https://doi.org/10.1016/j.agrformet.2009.02.012.
Guo, X., Shukla, M.K., Wu, D., Chen, S., Li, D., Du, T., 2021. Plant density, irrigation and nitrogen management: three major practices in closing yield gaps for agricultural sustainability in North-West China. Front. Agric. Sci. Eng. 8 (4), 525–544.
Haefele, S., Wopereis, M., Ndiaye, M., Barro, S., and Isselmou, M.O. (2003). Internal nutrient efficiencies, fertilizer recovery rates and indigenous nutrient supply of irrigated lowland rice in Sahelian West Africa.
Hajjarpoor, A., Soltani, A., Zeinali, E., Kashiri, H., Aynehband, A., Vadez, V., 2018. Using boundary line analysis to assess the on-farm crop yield gap of wheat. Field Crops Res. 225, 64–73. https://doi.org/10.1016/j.fcr.2018.06.003.
Harris, E.F., Smith, R.N., 2009. Accounting for measurement error: a critical but often overlooked process. Arch. Oral. Biol. 54. https://doi.org/10.1016/j.archoralbio.2008.04.010.
Hoogmoed, M., Neuhaus, A., Noack, S., Sadras, V.O., 2018. Benchmarking wheat yield against crop nitrogen status. Field Crops Res. 222, 153–163.
Huang, X., Wang, L., Yang, L., Kravchenko, A.N., 2008. Management effects on relationships of crop yields with topography represented by wetness index and precipitation. Agron. J. 100 (5), 1463–1471. https://doi.org/10.2134/agronj2007.0325.
Jahan, M., Amiri, M.B., 2018. Optimizing application rate of nitrogen, phosphorus and cattle manure in wheat production: an approach to determine optimum scenario using response-surface methodology. J. Soil Sci. Plant Nutr. 18 (1), 13–26.
Kindred, D.R., Milne, A.E., Webster, R., Marchant, B.P., Sylvester-Bradley, R., 2015. Exploring the spatial variation in the fertilizer-nitrogen requirement of wheat within fields. J. Agric. Sci. 153, 25–41. https://doi.org/10.1017/S0021859613000919.
Kintché, K., Hauser, S., Mahungu, N.M., Ndonda, A., Lukombo, S., Nhamo, N., 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. https://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. https://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. https://doi.org/10.1111/ejss.12318.
Li, T., Hao, X., Kang, S., 2017. Spatial variability of grape yield and its association with soil water depletion within a vineyard of arid North West China. Agric. Water Manag. 179, 158–166. https://doi.org/10.1016/j.agwat.2016.05.006.
Liebig, J., 1840. Organic chemestry in its application to agriculture and physiology. Friedrich Vieweg und Sohn Publ. Co, Braunschweig, Germany.
Liebscher, G., 1895. Untersuchungen über die bestimmung des düngerbedürfnisses der ackerböden und kulturpflanzen. J. Landwirtsch. 43, 49–125.
Lollato, R.P., Edwards, J.T., Ochsner, T.E., 2017. Meteorological limits to winter wheat productivity in the US Southern Great Plains. Field Crops Res. 203, 212–226.
Luo, N., Wang, X., Hou, J., Wang, Y., Wang, P., Meng, Q., 2020. Agronomic optimal plant density for yield improvement in the major maize regions of China. Crop Sci. 60 (3), 1580–1590. https://doi.org/10.1002/csc2.20000.
Makowski, D., Doré, T., Monod, H., 2007. A new method to analyse relationships between yield components with boundary lines. Agron. Sustain. Dev. 27, 119–128. https://doi.org/10.1051/agro:2006029.
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. https://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. https://doi.org/10.1111/j.1744-7348.2006.00085.x.
Mohammadi-Kashka, F., Pirdashti, H., Tahmasebi-Sarvestani, Z., Motevali, A., Nadi, M., Aghaeipour, N., 2023. Integrating life cycle assessment (LCA) with boundary line analysis (bla) to reduce agro-environmental risk of crop production: a case study of soybean production in Northern Iran. Clean Technol. Environ. Policy 1–20.
Mueller, N.D., Binder, S., 2015. Closing yield gaps: consequences for the global food supply, environmental quality & food security. Daedalus 144 (4), 45–56.
Myers, R.H., Montgomery, D.C., Anderson-Cook, C.M., 2016. Response surface methodology: process and product optimization using designed experiments. John Wiley & Sons.
Myung, I.J., 2003. Tutorial on maximum likelihood estimation. J. Math. Psychol. 47 (1), 90–100.
Ndabamenye, T., Asten, P.J.V., Blomme, G., Vanlauwe, B., Uzayisenga, B., Annandale, J. G., Barnard, R.O., 2013. Nutrient imbalance and yield limiting factors of low input East African highland banana (Musa spp. AAA-EA) cropping systems. Field Crops Res. 147, 68–78. https://doi.org/10.1016/j.fcr.2013.04.001.
Nehbandani, A., Soltani, A., Hajjarpoor, A., Dadrasi, A., Nourbakhsh, F., 2020. Comprehensive yield gap analysis and optimizing agronomy practices of soybean in Iran. J. Agric. Sci. 158, 739–747. https://doi.org/10.1017/S0021859621000241.
Neuhaus, A., Sadras, V.O., 2018. Relationship between rainfall-adjusted nitrogen nutrition index and yield of wheat in Western Australia. J. Plant Nutr. 41 (20), 2637–2643.
Nezamzade, E., Soltani, A., Dastan, S., Ajamnoroozi, H., 2020. View of Factors causing yield gap in rape seed production in the East of Mazandaran Province, Iran. Ital. J. Agron. 15, 10–19. https://doi.org/10.4081/ija.2020.1280.
Niang, A., Becker, M., Ewert, F., Dieng, I., Gaiser, T., Tanaka, A., others, 2017. Variability and determinants of yields in rice production systems of West Africa. Field Crops Res. 207, 1–12.
Patrignani, A., Lollato, R.P., Ochsner, T.E., Godsey, C.B., Edwards, J.T., 2014. Yield gap and production gap of rainfed winter wheat in the Southern Great Plains. Agron. J. 106 (4), 1329–1339. https://doi.org/10.2134/agronj14.0011.
Poorter, H., Anten, N.P., Marcelis, L.F., 2013. Physiological mechanisms in plant growth models: do we need a supra-cellular systems biology approach? Plant Cell Environ. 36 (9), 1673–1690.
Rhebergen, T., Fairhurst, T., Whitbread, A., Giller, K.E., Zingore, S., 2018. Yield gap analysis and entry points for improving productivity on large oil palm plantations and smallholder farms in Ghana. Agric. Syst. 165, 14–25. https://doi.org/10.1016/j.agsy.2018.05.012.
Riar, A., Gill, G., McDonald, G., 2016. Effect of post-sowing nitrogen management on co-limitation of nitrogen and water in canola and mustard. Field Crops Res. 198, 23–31.
Rizzo, G., Monzon, J.P., Ernst, O., 2021. Cropping system-imposed yield gap: Proof of concept on soybean cropping systems in Uruguay. Field Crops Res. 260, 107944.
Rousseeuw, P.J., Ruts, I., Tukey, J.W., 1999. The bagplot: a bivariate boxplot. Am. Stat. 53 (4), 382–387. https://doi.org/10.1080/00031305.1999.10474494.
Sadras, V.O., Angus, J.F., 2006. Benchmarking water-use efficiency of rainfed wheat in dry environments. Aust. J. Agric. Res. 57, 847–856. https://doi.org/10.1071/AR05359.
Salawu, I., Adeyemi, R., Aremu, T., 2007. Modified inverse polynomial and ordinary polynomial as a response surface model: A case study of nitrogen, phosphate, and potassium level on the yield of maize. Int. J. Pure Appl. Sci. (IJPAS) 1 (3), 18–24.
Scarlato, M., Giménez, G., Lenzi, A., Borges, A., Bentancur, Ó., Dogliotti, S., et al., 2017. Analysis and hierarchization of factors explaining strawberry cultivation yield gap in Southern Uruguay. Agrociencia (Montevideo) 21 (1), 43–57.
Schmidt, U., Thöni, H., Kaupenjohann, M., 2000. Using a boundary line approach to analyze N_2 O flux data from agricultural soils. Nutr. Cycl. Agroecosystems 57, 119–129. https://doi.org/10.1023/A:1009854220769.
Schnug, E., Heym, J.M., and Murphy, D.P.L. (1995). Boundary line determination technique (bolides).In: Robert, P.C., Rust, R.H., and Larson, W.E., (Eds.), site specific Management for Agricultural Systems, 899-908.Wiley Online Library.10.2134/1995.site-specificmanagement.c66.
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.
Shatar, T.M., McBratney, A.B., 2004. Boundary-line analysis of field-scale yield response to soil properties. J. Agric. Sci. 142, 553–560. https://doi.org/10.1017/S0021859604004642.
Silva, J.V., Baudron, F., Reidsma, P., Giller, K.E., 2019. Is labour a major determinant of yield gaps in sub-Saharan Africa? A study of cereal-based production systems in Southern Ethiopia. Agric. Syst. 174, 39–51. https://doi.org/10.1016/j.agsy.2019.04.009.
Silva, J.V., Reidsma, P., van Ittersum, M.K., 2017b. Yield gaps in Dutch arable farming systems: analysis at crop and crop rotation level. Agric. Syst. 158, 78–92. https://doi.org/10.1016/j.agsy.2017.06.005.
Silva, J.V., Reidsma, P., Baudron, F., Jaleta, M., Tesfaye, K., van Ittersum, M.K., 2021. Wheat yield gaps across smallholder farming systems in Ethiopia. Agron. Sustain. Dev. 41. https://doi.org/10.1007/s13593-020-00654-z.
Silva, J.V., Reidsma, P., Laborte, A.G., van Ittersum, M.K., 2017a. Explaining rice yields and yield gaps in Central Luzon, Philippines: an application of stochastic frontier analysis and crop modelling. Eur. J. Agron. 82, 223–241. https://doi.org/10.1016/j.eja.2016.06.017.
Skiena, S.S., 2008. The Algorithm Design Manual. Springer.
Sprengel, C., 1826. About plant humus, humic acid and salts of humic acids. Arch. Gesammt Nat. 8, 145–220.
Tagliapietra, E.L., Streck, N.A., daRocha, T.S.M., Richter, G.L., daSilva, M.R., Cera, J.C., Zanon, A.J., 2018. Optimum leaf area index to reach soybean yield potential in subtropical environment. Agron. J. 110 (3), 932–938. https://doi.org/10.2134/agronj2017.09.0523.
Tasistro, A., et al., 2012. Use of boundary lines in field diagnosis and research for Mexican farmers. Better Crops Plant Food 96 (2), 11–13.
Timsina, J., Wolf, J., Guilpart, N., van Bussel, L.G., Grassini, P., van Wart, J., van Ittersum, M.K., 2018. Can Bangladesh produce enough cereals to meet future demand? Agric. Syst. 163, 36–44. https://doi.org/10.1016/j.agsy.2016.11.003.
Tittonell, P., Giller, K.E., 2013. When yield gaps are poverty traps: the paradigm of ecological intensification in African smallholder agriculture. Field Crops Res. 143, 76–90. https://doi.org/10.1016/j.fcr.2012.10.007.
van Ittersum, M.K., Bussel, L.G.V., Wolf, J., Grassini, P., Wart, J.V., Guilpart, N., Cassman, K.G., 2016. Can sub-Saharan Africa feed itself? Proc. Natl. Acad. Sci. USA 113, 14964–14969. https://doi.org/10.1073/pnas.1610359113.
van Ittersum, M.K., Cassman, K.G., Grassini, P., Wolf, J., Tittonell, P., Hochman, Z., 2013. Yield gap analysis with local to global relevance-a review. Field Crops Res. 143, 4–17. https://doi.org/10.1016/j.fcr.2012.09.009.
van Ittersum, M.K., Rabbinge, R., 1997. Field crops research concepts in production ecology for analysis and quantification of agricultural input-output combinations. Field Crops Res. 52, 197–208. https://doi.org/10.1016/S0378-4290(97)00037-3.
van Vugt, D., Franke, A.C., 2018. Exploring the yield gap of orange-fleshed sweet potato varieties on smallholder farmers' fields in Malawi. Field Crops Res. 221, 245–256. https://doi.org/10.1016/j.fcr.2017.11.028.
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. https://doi.org/10.1016/j.fcr.2010.02.010.
Wairegi, L.W., Bennett, M., Nziguheba, G., Mawanda, A., de los Rios, C., Ampaire, E., van Asten, P.J., 2018. Sustainably improving Kenya's coffee production needs more participation of younger farmers with diversified income. J. Rural Stud. 63, 190–199.

Walworth, J., Letzsch, W., Sumner, M., 1986. Use of boundary lines in establishing diagnostic norms. Soil Sci. Soc. Am. J. 50 (1), 123–128.

Wang, N., Chen, H., Ding, D., Zhang, T., Li, C., Luo, X., Siddique, K.H., 2022. Plastic film mulching affects field water balance components, grain yield, and water productivity of rainfed maize in the Loess Plateau, China: a synthetic analysis of multi-site observations. Agric. Water Manag. 266, 107570 https://doi.org/10.1016/j.agwat.2022.107570.

Wang, N., Jassogne, L., van Asten, P.J., Mukasa, D., Wanyama, I., Kagezi, G., Giller, K.E., 2015. Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. Eur. J. Agron. 63, 1–11. https://doi.org/10.1016/j.eja.2014.11.003.

Webb, R.A., 1972. Use of the boundary line in the analysis of biological data. J. Hortic. Sci. 47 (3), 309–319. https://doi.org/10.1080/00221589.1972.11514472.

Yousefian, M., Soltani, A., Dastan, S., Ajamnoroozie, H., 2021. Yield gap assessment in rice-grown fields using CPA and BLA approaches in Northern Iran. Int. J. Plant Prod. 15, 203–217. https://doi.org/10.1007/s42106-020-00128-.

Zhang, Z., Cong, R., Ren, T., Li, H., Zhu, Y., Lu, J., 2020. Optimizing agronomic practices for closing rapeseed yield gaps under intensive cropping systems in China. J. Integr. Agric. 19, 1241–1249. https://doi.org/10.1016/S2095-3119(19)62748-6.

Zhang, D., Wang, C., Li, X., 2019. Yield gap and production constraints of mango (Mangifera indica) cropping systems in Tianyang County, China. J. Integr. Agric. 18, 1726–1736. https://doi.org/10.1016/S2095-3119(18)62099-4.

Zhang, D., Wang, C., Li, X., Yang, X., Zhao, L., Xia, S., 2019b. Correlation of production constraints with the yield gap of apple cropping systems in Luochuan County, China. J. Integr. Agric. 18 (8), 1714–1725.

Chapter 3

Exploration of Data for Analysis Using Boundary Line Methodology

The chapter has been published as:
Miti. C, Milne. A.E, Giller. K.E Sadras. V.O, Lark. RM (2024). Exploration of Data for Analysis using Boundary Lines. *Computers and Electronics in Agriculture Journal*, 219, 108794. https://doi.org/10.1016/j.compag.2024.108794.

3.1 Chapter 3 Overview

This chapter presents a study that addresses a critical but often overlooked step in boundary line analysis, Exploratory Data Analysis. It introduces a novel, objective method for evaluating whether a data set is suitable for boundary line modeling by examining the density of points in upper convex hull peels. This approach helps determine whether a plausible upper boundary is present and provides guidance for model selection and parameter initialization. The method is demonstrated using several agronomic and physiological data sets. In addition, simulated data is used to assess sample size requirements for detecting boundary structures in data when they exist. The findings highlight the importance of rigorous preliminary assessment before fitting boundary line models. This work has been published in ***Computers and Electronics in Agriculture Journal***. I contributed to this study by conceptualizing and developing the methodology, conducting the formal analysis, creating the visualizations, and drafting the original manuscript.

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

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.

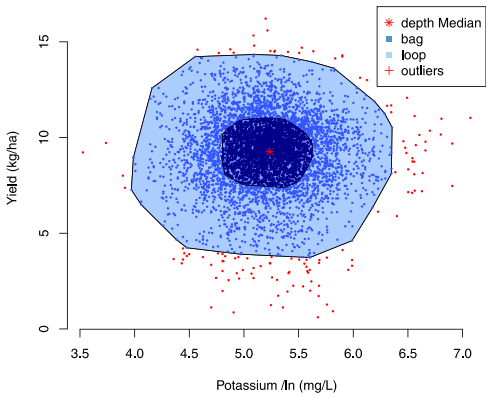


Fig. 1. An example of a bagplot components for a bivariate dataset of wheat yield against soil potassium concentration (Lark et al., 2020).

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.

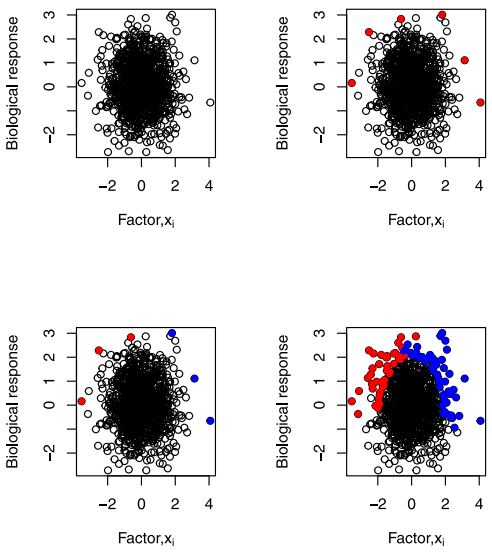


Fig. 2. The process of determining the successive peels in a dataset. (a) A scatterplot of a simulated bivariate (x, y) normal dataset. (b) The convex hull of the dataset determined as the first peel indicated with a red solid round dots . (c) The peel split into the left (red), and right (blue) sections. (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	$s\bar{d}_l$	p-value$_l$	sd_r	$s\bar{d}_r$	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, $s\bar{d}_l$ and $s\bar{d}_r$ 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

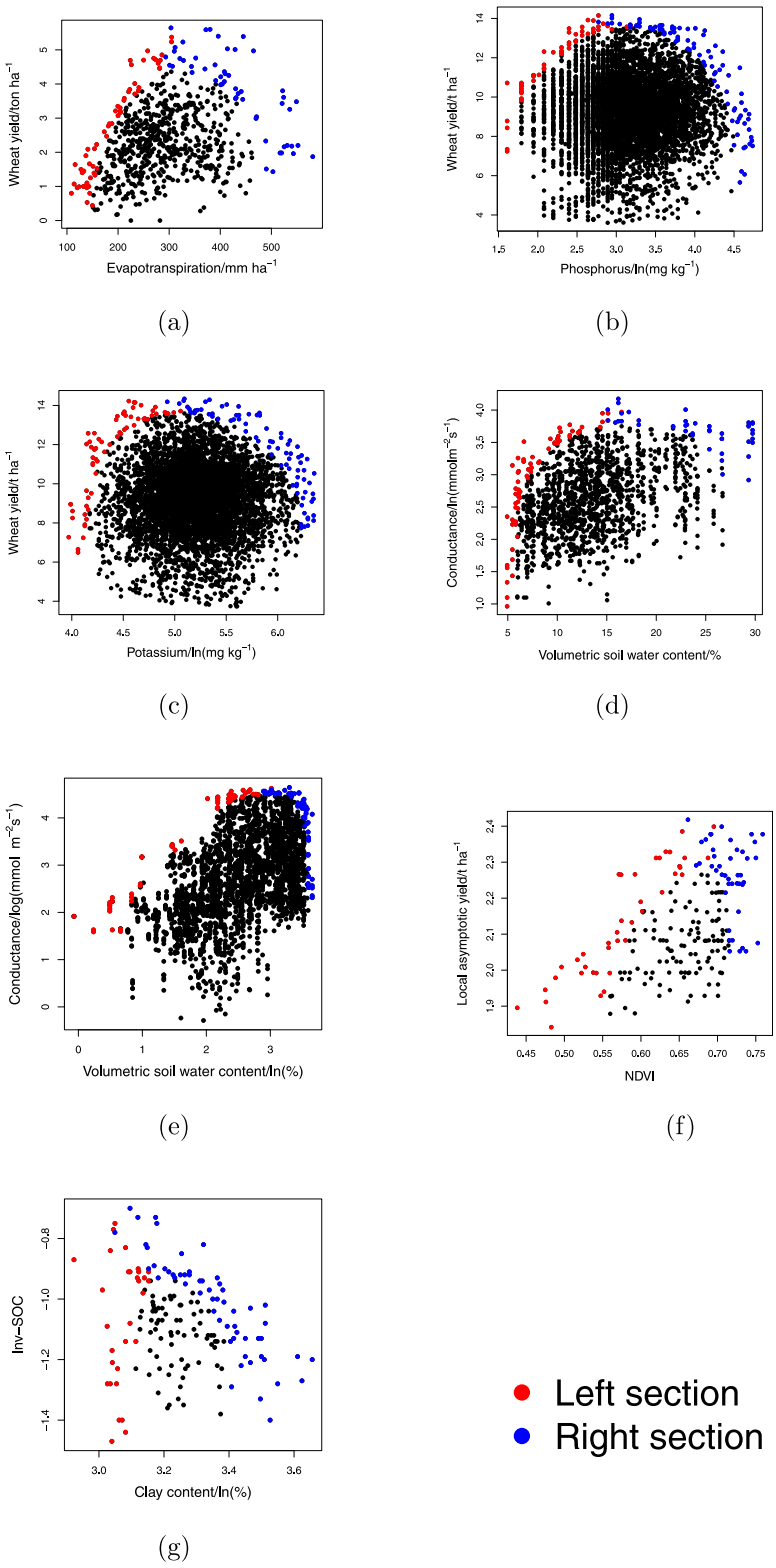


Fig. 3. Scatter plots of (a) wheat yield against evapotranspiration, (b) wheat yield against log phosphorus, (c) wheat yield against log potassium and, (d) beans log stomatal conductance against soil water content, (e) wheat log stomatal conductance against log volumetric water content, (f) local asymptotic 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

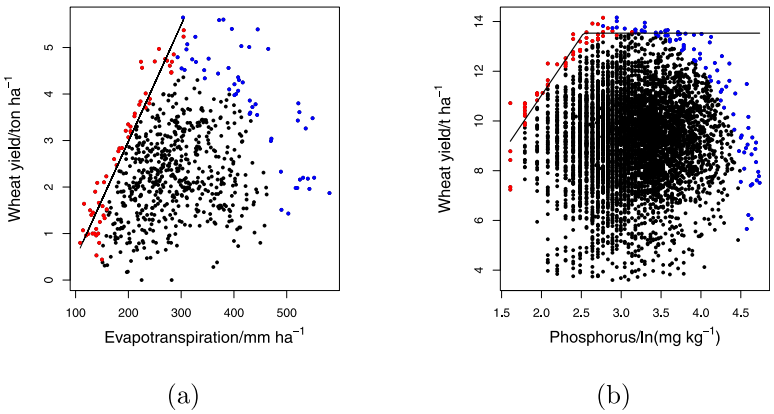


Fig. 4. Fitting appropriate forms of boundary model to datasets 1 (a) and 2 (b) based on peel density results.

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

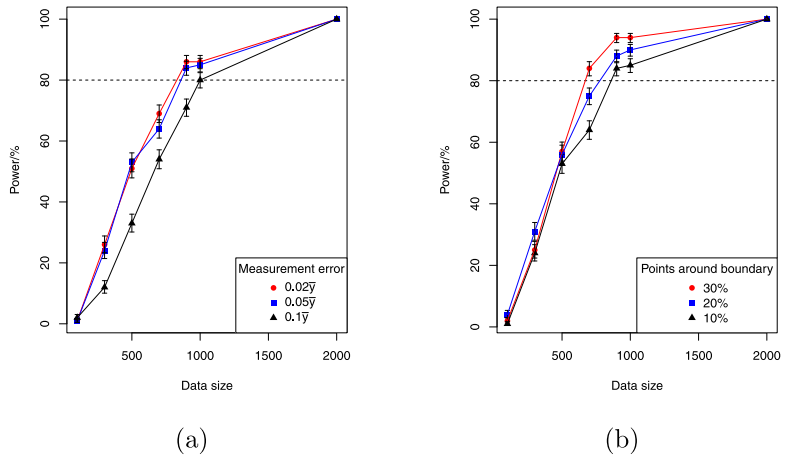


Fig. 5. Detection power with confidence interval for (a) varying measurement error and (b) varying concentration of points at the boundary of a dataset at 0.05 significance level.

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.

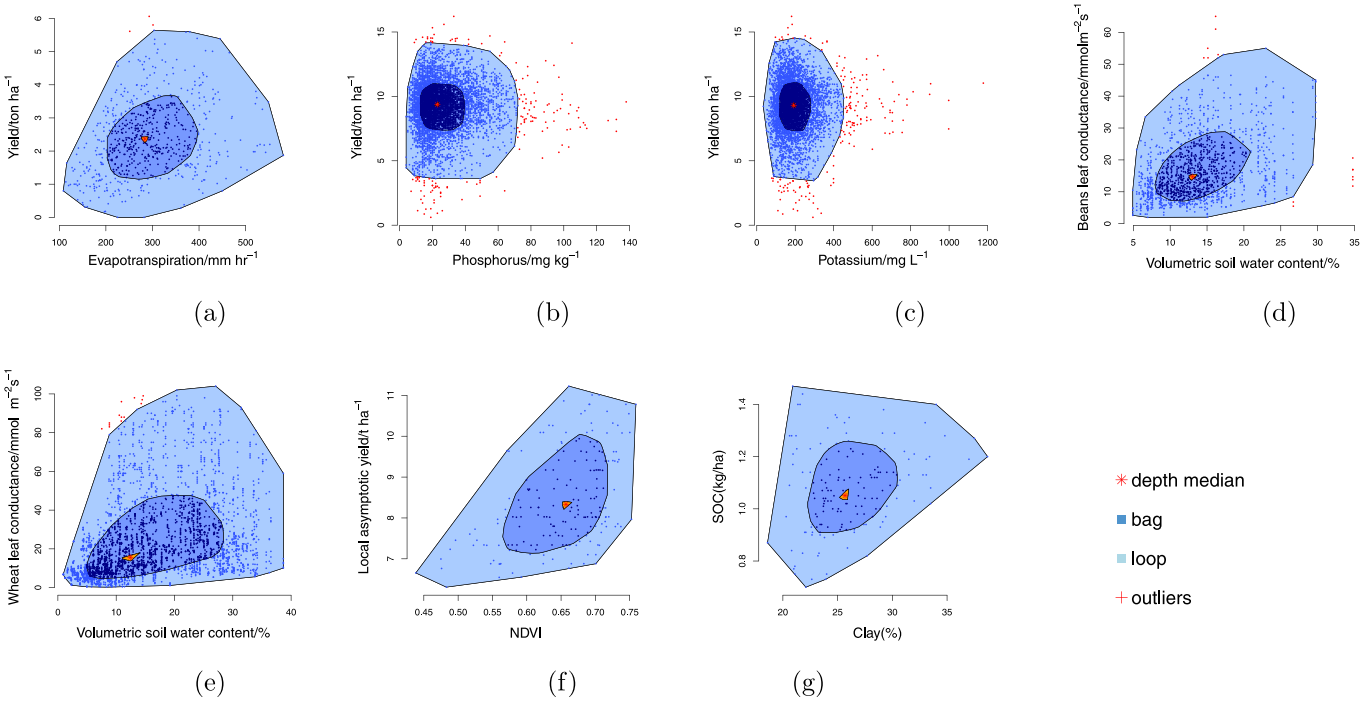


Fig. A.1. Bagplots of the datasets (a) wheat yield against evapotranspiration, (b) wheat yield against phosphorus, (c) wheat yield against potassium and, (d) beans stomatal conductance against soil water content, (e) wheat stomatal conductance against volumetric water content, (f) local asymptotic yield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

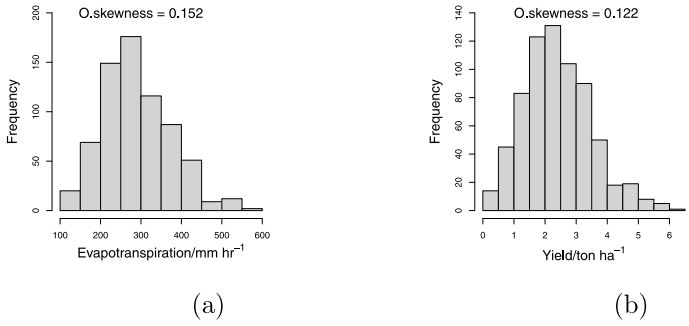


Fig. B.1. Histograms of the (a) evapotranspiration and (b) wheat yields from dataset 1.

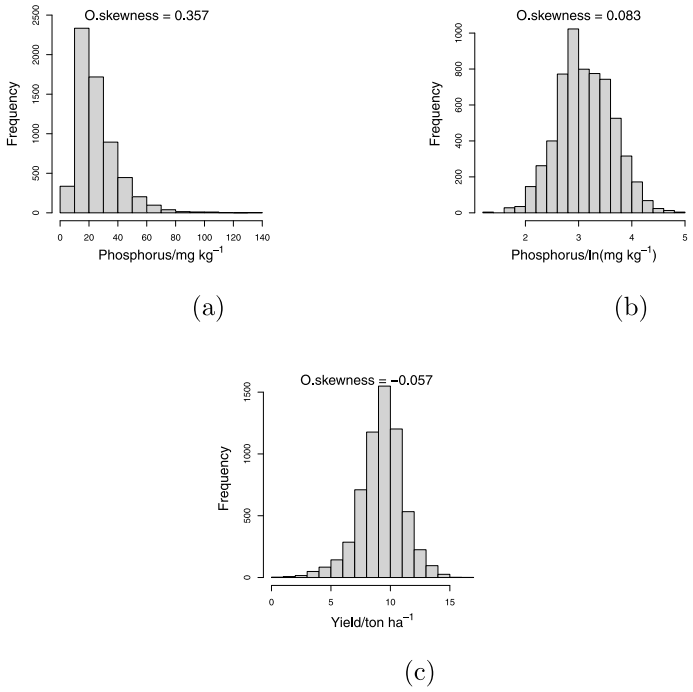


Fig. B.2. Histograms of the (a) soil phosphorus concentration, (b) log soil phosphorus concentration and (c) yield from dataset 2.

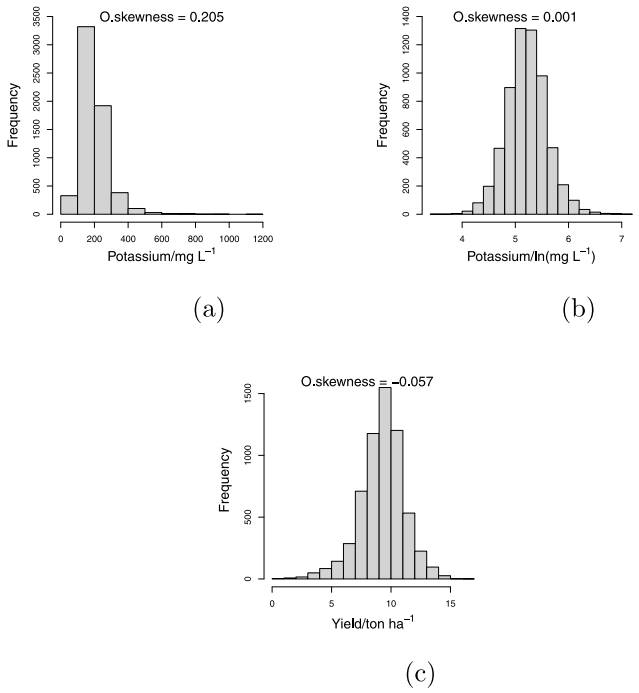


Fig. B.3. Histograms of the (a) soil potassium concentration, (b) log soil potassium concentration and (c) yield from dataset 3.

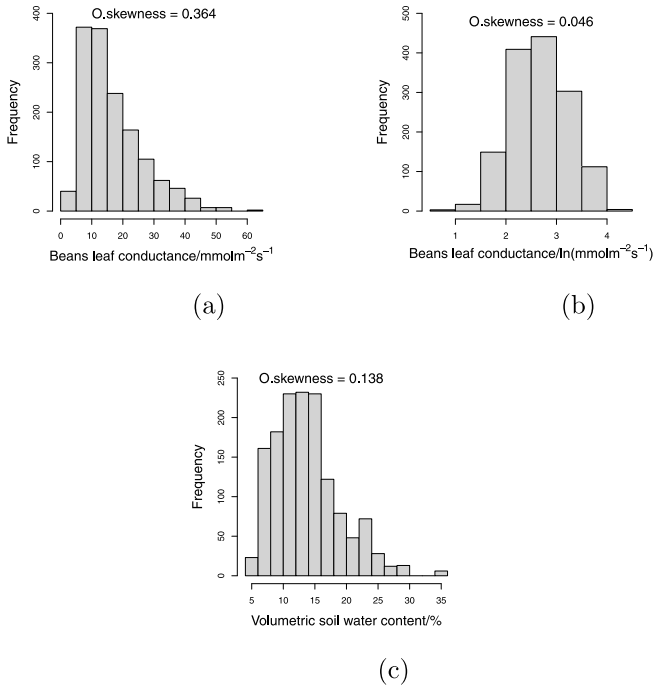


Fig. B.4. Histograms of the (a) bean leaf conductance, (b) natural log of bean leaf conductance and (c) soil volumetric moisture content from dataset 4.

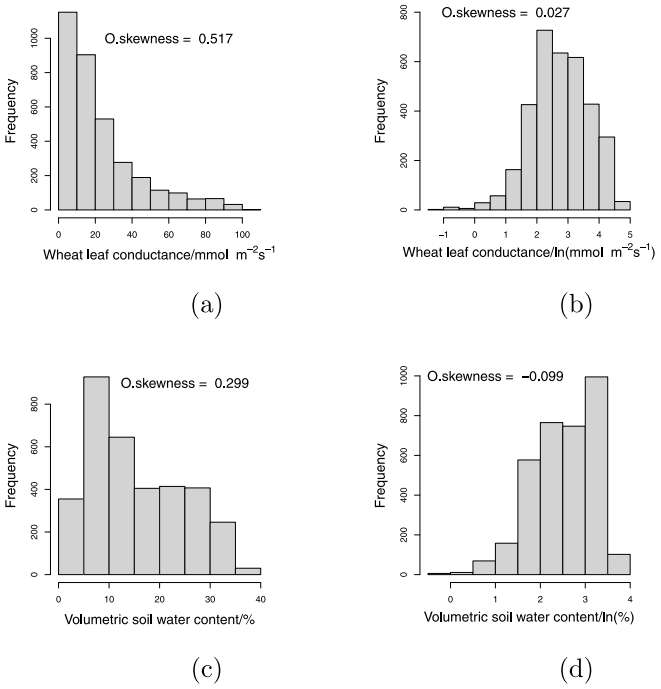


Fig. B.5. Histograms of the (a) Wheat leaf conductance, (b) natural log of wheat leaf conductance, (c) soil volumetric moisture content and (d) natural log of soil volumetric moisture content from dataset 5.

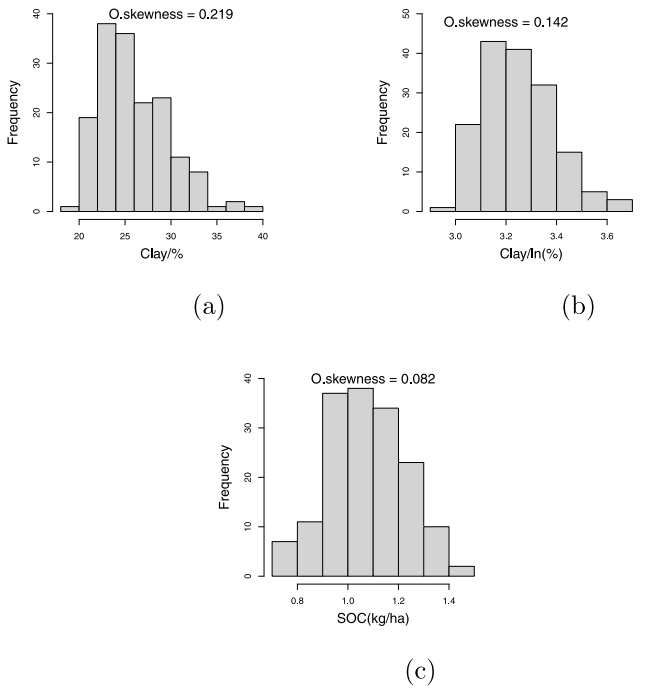


Fig. B.6. Histograms of the (a) soil clay content, (b) natural log of clay content and (c) soil organic content from dataset 6.

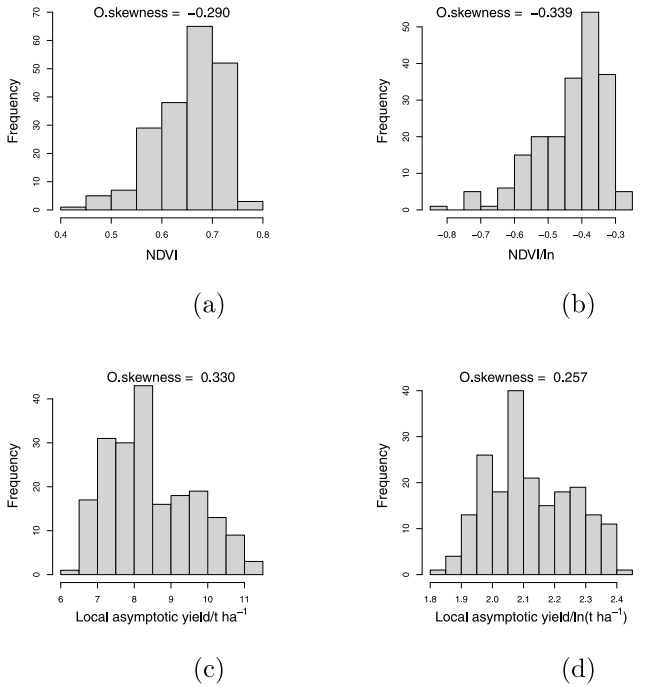


Fig. B.7. Histograms of the (a) NDVI, (b) natural log of NDVI, (c) local asymptotic yield and (d) natural log of local asymptotic yield from dataset 7.

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.

Chapter 4

Evaluating Boundary Line Fitting Approaches for Detecting Yield-Limiting Factors and Soil Nutrient Thresholds

The chapter has been submitted for publication in the European Journal of Agronomy as:

Miti. C, Milne. A, Giller. K, Lark. R (2025). Evaluating Boundary Line Fitting Approaches for Detecting Yield-Limiting Factors and Soil Nutrient Thresholds. *European Journal of Agronomy*, 170:127744. https://doi.org/10.1016/j.eja.2025.127744

4.1 Chapter 4 Overview

This chapter presents a comparative analysis of boundary line (BL) fitting methods for identifying yield-limiting factors and estimating critical soil nutrient concentrations. The study evaluates four BL fitting approaches including the binning, bolides, quantile regression, and the censored bivariate normal model (cbvn) across three datasets from England, East Africa, and Ethiopia. It assesses consistency in identifying the most-limiting factors using the Sprengel-Liebig Law of the Minimum and compares model-derived results with experimental findings from nutrient omission trials. The results highlight strengths and limitations of heuristic and statistical approaches, emphasizing the value of objective methods like cbvn for robust inference and uncertainty quantification. Additionally, feedback from stakeholder workshops provided insights into the usability and preferences for each method. This study has been published in ***European Journal of Agronomy***. I contributed to this work through conceptualization, study designing (including organizing the stakeholder workshops in Nairobi and Harare), conducting the formal analysis, creating visualizations, and writing the original manuscript draft.

European Journal of Agronomy 170 (2025) 127744

Contents lists available at ScienceDirect

European Journal of Agronomy

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

Evaluating boundary line fitting approaches for detecting yield-limiting factors and critical soil nutrient concentrations

C. Miti [a,b,*], A.E. Milne [b], K.E. Giller [c], 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

ARTICLE INFO

Dataset link: https://CRAN.R-project.org/package=BLA, https://data.cimmyt.org/dataset.xhtml?persistentId=hdl:11529/11015

Keywords:
Boundary lines
Most-limiting factor
Critical nutrient concentration

ABSTRACT

Closing the crop yield gap is critical to meeting rising global food demand driven by population growth. The boundary line (BL) methodology is widely used to assess yield gaps and identify its causes. However, the lack of a standard BL fitting method can lead to inconsistencies in outputs and recommendations. This study compared four BL fitting methods, binning, BOLIDES, quantile regression (QR), and the censored bivariate normal model (cbvn), in determining the most-limiting factor and critical values (x_{crit}) across three datasets from England (Dataset 1), East Africa (Dataset 2), and a nutrient omission-trial from Ethiopia (Dataset 3). The most-limiting factor was identified using the Law of the Minimum and experimentally via omission-trials. Agreement among BL fitting methods and between BL methodology and omission-trials was tested using Cohen/Fleiss κ-statistic. The consistency of x_{crit} from BL fitting methods was assessed using the 95% confidence interval (CI) of cbvn and compared to RB209 guidelines (Dataset 1 only). Additionally, stakeholder preferences/opinions on BL fitting methods were gathered via workshops in Nairobi and Harare. Results showed BL fitting methods generally identified the most-limiting factor consistently ($\kappa > 0.4$), but inconsistencies were observed for binning and QR methods. Experimentally-determined most-limiting factors were inconsistent with BL outputs ($\kappa < 0.2$). While most x_{crit} estimates fell within the cbvn CI, deviations occurred, especially in Dataset 2. BL fitting methods often underestimated x_{crit} compared to RB209 guidelines. Stakeholder exercise showed no evidence ($p = 0.56$) against the null hypothesis of uniform ranking of BL fitting methods. The study highlights that while BL fitting methods show general consistency, discrepancies with experimentally determined results exist. Despite consistent results, cbvn is recommended for critical nutrient estimation due to its uncertainty quantification, supporting probabilistic insights for agronomic decisions.

1. Introduction

Global increase in population coupled with the adverse effects of climate change and the reduction in agricultural land due to degradation threatens future global food security (Giller et al., 2021; Kopittke et al., 2019). To counter this, agricultural production must be increased. This can be achieved by agricultural intensification through closing crop yield gaps which can avoid the need for expansion of the land area under agriculture (Godfray et al., 2010; Foley et al., 2011). The process to close the yield gap should start with its quantification and the determination of its possible causes. Management options can be identified and policies adopted that favour the identified management options (FAO and DWFI, 2015). Various methods for bench-marking yield to determine the yield gaps have been described (FAO and DWFI, 2015). One method, which has been widely used in many agronomic studies, is the boundary line methodology (Sadras, 2020). This methodology allows for bench-marking yield especially, in cases where data are collected from non-experimental settings e.g. field surveys.

The boundary line methodology was initially proposed by Webb (1972) as a tool to model the most efficient response, y, to factor, x, when these are measured in situations where other causes of variations in y occur (Sadras, 2020). The boundary line model, therefore represents the maximum response of y, for any value of x. Webb's (1972) method involves, (i) creating a scatter plot of y against x, (ii) visually selecting the points at the upper edges assuming that they are the most efficient response and (iii) fitting a model to these selected points to represent the relationship of x and y. Over the years, this simple methodology has undergone various developments that include binning (Casanova et al., 1999), boundary line determination technique (BOLIDES) (Schnug et al., 1995), quantile regression

* 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.eja.2025.127744
Received 12 January 2025; Received in revised form 16 May 2025; Accepted 16 June 2025

methods, Makowski quantile regression (Makowski et al., 2007), the Bayesian segmented quantile regression (Andrade et al., 2023) and the censored bivariate normal model (cbvn) (Milne et al., 2006b). These methods can be classified into two groups, the heuristic methods (e.g. BOLIDES, Binning and quantile regression methods) which involve making subjective decisions (e.g. the quantile to regard as the boundary in quantile regression method) and the statistical methods which use strict statistical principles to fit boundary lines making them more objective and robust as they provide measures of uncertainty for the determined parameters (i.e. cbvn, Bayesian segmented quantile regression and Markowski quantile regression). For more information on the different boundary line fitting methods refer to Miti et al. (2024b).

The boundary line methodology is useful to identify appropriate agronomic practices to increase agricultural productivity. It has been applied in many studies to identify the most-limiting factors to production (Casanova et al., 1999; Fermont et al., 2009; Shatar and McBratney, 2004; Wang et al., 2015). Such information helps agronomists and agricultural land managers to prioritize which factors to address in order to improve productivity. The boundary line methodology has also been used for determination of critical nutrient concentrations (x_{crit}) to avoid limiting crop production (Andrade et al., 2023; Evanylo et al., 1987; Lark et al., 2020; Smith et al., 2024). This has economic and environmental benefits as application of excess fertilizer to soil which exceeds x_{crit} adds unnecessary financial costs of production and may also contaminate the environment (Schut and Giller, 2020). Reliable determination of the boundary line model parameters and their measures of uncertainty is therefore important if it is to be effectively used for yield gap analysis.

Despite the usefulness of the boundary line methodology in agronomy, there is still no standard agreed method for fitting a boundary line model to a dataset (Hajjarpoor et al., 2018; Smith et al., 2024; Miti et al., 2024b). The use of different boundary line fitting methods may lead to differences in the boundary line parameters and, therefore, different post-hoc interpretations and inconsistencies in agronomic recommendations. The use of statistical methods may be advantageous as they provide a more objective way of fitting the boundary line, making the procedure of boundary line fitting more consistent and also provide measures of uncertainty. However, they may be more complex to use which may make them difficult to implement for researchers with a limited statistical background (Harris and Smith, 2009). Statistical methods as opposed to heuristic methods may also require larger data sets to implement especially those that rely on the maximum likelihood approach. This means a lot of data is needed to establish the boundary with confidence. We are not aware of any studies that compare the results and interpretations for boundary line models fitted by different methods yet this type of comparative analysis is needed to give a better understanding of which methods should be recommended for future agronomic studies. This gap highlights the need for such comparative analysis, which our study addresses as a novel contribution.

The aim of this study is to compare the output of a boundary line analysis using different boundary line fitting methods for the purposes of (i) identification of the most-limiting factor, (ii) determination of critical soil nutrient concentration, and (iii) how these impact on agronomic interpretation and recommendations using dataset from three different studies. We also present results of a stakeholder consultation with agronomists and other researchers' opinions on the use of different boundary line methods as tools for yield gap analysis that were elicited during two hands-on workshops.

2. Methodology

2.1. Datasets used

Three datasets consisting of crop yields and potentially limiting factors were used for this study. Two of these datasets have previously been used in yield gap analysis studies using the boundary line methodology (Fermont et al., 2009; Lark et al., 2020).

The first dataset, Dataset 1, was assembled by AgSpace Agriculture Ltd and includes measures of wheat yield as well as soil variables pH, potassium (K), phosphorus (P), and magnesium (Mg), all of which were taken in various management units across England. For more details see Lark et al. (2020). AgSpace Agriculture Ltd uses pre-identified management zones inside each field to guide soil sampling for its clients. Each management unit served as the foundation for the sampling zone and was defined by skilled soil scientists utilizing a free survey. In each sampling zone, 24 soil cores were taken to a depth of 15 cm, and then pooled to create a bulk sample. A sub-sample was taken from the bulk sample for laboratory analyses for P, Mg, K and pH. The Olsen's method was used to extract P while K and Mg were extracted using 1M ammonium nitrate. Soil pH was measured in 1:2.5 soil to water suspension with a pH meter. 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. In our study, we applied the boundary line methodology to wheat yield and the variables P, K, Mg and pH.

The second dataset, Dataset 2, was compiled by Fermont et al. (2009) from farm surveys and agronomic on-farm and research station trials in a study on closing the cassava yield gap for smallholder farms in East Africa. Data on main production constraints, socio-economic settings, farm management, and cassava crop management was collected in on-farm surveys. In each field at a farm, composite soil samples (depth of $0-20$ cm) were taken, oven-dried, sieved through a 2-mm sieve, and analysed for pH, available P, exchangeable K, Ca, Mg, total N, soil organic carbon (SOC) and soil texture. Daily precipitation data were recorded using rainfall gauges at all sites. Research technicians scored overall weed management as well as disease incidence (yes/no) and severity. The boundary line methodology was applied to this dataset by Fermont et al. (2009) to identify the most-limiting factor in each field. In our study, we applied the boundary line methodology to cassava yield and the soil pH, available P, exchangeable K, Ca, Mg, total N and SOC.

The third dataset, Dataset 3, was compiled by CIMMYT for wheat nutrient omission trials conducted in two zones (West Showa and Jimma) in Ethiopia in 2015 and 2016 (Craufurd, 2017). The trials comprised six nutrient management treatments, namely Control (zero fertilizer), PK (N omitted), NK (P omitted), NP (K omitted), NPK, and NPK+Ca+Mg+Zn+B laid out in a randomized complete design replicated across individual farmers' fields. Each treatment plot measured 8 meters by 8 m, with maize planted at a spacing of 75 cm (inter-row spacing) and 25 cm (intra-row spacing). The nutrients N, P and K were applied at rates of 120, 40 and 40 kg/ha respectively using urea (46%), triple super phosphate (P_2O_5) and muriate of potash(K_2O) as sources. Nitrogen was applied in three splits, the first as a basal application, the second as a topdressing 21 days after emergence, and the third was applied as a topdressing approximately 42 days after emergence. All other nutrients (P and K) were applied as basal at the time of planting. Apart from fertilizer application, all plots received the same management practices. The plots were weeded manually twice during the cropping season. First at 2 weeks after planting and the second at 4 weeks after planting. Key soil properties SOC, total N, available P, exchangeable K, exchangeable Ca, exchangeable Mg, pH, soil texture, exchangeable acidity (H + Al) and Micro-nutrients (Zn, Fe, Cu, Mn and B) were measured in each field on soil samples to depth of 20 cm before the trial. SOC was determined using the modified Walkley and Black method while the micro-Kjeldahl digestion was used to determine total N. The Mehlich-3 extraction procedure preceding inductively coupled plasma optical emission spectroscopy was used to determine available P, K, Ca, Mg and micro-nutrients (Zn, Fe, Cu, Mn and B). Exchangeable acidity (H + Al) was determined by extracting soil with 1N potassium chloride and titration of the supernatant with 0.5M sodium hydroxide. At the end of the trial, biomass and grain yields were collected from each plot. We used the data from the control plot to fit the boundary lines to determine the most-limiting factor in the absence of fertilizer application. A total of 148 data points were used. In this study, boundary line models were fitted to SOC, N, P, K, Ca, Mg and pH data.

2.2. Fitting boundary line to data

2.2.1. Initial exploratory analysis

An exploratory analysis was conducted on all the factors used to fit the boundary line models in the datasets. Since the censored bivariate normal model works on the assumption that data are from a bivariate normal distribution with a censoring boundary frontier, exploratory analysis was conducted using histograms and summary statistics to check for the assumption of the variable being from a normal distribution possibly showing censoring of the yield variable. The octile skewness, a robust measure of skewness, was used (Brys et al., 2008). Variables with an octile skewness between −0.2 to 0.2 were assumed to be from a normal distribution. Otherwise they were log-transformed to meet the assumption. Because boundary line analysis is sensitive to outliers, we identified and removed bivariate outliers—data points that deviate markedly from the central distribution in two dimensions. To detect these, we used the bagplot, a bivariate extension of the univariate boxplot based on halfspace depth (Tukey, 1975). The central 50% of the data, referred to as the 'bag', was computed following the procedure of Rousseeuw et al. (1999), and the outer fence was defined by expanding this bag by a factor of 3. Observations falling outside the outer fence were classified as bivariate outliers and excluded from further analysis. Prior evidence for a limiting yield boundary in the response to different factors was obtained using the peel cluster approach described by Miti et al. (2024c).

2.2.2. Proposed models forms

Piece-wise broken-stick boundary line models were fitted to scatter plots for response, y as a function of factor x for the variables in the three datasets, as have been used in previous studies (Andrade et al., 2023; Lark et al., 2020). This enabled the determination of critical values (x_{crit}) in the independent variable (i.e. critical nutrient concentration) from the inflection points of the models. Two model forms, linear-plateau and trapezium boundary line models were used depending on the shape of the upper bound of a data cloud. The linear-plateau is a two-piece model that consists of a linear component that represents the functional change in response, y, with incremental changes in the independent factor, x, and a plateau component that represents the section of no change in y with change in x. It is expressed as

$$y = \min(\beta_0,\ \beta_1 + \beta_2 x), \tag{1}$$

where β_0 is the plateau response, β_1 is the y-intercept and β_2 is the slope. The trapezium model is a three piece model consisting of two linear components (with positive and negative slopes respectively) separated by a plateau component. It can be expressed as

$$y = \min(\beta_0,\ \beta_1 + \beta_2 x,\ \beta_3 + \beta_4 x), \tag{2}$$

where β_0 is the plateau response, β_1 and β_3 are the y-intercepts for the two linear components, and β_2 and β_4 are the slopes with positive and negative values respectively.

2.2.3. Fitting methods

Four commonly-used boundary line fitting methods were used to fit boundary lines to the data. These were the binning methodology (Casanova et al., 1999), BOLIDES algorithm (Schnug et al., 1995), quantile regression (Baudron et al., 2019) and the cbvn (Milne et al., 2006a).

For the binning methodology, a data cloud of the response variable, y, against a factor of interest, x, was divided into 10 sections in the x-axis as done by Casanova et al. (1999). In each section, a boundary point was selected as the point that corresponds to a given percentile of y. In this study we used the 95th and 99th percentile (τ). These are the most commonly used τ values in the literature for fitting boundary lines to data, although we note that they are somewhat arbitrary. The boundary line models were then fitted to the selected boundary points by the method of least squares.

The BOLIDES algorithm operates in several steps. First, the data relating factor x and response y were arranged in ascending order with respect to x. Three benchmark x values were defined as x_{min} for the minimum value of factor x, x_{max} for the maximum value of factor x and x_{ymax} for the value of factor x that corresponds to the largest response in the dataset. The boundary points were then selected as follows. The first boundary point was selected as the point at x_{min} with the largest y. The next point was the data point at the next x value that had the largest response greater than the previous boundary point. Subsequent boundary points were selected in this way until x_{ymax} was reached. This process was then repeated starting from x_{max}, this time moving in the opposite direction. Agronomically plausible boundary line models were then fitted to the selected boundary points by the least squares method.

The quantile regression method was implemented by fitting a boundary line model of the response y as a function of x to the data conditional of a quantile, τ, which can range from $0 - 1$ (Cade et al., 1999). Parameters of the boundary model were determined by minimizing the weighted sum difference between the modelled response and actual response (Koenker and D'Orey, 1987). In this procedure, data points that lay above the modelled response are given more weight (τ) than those that fall below it ($1 - \tau$). Two quantile values, $\tau = 0.95$ and $\tau = 0.99$, commonly used in literature to fit boundary lines were considered in this study.

For the cbvn, the boundary line was fitted using the maximum likelihood approach following the procedure described by Lark and Milne (2016). In brief, this method assumes that the observed data is from a bivariate normal distribution of x and y, which has an upper censor (boundary) which y cannot exceed. Fitting the boundary line in this case involves finding the maximum likelihood of a censored bivariate normal distribution of the data given its distribution properties (means of x and y, $\boldsymbol{\mu}$ and variance–covariance matrix, $\mathbf{C}$), the parameters of the censor/boundary line, $\boldsymbol{\beta}$ and a parameters describing the measurement error in y (σ_{me}). The likelihood of the distribution is given as the product of the likelihoods for the individual data points. As the product of a large number of small likelihoods can easily underflow the numerical precision of the computer, the log of the likelihood values was taken and the sum of the log likelihoods was computed instead. Direct estimates of σ_{me}, a fixed parameter in the cbvn, were not available and were therefore, estimated using the σ_{me} likelihood profile (Royall, 2017; Lark and Milne, 2016). This was done by fixing σ_{me} at each of a set of values in turn, and remaining parameters were estimated conditional on σ_{me} by maximum likelihood. The maximized likelihoods for the sequence of values constitute a likelihood profile. The value of σ_{me} where the profile is maximized was selected as the estimate of σ_{me}.

To determine the parameters of the boundary line, an objective negative log-likelihood function was written in `R` software (version 4.4.0) (R Core Team, 2022) and the `optim()` function was used to find the parameters that minimized the objective function using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization methods. There is a tendency for optimization algorithms to settle at a local rather than global minima. To avoid this, several starting values were used for the optimization and the output with the lowest negative log-likelihood value was chosen. In addition, the uncertainty (standard error) of the parameters was determined using the Hessian matrix, $\mathbf{H}$ (which is part of the output of the `optim()` function), following the procedure described by Dobson and Barnett (2018). This is done by taking the inverse of $\mathbf{H}$ and then taking the square root of the corresponding terms on the main diagonal.

Evidence for the validity of the cbvn compared to a bivariate normal model was assessed by computing Akaike's information criteria (AIC) for the censored and uncensored bivariate normal distribution models fitted to the data (Lark and Milne, 2016). The AIC values for a model was calculated as

$$\Lambda = 2p + 2\ell, \tag{3}$$

where Λ is the criterion, p is the number of model parameters and ℓ is the maximized log-likelihood value. This information criterion compares the models on their likelihood, with a penalty for the greater complexity of the cbvn model. In any comparison the model with smallest AIC is selected. Although the AIC is not a formal significance test, if one selects the model with smaller AIC this minimizes the expected information loss through the selection decision (Verbeke and Molenberghs, 2000).

In general we selected the cbvn model if its value of AIC, Λ_{bl}, was smaller than that of the uncensored bivariate normal model, Λ_{bvn}. Further information on the strength of this evidence for the model was obtained by computing the Akaike weight for each, as done by Lark et al. (2020). If the value of Λ for the i^{th} model, Λ_i, in a set of m exceeds the minimum value of Λ_j over the set of models considered by Δ_i then one may compute the Akaike weight for that model as

$$w_i = \frac{\exp\{-\Delta_i/2\}}{\sum_{j=1}^{m}\exp\{-\Delta_j/2\}}. \quad (4)$$

The value w_i can be interpreted as the probability that the model is the best one for the variable in the sense that the Kullback–Leibler divergence is minimal for the i^{th}, over the set of m models considered (Burnham and Anderson, 2004).

2.2.4. Determination of the most-limiting factor

For each point in the three datasets, the most-limiting factor was identified according to Sprengel-Liebig Law of the Minimum (van der Ploeg et al., 1999) as done by Fermont et al. (2009) using the equation:

$$y = \min(f_1(x_1), f_2(x_2), \ldots, f_n(x_n)), \quad (5)$$

where y is the response variable and x_i are the potential limiting factors on y according to the boundary line function $f_i(x_i)$, where $i = 1, 2, \ldots, n$. The most-limiting factor is identified as i if $y = f_i(x_i)$. However, if a point has an x value greater than the critical value for all factors and has a yield gap, the cause of the yield gap was defined as "Unknown" because all the studied factors predict that yield should be maximum. The most-limiting factor was modelled using each of the four boundary line fitting methods.

For Dataset 3, this method was applied to interpret the boundary models fitted to the yield data from control plots only. An experimental determination of the most-limiting factor was also done considering the additional plots from the nutrient omission trial. The most-limiting factor was identified for each experimental farm in the omission trials by examining the yields for all non-control plots (PK, PN, NK, NPK). The most limiting factor was identified by finding, among the plots with all, or all but one, nutrient applied, which had the smallest yield. This is the nutrient which, on addition, has the largest effect on yield. In cases where the smallest yield was observed in the treatment that received all the nutrients (NPK), the limiting factor was classified as "Unknown". Additionally, if the difference between the treatment with the smallest yield and NPK is less than the critical difference at 95% confidence, we concluded that the NPK yield was not significantly large and hence classified the limiting factor as "Unknown". The critical difference with 95% confidence was determined as

$$\Delta_{\text{crit}} = 1.65 \times \sqrt{2\sigma^2_{\text{me}}}, \quad (6)$$

where Δ_{crit} is the critical difference at 95% confidence and σ^2_{me} is the variance of measurement error (determined using the likelihood profile described in Section 2.2.3). The variance of difference between two measurements is given as $2\sigma^2_{\text{me}}$ and so the standard error of difference is equal to its square root. The 1.65 is a one-tailed t-statistic value at 95% confidence. For example, given that the yields for the treatments PK(−N), NP(−K) NK(−P) and NPK are 2, 6, 5.21 and 6.5 respectively, and the Δ_{crit} is equal to 3. The most-limiting factor is identified as N because it has the smallest yield and its difference with NPK is larger than Δ_{crit}. However, if the yield for the treatment PK(−N) was 5, its difference with NPK treatment was less than Δ_{crit} (no significant difference). In this case, the most-limiting factor was classified as unknown.

Table 1
Joint proportions of rating by two boundary line fitting methods with j limiting factor categories.

		Method 1				
		factor 1	factor 2	..	factor j	**Total**
Method 2	factor 1	π_{11}	π_{12}	..	π_{1j}	$\pi_{1.}$
	factor 2	π_{21}	π_{22}	..	π_{2j}	$\pi_{2.}$
	:	:	:		:	:
	factor j	π_{j1}	π_{j2}	..	π_{jj}	$\pi_{j.}$
	Total	$\pi_{.1}$	$\pi_{.2}$	..	$\pi_{.j}$	1

2.3. Statistical analysis

2.3.1. Comparing the most-limiting factors predicted by different boundary line fitting methods

The Cohen's and Fleiss kappa (κ) statistics were used to check the agreement among the different boundary line fitting methods in the identification of the most-limiting factor. The κ statistic is a measure of inter-rater-reliability of classification methods (Fleiss, 1971). The Cohen's κ is used when two classifications are compared while the Fleiss κ is used when more than two classifications are compared. These statistics are particularly useful when dealing with categorical data or classification problems, where models are assigning items to categories or labels. It assesses whether the agreement between the models is better than what would be expected by chance alone. The κ statistic has been widely used in medicine and psychology studies, and it has also been applied in ecological studies (Monserud and Leemans, 1992; Tarkesh and Jetschke, 2012; Tang et al., 2009). It works on three basic assumptions that (a) the subjects being rated are independent (knowing the class of subject x_1 does not give any information about the class of subject x_i, where $i = 2, 3, \ldots, n$), (b) the categories of ratings are independent, mutually exclusive and collectively exhaustive, and (c) two raters operate independently. The formula for κ is given as:

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e}, \quad (7)$$

where π_o is the observed agreement (the proportion of times the models actually agree) and π_e is the expected agreement by chance (Fleiss et al., 2003). For a comparison of two boundary line fitting methods (though this can be extended to comparison of more than two methods) with j number of potential limiting factors modelled over n number of points, a contingency table of proportions of agreement is generated as shown in Table 1. Note that π is the proportion and not the count value.

As described by Fleiss et al. (2003), the value of π_o is calculated as:

$$\pi_o = \sum_{i=1}^{j}\pi_{ii}, \quad (8)$$

while π_e is calculated as:

$$\pi_e = \sum_{i=1}^{j}\pi_{i.}\pi_{.i}. \quad (9)$$

The κ coefficient values range from −1 to 1. A κ of 1 shows that there is perfect agreement while a κ equal to 0 indicates that the agreement is equivalent to chance. A κ of less than 0 indicates that agreement is worse than chance. This negative value suggests that the raters are not only disagreeing but are systematically disagreeing more often than what is expected by random chance alone. Generally, κ values greater than 0.75 are considered to represent excellent agreement beyond chance while those below 0.40 are considered poor agreement beyond chance. Values of κ that are between 0.40 and 0.75 are considered to represent fair to good agreement beyond chance (Monserud and Leemans, 1992).

The z-statistic is used to test null hypothesis (H_0) that the agreement is no better than would be obtained by chance agreement ($\kappa = 0$)

against the alternative hypothesis (H_a) that the agreement is different from chance agreement ($\kappa \neq 0$). The z-statistic is computed as:

$$z = \frac{\kappa - 0}{\text{se}(\kappa)}, \tag{10}$$

where se(κ) is the standard error of κ calculated based on the formula proposed by Fleiss et al. (1969):

$$\text{se}(\kappa) = \frac{1}{(1-\pi_e)\sqrt{n}} \sqrt{\pi_e + \pi_e^2 - \sum_{i=1}^{j} \pi_{i.}\pi_{.i}(\pi_{i.} + \pi_{.i})}. \tag{11}$$

The null hypothesis was rejected if the probability of getting the calculated z-value or larger was small. This process was done using the functions `kappa2()` and `kappam.Fleiss()` from the `irr` package version 0.84.1 (Gamer et al., 2012) to compare two and more than two models respectively.

The κ statistic was also used to compare the modelled most-limiting factor using boundary lines and that obtained experimentally using omission trials for Dataset 3. The assumption that outcome categories of the raters are collectively exhaustive for use of κ-statistic was not met as the omission trial had four possible outcomes (N, P, K and Unknown) while the boundary line models had eight possible outcomes (Ca, K, Mg, N, SOC, P, pH and Unknown). However, if the classifications from the boundary line methodology and omission trials are the same, we expect points that identified pH, Ca, Mg and SOC as the most-limiting factor using the boundary line models will correspond to the classification "Unknown" for the omission trial. We therefore converted the classifications pH, Ca, Mg and SOC for the boundary line methodology to Unknown to satisfy the assumption of raters being collectively exhaustive.

2.3.2. Comparing critical nutrient concentration obtained using various boundary line techniques

The critical nutrient concentrations for variables in datasets 1, 2 and 3 were computed using the boundary line fitting methods described above and the outcomes compared. The critical nutrient concentration was determined as the inflection point of the boundary line model where an increase in soil nutrient concentration does not result in an increase in yield. This was calculated using parameters obtained from the model in Eqs. (1) and (2) as:

$$x_{\text{crit}} = \frac{\beta_0 - \beta_1}{\beta_2}. \tag{12}$$

The estimates of x_{crit} by heuristic methods were compared to the estimate obtained by the cbvn method to check if they are consistent. This was done in `R` (version 4.4.0) by computing the uncertainty around the x_{crit} determined using cbvn and checking whether the x_{crit} obtained using heuristic methods were contained in this uncertainty range. To determine the uncertainty around x_{crit} from the cbvn, 10000 combinations of the boundary line parameters were generated using the `mvrnorm()` function from the `MASS` package version 7.3-65 (Ripley et al., 2013) with mean vector equal to the censored bivariate normal parameter estimates and the estimated covariance matrix, $\boldsymbol{\Sigma}$. The value of x_{crit} was computed from each simulated set of boundary line parameters. As the values of x_{crit} were not symmetrically distributed, a 95% highest density interval (HDI) (Barry, 2011) was determined around x_{crit} as a measure of uncertainty using the `hdi()` function from the `HDInterval` package version 0.2.4 (Meredith and Kruschke, 2020). The HDI is the interval which contains the required mass such that all points within the interval have a higher probability density than points outside the interval. The HDI is particularly useful for summarizing uncertainty of a variable with an asymmetric distribution. A 95% HDI, means that there is a 95% probability that the true value of the parameter lies within that interval. Once the 95% HDI of x_{crit} was determined, we checked whether the x_{crit} obtained using the heuristic methods were contained within this range.

For Dataset 1, the estimated x_{crit} values were further compared to the standard index concentrations for soil nutrients proposed by RB209 (AHDB, 2023) for fertilizer advice in the UK. The RB209 guidelines classifies laboratory soil analysis result ranges for P, K and Mg into soil indices for each nutrient according to cereal yield response in field trials (AHDB, 2023). For P, K and Mg, yield response is expected in index 0 and 1 with application of P, K and Mg fertilizer. In index 2, 2- and 2 for P, K and Mg respectively, yield response to addition of P, K and Mg is not expected as this is the optimum range. It is therefore, expected that the x_{crit} values determined from the boundary line methodology for each nutrient to fall in these respective indices. At index 3 and above, application of P, K and Mg is not required.

2.4. Stakeholder feedback exercises

It is important that methods for yield gap analysis are accessible to scientists who may wish to use them. For this reason workshops were organized at CGIAR centres in Kenya and Zimbabwe. Participants with various specialities and levels of experience in statistical analysis were drawn from various institutions including International Institute of Tropical Agriculture (IITA), International Centre of Insect Physiology and Ecology (ICIPE), Alliance of Bioversity International and CIAT, Kenya Agricultural and Livestock Research (KALRO), University of Nairobi, French Agricultural Research Centre for International Development (CIRAD) and International Maize and Wheat Improvement Centre (CIMMYT). These were recruited by CGIAR research leads in Harare (CIMMYT) and Nairobi (IITA) who contacted potential users of the boundary line methodology for yield gap analysis. The criteria used to identify potential participants was that they work in research in agricultural crop production. A total of 32 participants took part in the workshops, 14 in Kenya and 18 in Zimbabwe. A comprehensive yield gap analysis was demonstrated to the participants using the four boundary line fitting methods, binning, BOLIDES, quantile regression and cbvn, in what was then a prototype of the BLA library (Miti et al., 2024a). `R` scripts for each of the boundary line fitting methods were provided and analysis was demonstrated on Dataset 1. Participants were then given Dataset 2 to carry out the yield gap analysis independently using the different boundary line fitting methods.

An anonymous structured questionnaire was given to participants at the end of the session to elicit their views on the various boundary line fitting methods and yield gap analysis in general. Each questionnaire had a consent form attached for participants to indicate informed consent to participate in the study. A positive ethical option (SBREC202324010FEO - SB202223/35 PhD - Miti (Lark)) was provided by the School of Biosciences Research Ethics Committee (University of Nottingham) prior to this study taking place. The questionnaire was composed of two sections. The first section consisted of questions intended to collect information on participants' research areas, their interaction with farmers and level of statistical experience. The second part contained questions to collect information on their past use of the boundary line methodology, their preferred methods (through ranking) and their views on boundary line methodology for yield gap analysis in general. The complete structure of the questionnaire is provided in Appendix D.

The question in section two required participants to rank the boundary line fitting methods presented in the workshop according to which they found easier to use from the easiest (rank 1) to the least easy (rank 4). The dataset was tested for uniformity in ranking to see if there was influence of location, statistical experience and past use of boundary line methods. When a ranking dataset is uniform, it means that all possible rankings have the same probability of being observed. To test the uniformity of ranking in the dataset, a hypothesis test was conducted on the mean rank according to the procedure of Marden (1996). Under the null hypothesis of uniform ranking for k items, the expected value of the mean rank is given as:

$$\mu_{\text{rank}} = (k+1)/2, \tag{13}$$

Table 2
Parameters of fitted boundary lines of the potential limiting factors for Dataset 1.

Method	Factor	Model	β_1	β_2	β_0	β_3	β_4	Λ_{bl}	Λ_{bvn}	ω_{bv}	ω_{bl}
Binning ($\tau = 0.95$)	P	trapezium	7.31	1.89	12.14	19.18	−1.82	–	–	–	–
Binning ($\tau = 0.95$)	Mg	trapezium	5.52	5.31	12.13	26.48	−2.67	–	–	–	–
Binning ($\tau = 0.95$)	pH	linear-plateau	18.88	3.53	11.97	–	–	–	–	–	–
Binning ($\tau = 0.95$)	K	linear-plateau	–	–	12.24	15.60	−0.67	–	–	–	–
Binning ($\tau = 0.99$)	P	trapezium	4.74	3.38	13.34	22.16	−2.32	–	–	–	–
Binning ($\tau = 0.99$)	Mg	trapezium	−15.17	8.66	13.32	34.46	−3.95	–	–	–	–
Binning ($\tau = 0.99$)	pH	linear-plateau	24.05	5.44	13.34	–	–	–	–	–	–
Binning ($\tau = 0.99$)	K	trapezium	−1.52	3.42	13.61	23.18	−1.77	–	–	–	–
BOLIDES	P	trapezium	4.76	3.46	13.57	108.35	−21.26	–	–	–	–
BOLIDES	Mg	trapezium	−18.24	9.34	13.94	40.94	−5.14	–	–	–	–
BOLIDES	pH	linear-plateau	35.01	10.21	13.97	–	–	–	–	–	–
BOLIDES	K	trapezium	−76.01	21.16	13.95	42.86	−5.04	–	–	–	–
Q.reg ($\tau = 0.99$)	P	trapezium	7.82	2.03	13.37	27.41	−3.46	–	–	–	–
Q.reg ($\tau = 0.99$)	Mg	trapezium	−15.34	8.50	13.34	37.37	−4.47	–	–	–	–
Q.reg ($\tau = 0.99$)	pH	linear-plateau	17.59	2.17	13.36	–	–	–	–	–	–
Q.reg ($\tau = 0.99$)	K	trapezium	2.04	2.53	13.42	41.31	−4.76	–	–	–	–
Q.reg ($\tau = 0.95$)	P	trapezium	7.83	1.65	12.14	15.26	−0.85	–	–	–	–
Q.reg ($\tau = 0.95$)	Mg	trapezium	−8.37	6.07	12.14	26.11	−2.58	–	–	–	–
Q.reg ($\tau = 0.95$)	pH	linear-plateau	13.74	1.12	12.13	–	–	–	–	–	–
Q.reg ($\tau = 0.95$)	K	trapezium	−6.62	5.80	12.10	44.10	−5.23	–	–	–	–
cbvn	P	trapezium	5.75 (0.88)	2.82 (0.37)	13.43 (0.09)	82.57 (12.66)	−15.77 (2.71)	**32376.46**	32 429.55	1.0	0.0
cbvn	Mg	trapezium	−15.33 (2.61)	8.13 (0.84)	13.20 (0.19)	40.62 (1.90)	−5.15 (0.31)	**33 145.73**	33 189.39	1.0	0.0
cbvn	pH	linear-plateau	26.93 (1.88)	6.98 (0.81)	13.61 (0.10)	–	–	**41 073.00**	41 093.24	0.92	0.08
cbvn	K	trapezium	−21.88 (0.04)	8.12 (0.16)	13.35 (0.19)	46.23 (7.03)	−5.67 (1.16)	**28 464.52**	28 476.13	0.99	0.01

cbvn, censored bivariate normal model; Q.reg, quantile regression model; Λ_{bl} and Λ_{bvn}, AIC values for the censored and uncensored bivariate normal models; ω, Aikaike weight for the fitted model.
Bold values of Λ_{bl} and Λ_{bvn} indicate the smaller of the two values.

and the evidence against this null hypothesis is measured using the test statistic:

$$\frac{12n}{k(k+1)} \sum_{i=1}^{k} \{m_i - \mu_{\text{rank}}\}^2, \tag{14}$$

where m_i is the mean rank of the i^{th} item, and n is the number of rankings. Under the null hypothesis, this statistic is distributed as χ^2 with $n - 1$ degrees of freedom (Lee and Yu, 2013). Note that whilst participants ranked the methods 1 (favoured) to 4 (least favoured), mean rank was evaluated on reversed ranking i.e. the method which was ranked as the easiest to use was given a weight of 4 while the second, third and fourth easiest were given weights of 3, 2 and 1 respectively. Only those responses which had complete ranking provided by the respondent were considered for this test.

3. Results

3.1. Fitted boundary lines to data sets

The summary statistics and plots for the variables in the datasets used in this study are provided in Appendix A. From the initial exploratory analysis, the variable P, K and Mg from Dataset 1, SOC, K, P, Mg, Ca and N from Dataset 2 and, P and Ca from Dataset 3 were transformed to natural logs to make the assumption of underlying normality plausible. For pH in Dataset 1, the assumption of normality was achieved with Yeo-Johnson transformation (Yeo and Johnson, 2000). The test for evidence of a limiting boundary (Miti et al., 2024c) in the datasets showed that a boundary was likely (small p-values) in Dataset 1 for the variables P, Mg, K and pH while for Datasets 2 and 3, none of the variables exhibited any evidence of a boundary (See Appendix B).

All the four boundary line fitting methods were applied for yield responses to variables in Datasets 1 and 2. For Dataset 3, the binning methodology was not applied because the selected boundary data points could not represent the boundary well with the proposed bin sizes (10 bins) for all the variables. The standard deviation of the measurement error, a fixed parameter in the cbvn, was estimated to be 0.4 t ha^{-1}, 0.6 t ha^{-1} and 0.15 t ha^{-1} for Datasets 1, 2 and 3 respectively using the likelihood profiling methods (see Appendix C). The various model parameters for Dataset 1, 2 and 3 for the various boundary line fitting methods are presented in Tables 2, 3 and 4 respectively. The standard errors of the parameters are given in the brackets for the cbvn. The AIC values for the censored (Λ_{bl}) and uncensored bivariate normal (Λ_{bvn}) model for variables in Datasets 1, 2 and 3 all indicated that the boundary model was appropriate for all the variables in the datasets as the Λ_{bl} was always less than Λ_{bvn} except for the variables N and Mg in Dataset 3. However, the Aikaike weights for the boundary line models of these two variables showed a larger probability (close to 50%) for the boundary line model. The determined critical values for variable which were transformed are converted back to the original scale (see Fig. 1).

3.2. Comparison of the modelled most-limiting factors

The boundary line parameters reported in Section 3.1 were used to model the most-limiting factors for each point in the datasets using Eq. (5) described in Section 2.2.4. Fig. 2 shows the proportions of the identified most-limiting factors modelled by the different boundary line fitting methods for the three datasets. For Dataset 1, the identified most-limiting factor proportions for K, Mg, P and pH were similar for all methods. However, the binning method at $\tau = 95$ identified K and pH as most-limiting factors more often than the other methods. This difference is also reflected by the limiting factor being unknown for a smaller proportion of the variables for the binning method ($\tau = 0.95$). For Dataset 2, the binning($\tau = 0.95$) and quantile regression($\tau = 0.95$) methods generally differed from the other methods in identification of the most-limiting factor, particularly for N, P, pH and SOC. In instances where other methods did not identify a limiting factor, the quantile regression($\tau = 0.95$) largely allocated it to N and pH. For binning ($\tau = 0.95$), the smaller proportion of P identified as most-limiting factor compared to other methods was compensated by a larger proportion of N and SOC as the identified most-limiting factor.

For Dataset 3, the proportions of the identified most-limiting factors were similar for the four boundary line fitting methods. However, there was a notable difference for the proportion of Mg identified as the limiting factor for the BOLIDES method. While other methods identified Mg as the most-limiting factor in about 5% of the cases, the BOLIDES identified it in 60% of the cases. This was also reflected by a smaller

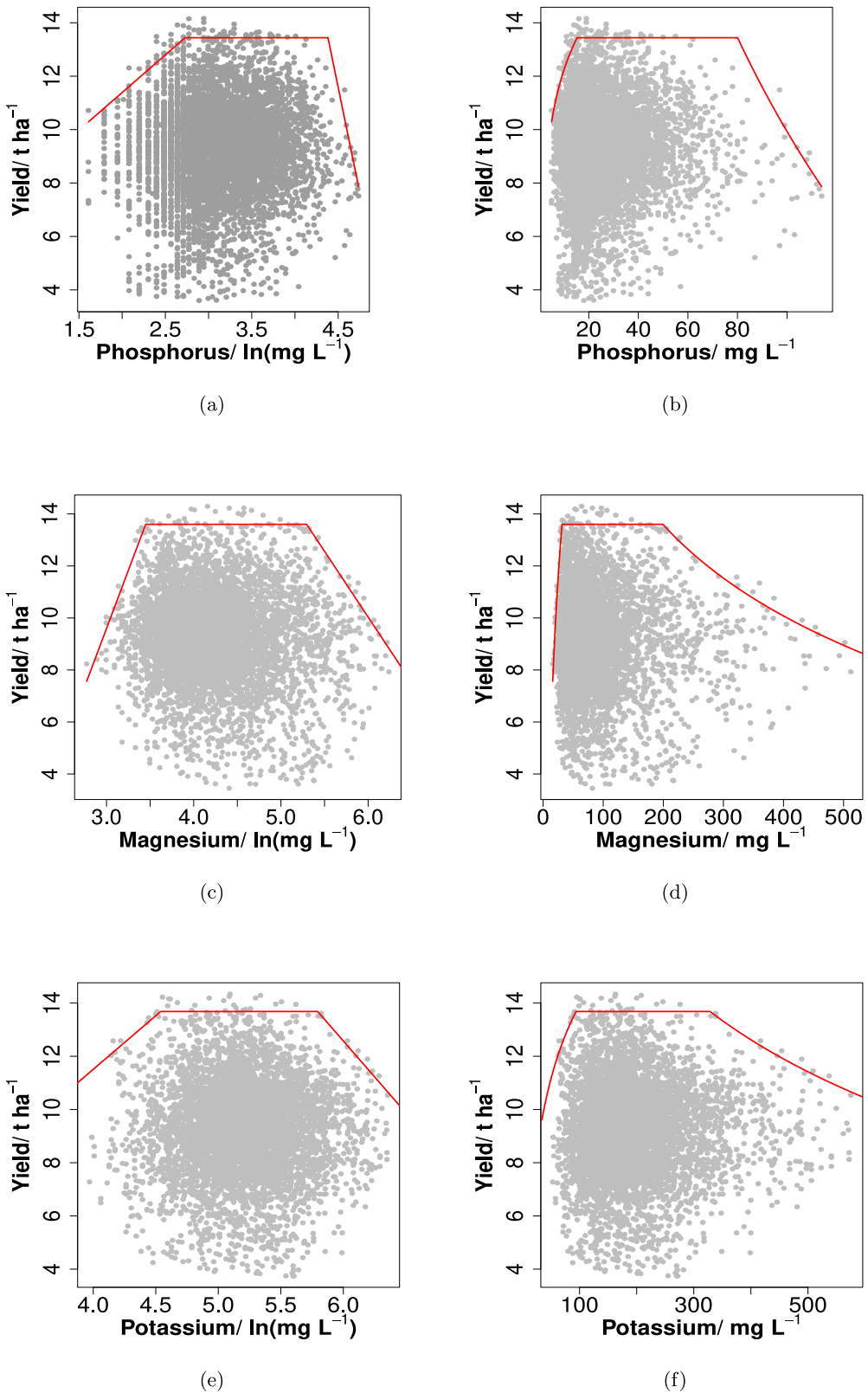


Fig. 1. Boundary lines fitted using the censored bivariate normal model for the variables P, Mg, and K from Dataset 1, displayed on the log-transformed scale (panels a, c, and e) and the original scale (panels b, d, and f) respectively.

allocation to the unknown limiting factor for BOLIDES (5%) compared to the other methods (about 60%).

Table 5 shows the κ values for comparison of the identified most-limiting factors modelled by different boundary line fitting methods with the corresponding z-statistic values and p-values for Datasets 1, 2 and 3. For Dataset 1, overall agreement (when all models were compared) was 0.52 which is fair to good agreement. This is similar to when only heuristic methods were compared (κ = 0.50). The comparison

Table 3
Parameters of fitted boundary lines of the potential limiting factors for Dataset 2.

Method	Factor	Model	β_1	β_2	β_0	β_3	β_4	Λ_{bl}	Λ_{bvn}	ω_{bl}	ω_{bvn}
Binning ($\tau = 0.95$)	SOC	Linear Plateau	21.75	41.86	22.37	–	–	–	–		
Binning ($\tau = 0.95$)	N	Linear Plateau	49.68	10.21	25.07	–	–	–	–	–	–
Binning ($\tau = 0.95$)	pH	Linear Plateau	−232.37	50.48	23.44	–	–	–	–	–	–
Binning ($\tau = 0.95$)	P	Trapezium	−1.87	33.97	23.71	35.40	−5.73	–	–	–	–
Binning ($\tau = 0.95$)	K	Linear Plateau	67.18	21.98	23.41	–	–	–	–	–	–
Binning ($\tau = 0.95$)	Mg	Linear Plateau	36.48	7.43	23.52	–	–	–	–	–	–
Binning ($\tau = 0.95$)	Ca	Linear Plateau	29.12	6.76	22.50	–	–	–	–	–	–
Binning ($\tau = 0.99$)	SOC	Linear Plateau	22.56	43.87	23.44	–	–	–	–	–	–
Binning ($\tau = 0.99$)	N	Linear Plateau	58.38	12.40	25.03	–	–	–	–	–	–
Binning ($\tau = 0.99$)	pH	Linear Plateau	−263.79	57.05	24.81	–	–	–	–	–	
Binning ($\tau = 0.99$)	P	Trapezium	0.29	27.99	25.64	45.29	−9.10	–	–	–	–
Binning ($\tau = 0.99$)	K	Linear Plateau	74.95	24.92	24.55	–	–	–	–	–	–
Binning ($\tau = 0.99$)	Mg	Linear Plateau	40.57	8.47	24.16	–	–	–	–	–	–
Binning ($\tau = 0.99$)	Ca	Linear Plateau	32.34	8.00	23.23	–	–	–	–	–	–
BOLIDES	SOC	Linear Plateau	20.50	32.10	26.17	–	–	–	–	–	–
BOLIDES	N	Linear Plateau	56.54	10.97	27.05	–	–	–	–	–	–
BOLIDES	pH	Linear Plateau	−260.39	55.44	26.66	–	–	–	–	–	–
BOLIDES	P	Trapezium	4.24	19.36	26.81	58.82	−13.89	–	–	–	–
BOLIDES	K	Linear Plateau	63.53	21.66	27.05	–	–	–	–	–	–
BOLIDES	Mg	Linear Plateau	36.24	6.91	26.81	–	–	–	–	–	–
BOLIDES	Ca	Linear Plateau	29.25	5.80	26.67	–	–	–	–	–	–
Q.reg ($\tau = 0.95$)	SOC	Linear Plateau	19.91	10.75	26.09	–	–	–	–	–	–
Q.reg ($\tau = 0.95$)	N	Linear Plateau	61.99	11.83	25.50	–	–	–	–	–	–
Q.reg ($\tau = 0.95$)	pH	Linear Plateau	−254.82	54.42	25.25	–	–	–	–	–	–
Q.reg ($\tau = 0.95$)	P	Trapezium	4.61	19.86	25.83	50.29	−10.54	–	–	–	–
Q.reg ($\tau = 0.95$)	K	Linear Plateau	73.56	25.07	25.25	–	–	–	–	–	–
Q.reg ($\tau = 0.95$)	Mg	Linear Plateau	43.64	8.38	25.25	–	–	–	–	–	–
Q.reg ($\tau = 0.95$)	Ca	Linear Plateau	30.56	6.33	25.50	–	–	–	–	–	–
Q.reg ($\tau = 0.99$)	SOC	Linear Plateau	20.10	18.51	27.33	–	–	–	–	–	–
Q.reg ($\tau = 0.99$)	N	Linear Plateau	53.17	9.05	27.33	–	–	–	–	–	–
Q.reg ($\tau = 0.99$)	pH	Linear Plateau	−254.83	54.42	26.53	–	–	–	–	–	–
Q.reg ($\tau = 0.99$)	P	Trapezium	4.46	20.98	27.33	50.29	−10.54	–	–	–	–
Q.reg ($\tau = 0.99$)	K	Linear Plateau	74.36	25.36	26.60	–	–	–	–	–	–
Q.reg ($\tau = 0.99$)	Mg	Linear Plateau	44.10	8.48	27.33	–	–	–	–	–	–
Q.reg ($\tau = 0.99$)	Ca	Linear Plateau	30.58	6.33	27.33	–	–	–	–	–	–
cbvn	SOC	Linear Plateau	20.58 (0.59)	32.23 (1.82)	26.33 (0.42)	–	–	**790.10**	799.48	0.99	0.01
cbvn	N	Linear Plateau	61.84 (2.18)	11.82 (0.59)	26.48 (0.43)	–	–	**887.53**	893.40	0.95	0.05
cbvn	pH	Linear Plateau	−255.49 (12.62)	54.53 (2.54)	26.48 (0.42)	–	–	**852.25**	858.15	0.95	0.05
cbvn	P	Trapezium	4.47 (0.67)	20.08 (0.84)	26.48 (0.44)	49.51 (12.44)	−10.24 (4.84)	**879.60**	893.61	0.99	0.01
cbvn	K	Linear Plateau	54.99 (0.15)	15.31 (0.29)	26.47 (0.44)	–	–	**934.84**	935.44	0.57	0.43
cbvn	Mg	Linear Plateau	44.09 (1.40)	8.48 (0.36)	26.48 (0.44)	–	–	**1060.17**	1065.42	0.93	0.07
cbvn	Ca	Linear Plateau	29.54 (1.42)	6.07 (0.75)	26.33 (0.43)	–	–	**1086.76**	1092.32	0.94	0.06

cbvn, censored bivariate normal model; Q.reg, quantile regression model; Λ_{bl} and Λ_{bvn}, AIC values for the censored and uncensored bivariate normal models; ω, Aikaike weight for the fitted model.
Bold values of Λ_{bl} and Λ_{bvn} indicate the smaller of the two values.

Table 4
Parameters of fitted boundary lines of the potential limiting factors for Dataset 3.

Method	Factor	Model	β_1	β_2	β_0	Λ_{bl}	Λ_{bvn}	ω_{bl}	ω_{bvn}
BOLIDES	N	linear-plateau	−45.19	396.10	9.05				
BOLIDES	pH	linear-plateau	−37.12	9.23	9.25				
BOLIDES	SOC	linear-plateau	−8.99	9.42	9.05				
BOLIDES	P	linear-plateau	6.19	2.29	9.17				
BOLIDES	Ca	linear-plateau	−5.58	8.77	9.17				
BOLIDES	Mg	linear-plateau	1.67	2.75	9.02				
BOLIDES	K	linear-plateau	3.51	7.64	9.05				
Q.reg ($\tau = 0.99$)	N	linear-plateau	0.14	47.34	9.25				
Q.reg ($\tau = 0.99$)	pH	linear-plateau	−35.02	8.81	9.25				
Q.reg ($\tau = 0.99$)	SOC	linear-plateau	−3.96	6.66	9.25				
Q.reg ($\tau = 0.99$)	P	linear-plateau	6.48	2.38	9.25				
Q.reg ($\tau = 0.99$)	Ca	linear-plateau	−5.58	8.77	9.25				
Q.reg ($\tau = 0.99$)	Mg	linear-plateau	4.84	1.49	9.25				
Q.reg ($\tau = 0.99$)	K	linear-plateau	3.51	8.29	9.25				
Q.reg ($\tau = 0.95$)	N	linear-plateau	0.15	47.33	8.66				
Q.reg ($\tau = 0.95$)	pH	linear-plateau	−31.21	7.98	8.65				
Q.reg ($\tau = 0.95$)	SOC	linear-plateau	−3.96	6.66	8.67				
Q.reg ($\tau = 0.95$)	P	linear-plateau	6.43	2.27	9.02				
Q.reg ($\tau = 0.95$)	Ca	linear-plateau	−5.53	8.72	8.65				
Q.reg ($\tau = 0.95$)	Mg	linear-plateau	1.43	3.06	8.67				
Q.reg ($\tau = 0.95$)	K	linear-plateau	3.51	8.29	8.67				
cbvn	N	linear-plateau	0.14 (0.22)	47.34 (0.08)	9.06 (0.24)	99.25	**98.76**	0.44	0.56
cbvn	pH	linear-plateau	−36.95 (0.01)	9.19 (0.04)	8.97 (0.20)	**377.62**	382.35	0.91	0.09

(*continued on next page*)

Table 4 (*continued*).

Method	Factor	Model	β_1	β_2	β_0	Λ_{bl}	Λ_{bvn}	ω_{bl}	ω_{bvn}
cbvn	SOC	linear-plateau	−8.07 (0.06)	9.13 (0.11)	8.99 (0.22)	**391.44**	394.22	0.8	0.2
cbvn	P	linear-plateau	6.33 (0.25)	2.35 (0.11)	9.06 (0.22)	**482.91**	485.55	0.79	0.21
cbvn	Ca	linear-plateau	−5.58 (3.73)	8.77 (2.52)	9.07 (0.23)	**337.68**	337.77	0.51	0.49
cbvn	Mg	linear-plateau	1.44 (1.71)	3.05 (0.95)	8.98 (0.22)	506.20	**505.37**	0.4	0.6
cbvn	K	linear-plateau	3.41 (0.39)	8.27 (1.04)	9.07 (0.26)	**374.68**	377.80	0.83	0.17

cbvn, censored bivariate normal model; Q.reg, quantile regression model; Λ_{bl} and Λ_{bvn}, AIC values for the censored and uncensored bivariate normal models; ω, Aikaike weight for the fitted model.
Bold values of Λ_{bl} and Λ_{bvn} indicate the smaller of the two values.

Table 5
The κ coefficients for most-limiting factors modelled by different boundary line fitting methods for Dataset 1, 2 and 3.

Dataset	Model comparison	raters	κ	*z*-statistic	*p*-value
1	All models	6	0.52	258.00	<0.01
1	heuristic models	5	0.50	203.00	<0.01
1	cbvn and BOLIDES	2	0.64	78.00	<0.01
1	cbvn and Binning ($\tau = 0.99$)	2	0.55	64.00	<0.01
1	cbvn and Q.Reg ($\tau = 0.99$)	2	0.77	92.60	<0.01
1	cbvn and Binning ($\tau = 0.95$)	2	0.32	51.20	<0.01
1	cbvn and Q.Reg ($\tau = 0.95$)	2	0.63	77.20	<0.01
1	Binning ($\tau = 0.99$) and Binning ($\tau = 0.95$)	2	0.39	54.70	<0.01
1	Q.Reg ($\tau = 0.95$) and Q.Reg ($\tau = 0.99$)	2	0.67	81.60	<0.01
2	All models	6	0.59	52.60	<0.01
2	heuristic models	5	0.55	41.20	<0.01
2	cbvn and BOLIDES	2	0.82	17.50	<0.01
2	cbvn and Binning($\tau = 0.99$)	2	0.82	17.50	<0.01
2	cbvn and Binning($\tau = 0.95$)	2	0.40	8.58	<0.01
2	cbvn and Q.Reg ($\tau = 0.99$)	2	0.81	17.20	<0.01
2	cbvn and Q.Reg ($\tau = 0.95$)	2	0.39	9.08	<0.01
2	Binning($\tau = 0.99$) and Binning($\tau = 0.95$)	2	0.46	10.10	<0.01
2	Q.Reg ($\tau = 0.95$) and Q.Reg ($\tau = 0.99$)	2	0.40	9.80	<0.01
3	heuristic models	3	0.37	14.10	<0.01
3	cbvn and BOLIDES	2	0.31	14.00	<0.01
3	cbvn and Q.Reg ($\tau = 0.99$)	2	0.86	18.40	<0.01
3	cbvn and Q.Reg ($\tau = 0.95$)	2	0.93	18.90	<0.01
3	Q.Reg ($\tau = 0.95$) and Q.Reg ($\tau = 0.99$)	2	0.86	18.30	<0.01
3	Omission trial and BOLIDES	2	0.01	0.31	0.75
3	Omission trial and Q.Reg($\tau = 0.99$)	2	0.00	0.17	0.87
3	Omission trial and Q.Reg($\tau = 0.99$)	2	0.00	0.00	0.99
3	Omission trial and cbvn	2	0.00	0.04	0.97

between the cbvn and any of the heuristic methods produced fair to good agreements with the exception of binning ($\tau = 0.95$) which had poor agreement. For the methods that require a subjective choice of τ, the comparison of the results using the two values had poor agreement for binning and fair to good for quantile regression. For Dataset 2, the overall agreement was fair to good when all methods were compared and similar when all the heuristic methods were compared. Good to excellent agreement was observed for comparison of the cbvn and each of the heuristic methods. The comparison between the two quantile regression methods as well as the two binning methods ($\tau = 0.95$ and $\tau = 0.99$) had fair to good agreements. For Dataset 3, only the cbvn, BOLIDES, quantile regression ($\tau = 95$) and quantile regression ($\tau = 99$), were compared. The overall agreement was fair to good agreement when all methods were compared. A poor agreement was observed when only heuristic methods were compared. The comparison between cbvn and any of the heuristic methods had excellent agreements except for the BOLIDES method which had poor agreement. The comparison of the two quantile regression methods produced excellent agreement.

Comparing the identified most-limiting factor using the boundary line methodology and the omission trials for Dataset 3, poor agreements were observed ($\kappa < 0.1$ for all comparisons). The *p*-values were large indicating that these κ were not significantly different from zero. The omission trial identified N as the most-limiting factor in 71.6% of the cases, P in 12.8% of the cases, K in 3.4% of the cases and 12.2% of the cases did not identify any of the studied factors as most limiting (Fig. 2(d)). On the other hand, the boundary line methods on average identified N as limiting factor in 1.4% of cases, P in 15.7% of cases, K in 3.2% of cases, Mg in 19.3% of cases, Ca in 0.8% of cases, SOC in 7.8% of cases, pH in 3.2% of cases and 48.6% of cases were unknown (Fig. 2(c)).

3.3. Comparison of critical nutrient concentration from various boundary line techniques

Table 6 shows the critical values associated with the variables in the three datasets. The upper and lower bounds of 95% confidence intervals (highest density interval) determined by the cbvn are also given. Note that the critical values for the transformed variables were back transformed to the original scale after fitting the boundary lines for easy comparison with recommended RB209 index values.

For all datasets, the critical values obtained using the quantile regression methods fell within the 95% CI of the cbvn in most instances except for Mg, K and pH in Datasets 1 and SOC in Dataset 2. However, the critical Mg values just fell at the border of the lower limit of the CI. The values obtained using the BOLIDES were consistently within the 95% CI of cbvn except for few instances mostly in Dataset 2. The critical values obtained using the binning methods mostly fell outside the 95% CI of the cbvn. Comparing the estimated critical nutrient values for Dataset 1 with the RB209 guideline indices, the critical P concentrations determined by heuristic methods all fell in the upper section of index 1 with the exception of the quantile regression ($\tau = 0.99$) which, like the cbvn, fell in the lower section of index 2. However the 95% CI of cbvn stretches further into index 2 (Fig. 3(a)). The critical Mg concentrations fell in the lower to middle section of index 1 for all methods except for the binning ($\tau = 0.95$) which fell in index 0. It is worth noting that for the variables Mg and K, there was a small number data points in the index 0. The fitted boundary line models for K using the binning and quantile regression methods at $\tau = 0.95$ (i.e initially plateau then linear model with negative slope) indicate that all values of K were above the critical concentration i.e no value of

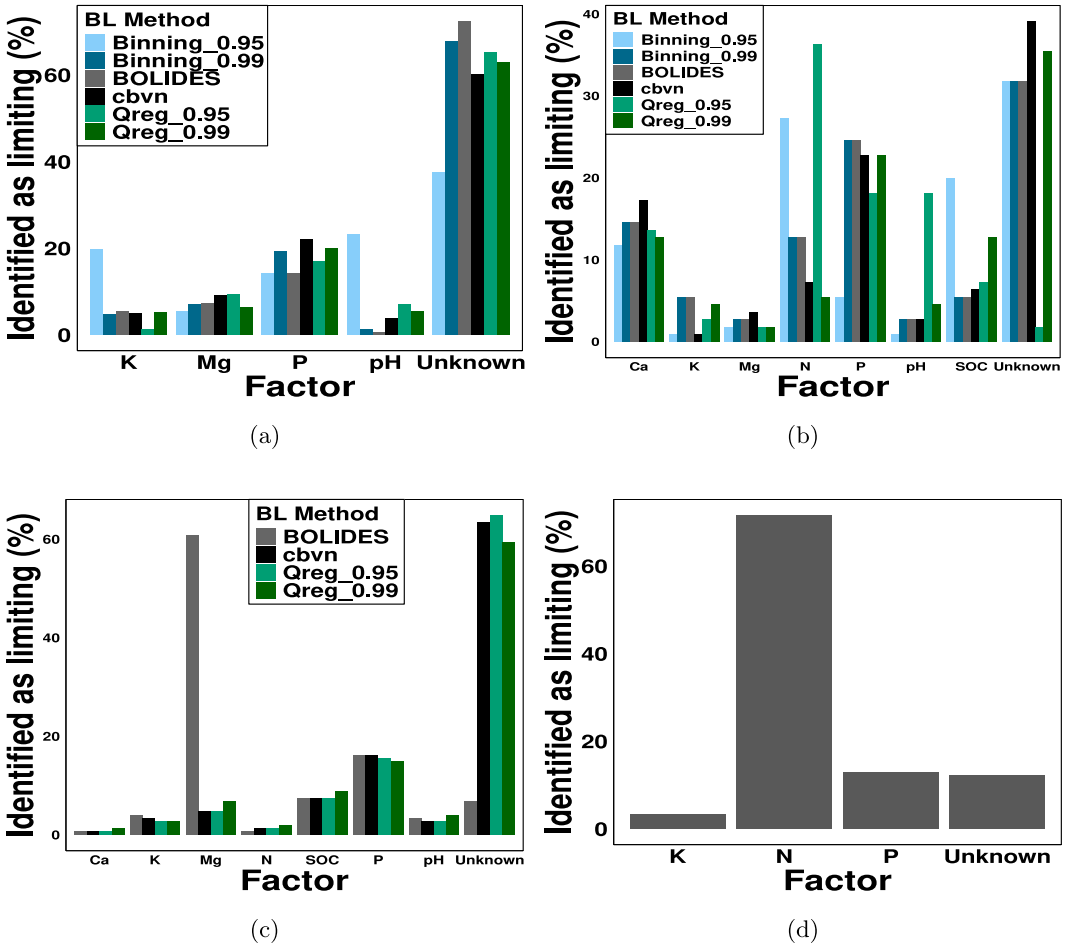


Fig. 2. Proportions of identified limiting factors modelled using four boundary line fitting methods for (a) Dataset 1, (b) Dataset 2 (c) Dataset 3, and (d) limiting factor identified using nutrient omission trials (Dataset 3).

Table 6
Critical values obtained from boundary line parameters and the upper and lower limit from the censored bivariate normal model.

Data	var	Units	cbvn	Lower	Upper	BOLIDES	Bin95	Bin99	QR95	QR99
1	P	mg L^{-1}	15.29	12.65	17.80	12.76	12.88	12.74	13.63	15.39
1	Mg	mg L^{-1}	33.42	30.38	37.47	31.36	**3.47**	**26.84**	**29.34**	**29.20**
1	pH	–	6.06	5.93	6.23	**5.84**	5.99	5.97	6.00	**6.59**
1	K	mg L^{-1}	76.61	66.19	90.94	70.21	–	83.43	**25.22**	89.84
2	SOC	g kg^{-1}	1.20	1.13	1.27	1.19	**1.02**	**1.02**	**1.77**	**1.48**
2	N	%	0.05	0.05	0.06	**0.07**	**0.09**	**0.07**	0.05	0.06
2	pH	–	5.17	5.14	5.20	5.18	**5.07**	**5.06**	5.15	5.17
2	P	mg kg^{-1}	2.99	2.81	3.19	**3.21**	**2.12**	**2.47**	2.91	2.98
2	K	cmol kg^{-1}	0.16	0.13	0.19	0.19	0.14	0.13	0.15	0.15
2	Mg	cmol kg^{-1}	0.13	0.10	0.15	**0.26**	**0.18**	0.14	0.11	0.14
2	Ca	cmol kg^{-1}	0.59	0.42	0.89	0.64	**0.38**	**0.32**	0.45	0.60
3	N	%	0.19	0.17	0.21	**0.14**	–	–	0.19	0.18
3	pH	–	5.00	4.87	5.17	5.03	–	–	5.03	5.00
3	SOC	%	1.87	1.75	2.01	1.92	–	–	1.98	1.90
3	P	mg kg^{-1}	3.20	2.47	4.10	3.67	–	–	3.21	3.14
3	Ca	cmol kg^{-1}	5.32	4.73	6.38	5.38	–	–	5.43	5.08
3	Mg	cmol kg^{-1}	2.47	2.04	3.56	2.68	–	–	2.95	2.37
3	K	cmol kg^{-1}	0.68	0.58	0.82	0.73	–	–	0.69	0.62

All values are on the original measurement scale.
Values in **bold** are outside the 95% confidence interval of the censored bivariate normal model.
RB209 P Index: **0**, 0–9 mg L^{-1}; **1**, 10–15 mg L^{-1}; **2**, 16–25 mg L^{-1}; **3**, 26–45 mg L^{-1}; **4**, 46-70 mg L^{-1}
RB209 Mg Index: **0**, 0–25 mg L^{-1}; **1**, 26–50 mg L^{-1}; **2**,51–100 mg L^{-1}; **3**, 101–175 mg L^{-1}; **4**, 176–250 mg L^{-1}
RB209 K Index: **0**, 0–60 mg L^{-1}; **1**, 61–120 mg L^{-1}; **2**–, 121–180 mg L^{-1}; **2**+ , 181–240 mg L^{-1}; **3**, 241–600 mg L^{-1}

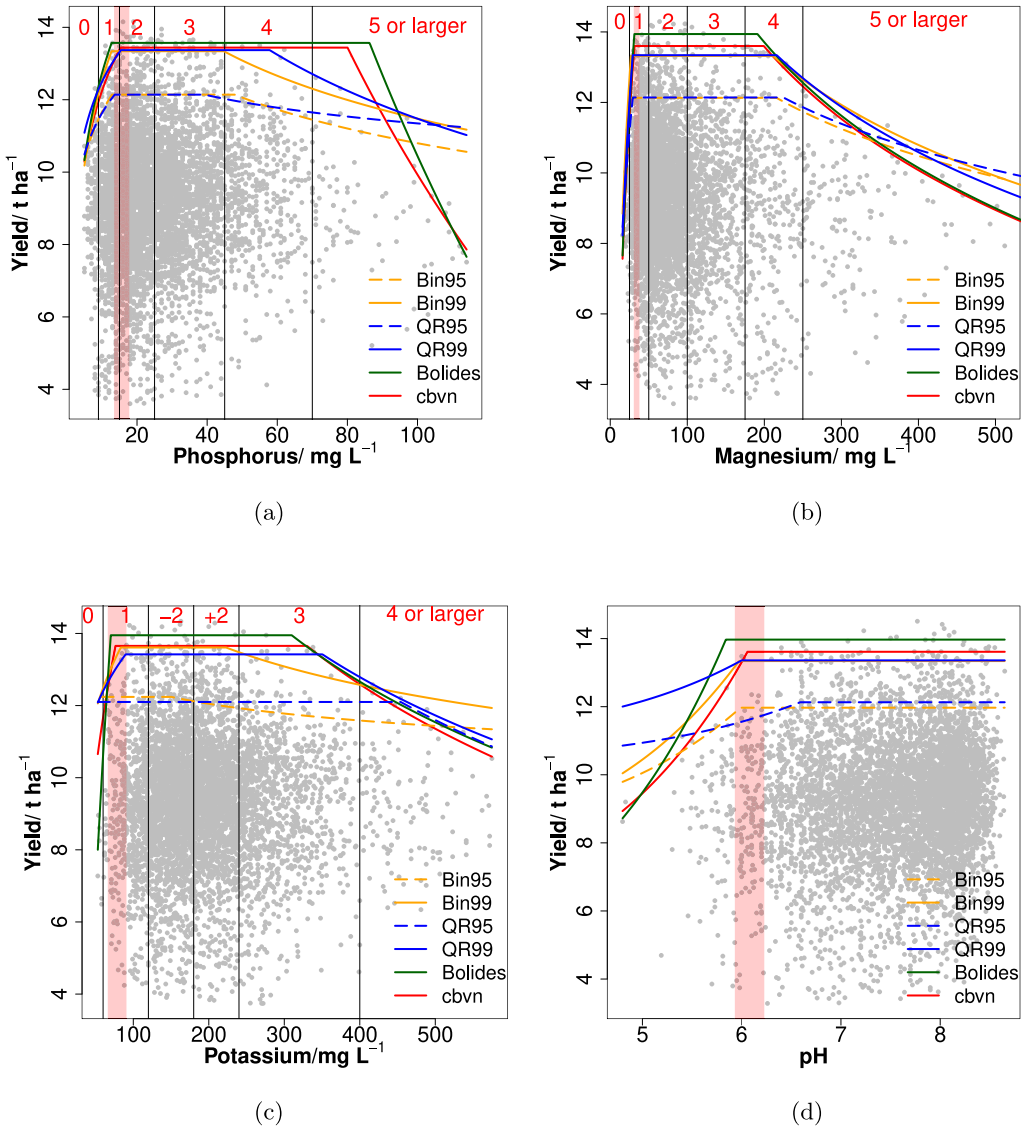


Fig. 3. Boundary line models for (a) soil P, (b) soil Mg, (c) soil K and (d) Soil pH fitted using the various methods. The reddish area is the 95% CI of x_{crit} from the cbvn model. The black vertical lines mark the boundaries used for the RB209 index categories: 0, 1, 2, 3, 4, and 5.

measured K could have limited yield by deficiency. According to the RB209 guidelines, the optimum pH for mineral soils of England is 6.5. This falls just outside the upper bound of the 95% CI of cbvn.

3.4. Stakeholder preferences

Fig. 4 shows the mean ranks of stakeholders' preferences for the different boundary line fitting methods from the study. The mean ranks for binning, BOLIDES, cbvn and quantile regression for all participants were 3.17, 2.45, 2.31 and 2.07 respectively. The test statistic (χ^2 = 2.71, df = 3) for all the participants showed that there was no evidence (p = 0.56) against the null hypothesis of uniform ranking. For adequacy of output from the different boundary line fitting methods for yield gap analysis (Fig. 4(b)), 4 participants stated all methods, 7 binning, 1 BOLIDES and 20 cbvn. Twenty three participants stated that information on uncertainty was an important output for interpretation while 9 did not give an opinion. For the adequacy of the boundary line methodology in general for yield gap analysis (Fig. 4(c)), 23 participants stated that it is adequate, 7 stated it is not adequate and 2 did not give an opinion.

The results from the elicitation exercise in the Nairobi and Harare workshops did not indicate any special preference of a particular method. The ranking of the boundary line fitting methods was uniform for the whole dataset and hence we did not further split the data by locations and statistical experience for further analysis. However, some useful information for the reasons of their ranking was obtained. Some participants stated that the input parameters for the cbvn may be unavailable and difficult to estimate (especially measurement error) as reason for ranking it lower. However, the information on uncertainty of parameters was recognized as an important output for boundary lines using the cbvn. While more participants stated that the boundary line methodology was sufficient for yield gap analysis as a whole, some participants stated that it is inadequate as it does not incorporate the interaction effects on biological response, and in the cases of the cbvn method, cannot be applied to categorical variables including social economic factors.

4. Discussion

The comparison of boundary line fitting methods for yield gap analysis was undertaken to check the consistency among the different methods to model the most-limiting factor and critical nutrient concentrations. The results for the three datasets show the importance of

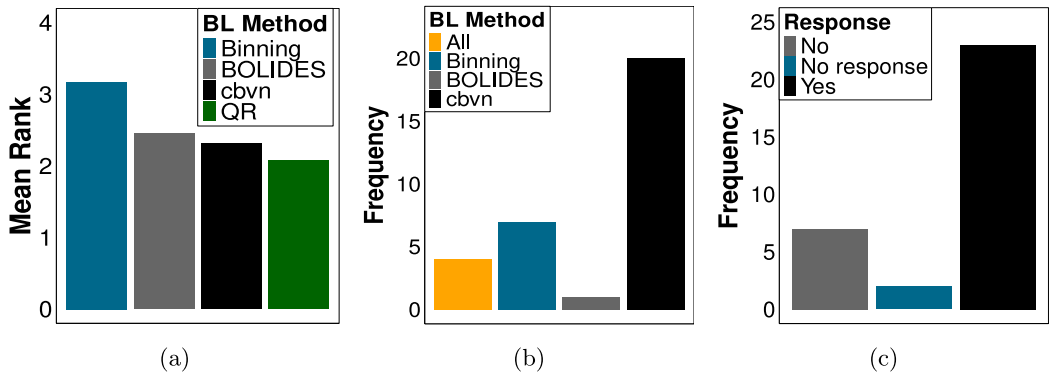


Fig. 4. (a) Ranking of the boundary line fitting methods according to user-friendliness for all participants, (b) Adequacy of output for interpretation (c) Adequacy of boundary line methodology as a whole for yield gap analysis.

considering whether a boundary model is justified, rather than automatically applying it. This is not generally done, and heuristic methods offer no basis for assessing whether the model is justified on the basis of its fit. One could, after all, draw lines through the upper fringes of any bivariate plot, and this is what the peel method aims to address at the exploratory stage, and what the AIC assessments of the cbvn model does more powerfully. Unlike heuristic methods, the statistical cbvn approach fits both bounded and unbounded models to the data, enabling model comparison. By evaluating AIC values, this method assesses whether the additional boundary parameters improve model fit. A lower AIC for the bounded model indicates that the boundary is statistically justified, supporting its interpretation. Without this step, any interpretation of the boundary would lack statistical validation and could be misleading. In addition, one can also check for evidence of clustering in the upper edges of a data cloud as justification to carry out boundary line analysis.

Dataset 1 showed evidence for boundary structure from the peel tests (*sd* mostly smaller than the bootstrap mean, and some small *p*-values) (see Appendix B) and this was supported by the AIC assessments with decisive evidence for the cbvn model with Akaike weights all above 0.9 and mostly above 0.99 (Table 2). With Dataset 2, the evidence from the peel tests was mostly positive, but larger *p*-values(Appendix B). This is consistent with smaller size of dataset and observations on limited range for some variables. Dataset 3 mostly provides no evidence for boundary structure (*sd* larger than bootstrap mean in 5 out of 7 cases for left bound and all cases for right) and assessment of the AIC rejects boundary model for N and Mg, and is much less decisive for other variables with Akaike weights for cbvn less than 0.9 for all but pH (Table 4). This shows that the AIC assessment of the cbvn model is more powerful than the peel test, but is consistent with it. The small size of Dataset 3 makes it difficult to establish evidence of boundary existence as there are less data points in the upper edges of the data cloud. These findings indicate that results based on boundary line models from Datasets 2 and 3 should be interpreted with caution, as there is no clear evidence of boundary structures in these datasets.

Agronomically plausible models were fitted to the datasets by accessing the upper boundary data structure in relation to the possible underlying biophysical and agronomic theories as emphasized by Sadras (2020). For example in Dataset 1, a piecewise trapezium model was fitted for the relationship of Mg and wheat yield. This is because an increase in soil Mg concentration might be expected to lead to increases in yield until the critical concentration above which yields do not increase further (AHDB, 2023). In general, soil Mg concentrations larger than the critical concentration do not damage crop growth, but may hinder the uptake of other cations such as K. So, where soil Mg is excessively high (index 4 and above of the RB209 guidelines), a reduction in yield may be observed due to limitation of other nutrients (PDA, 2017). This was consistent with our fitted model as the second inflection point fell within the range of index 4.

Apart from the *piece-wise* linear models, other complex models like the quadratic, logistic (Nelder, 1961; Oliver, 1964), double-logistic (Shabani et al., 2018), inverse logistic (Helidoniotis et al., 2011), monomolecular (Draper and Smith, 1998), Gompertz (Draper and Smith, 1998), Weibull (Myers and Myers, 1990), Schmidt (Schmidt et al., 2000) and Mitscherlich (Mitscherlich, 1909) models among others may be used to fit the boundary lines. In this study, we used the *piece-wise* linear models to enable the determination of the critical values at the inflection points. In cases where other complex models are used, they may produce similar fits and the simplest model should be adopted (Bargagli Stoffi et al., 2022). However, differences in model structure/form has been identified as one of the sources of uncertainty in determination of yield gap from process models (Schils et al., 2022). This may affect the boundary line model outputs in a similar way. There is need for further studies to evaluate the sensitivity of model form on the output of the boundary line methodology.

The κ statistic was used to check how consistent boundary line fitting approaches are in modelling the most-limiting factor (Tables 5). It is important to note that the κ statistic is a measure of consistency and not correctness. In general, the approaches are consistent in the determination of the most limiting factor but some inconsistencies were observed for the comparisons between the binning methods in Dataset 1 and, the comparison of quantile regression ($\tau = 0.95$) and cbvn in Dataset 2. These inconsistencies highlight the influence of the subjective decisions in the implementation of the heuristic binning and quantile regression methods, where the boundary is defined based on an arbitrary choice of τ. This emphasizes the need for the use of more objective methods to fit boundary line models if consistency is to be achieved. There was one notable instance for Dataset 3, where soil Mg was found to be limiting in substantially more cases using the BOLIDES than the other methods (Fig. 2(c)). This was also reflected in the poor agreement ($\kappa < 0.4$) for the heuristic methods. This may have been due to the effect of the data size and the arrangement of data points at the upper edge of the scatter. Unlike all other methods, the BOLIDES fits the boundary line model to the extreme points that cover the upper part of the cloud and therefore, has a higher sensitivity to the distribution of the boundary points. This means that a few extreme points, including points that are not considered as outliers, could determine the model parameters when the BOLIDES method is applied.

In the case of the omission trial, the limiting factors identified from the experimental results were not consistent with those identified from the boundary line methodology ($\kappa < 0.4$) (Table 5). This highlights the need for further work to examine the interpretation of boundary line models for yield gap analysis. As has been noted previously by Miti et al. (2024b), the boundary line need not be interpreted exclusively in terms of a limiting factor model in terms of Sprengel-Liebig, where factors have independent effects on the crop response, and only one can be limiting. However, before considering further experimental work

it is necessary to consider the potential reasons for these observed inconsistencies and the two issues which these results raise.

First, the results for exploratory analysis of Datasets 2 and 3 did not provide evidence for boundary structures in the data. In the case of Dataset 2 the comparison of the cbvn model with a null bivariate normal model did indicate that the boundary model was appropriate for all soil variables apart from Potassium (Table 3), with the Akaike weight for the cbvn model exceeding 0.93 in all cases. In contrast, for Dataset 3 from the omission trial (Table 4), the censored bivariate normal model was favoured for soil N and Mg. For other variables, the evidence for the boundary model was limited, with Akaike weights ranging from 0.51 (soil Ca) to 0.91 (soil pH). These results show weak evidence that the data exhibit a boundary which can be confidently modelled and interpreted to identify the most-limiting factor. In particular, there was no evidence for the boundary model in the case of soil N, whereas the largest group of sites in the omission trial was that for which N application had the largest effect on yield.

The small size of Dataset 3 likely contributed to the observed result. It has been emphasized that boundary line analysis requires relatively large datasets to ensure that any bounding behaviour is exhibited, and that there are sufficient observations near the bound to allow robust estimation of its parameters (Sadras, 2020; Shatar and McBratney, 2004; Miti et al., 2024c). Miti et al. (2024c) used simulated data from different distributions to examine the power of the peel cluster test to detect bounded behaviour; they found that a minimum of 400 data points were required. In the case of Dataset 3, the exploratory analysis and analysis with the cbvn model would provide grounds for not proceeding with an interpretation based on the boundary line. In most studies which use heuristic methods to fit the boundary line (Miti et al., 2024b) such an assessment cannot be made.

Another consideration, highlighted by the results for Dataset 3, is whether the appropriate soil variables have been measured to identify a limiting effect which would respond to a particular input. For a macronutrient such as P and K it is widely assumed that measurements of available nutrient concentration in the soil can be used to determine a quantitative recommendation for P or K fertilizer respectively. In the case of soil N, laboratory measurements of either total N or mineral N do not relate simply to a measure of soil N supply which is commensurate with rates of application of fertilizer N. Nitrogen in the soil undergoes constant transformations such as mineralization, nitrification and denitrification which are driven by microbial activity, temperature, moisture and other environmental factors, all of which influence the availability of plant-accessible N (Pruthviraj et al., 2024). It is therefore unlikely that a single soil N measurement will necessarily be indicative of the presence or absence of an N limitation on crop yield. It is notable that there was some evidence, albeit not strong, for a boundary line relationship between yield and SOC in Dataset 3, with an Akaike weight of 0.8. As mineralization of SOC is the primary source of the soil N supply in the absence of artificial fertilizer, this boundary relationship might be indicative of nitrogen limitation. It is necessary to consider carefully what soil measurements are most appropriate to identify and act on yield gaps.

The boundary line fitting approaches are generally consistent in the estimation of critical values. Therefore, the use of one method is not expected to produce results that differ markedly from others. In most instances, the critical values for variables estimated using the heuristic methods all fell within 95% CI of the cbvn or just outside the CI (e.g. Mg in Dataset 1 and N using the BOLIDES in Dataset 3) (Table 6). Most of the cases where the critical values fell outside the 95% CI of cbvn were observed in Dataset 2 especially when the binning methods were used. This may have been due to the combined effect of bin size, τ value considered as boundary and small size of the dataset. From this, we recommend that the boundary line methodology is applied to sufficiently large datasets with enough coverage and sufficient distribution of all possible conditions, and especially those that fulfil the assumption of boundary existence as earlier pointed out.

Generally, results from the different boundary line fitting methods are inconsistent with the RB209 guideline indices with the exception of the cbvn and quantile regression ($\tau = 0.99$) methods for soil P (Table 6). In most instances, the boundary line fitting methods underestimated the critical values placing them in index 1 and in a few case index 0. In these indices we still expect addition of P, K and Mg to increase yield (AHDB, 2023). The inconsistencies in results obtained using the different τ values for the binning and quantile regression methods indicate the impact of the subjective decision on the outcome of the boundary line analysis when heuristic method are used. Andrade et al. (2023) similarly found that varying the τ value for quantile regression had a large effect on the determined critical soil nutrient values. This stresses the need for a more objective way to determine the τ value to consider as the boundary. Makowski et al. (2007) proposed a method to determine a specific τ for quantile regression methods using expert knowledge and previous knowledge on the distribution of measurement errors using Bayesian statistics. However, this is only possible when there is enough information available on the uncertainty of the crop yield being studied. Despite, the inconsistency of the boundary line methodology with standard RB209 indices, similar studies have found that the boundary line methodology is an effective tool for understanding yield response to soil conditions (Smith et al., 2024; Andrade et al., 2023). In these studies, the obtained critical values fell within the recommended guidelines for soils of those regions.

The results from our study indicate that there is little difference in modelled limiting factors and critical nutrient concentrations using the different boundary line fitting methods. However, the statistical cbvn method offers a more objective and consistent approach to fitting boundary line models, with added robustness through uncertainty estimates for the critical nutrient concentration and a strong test for boundary presence. Schut and Giller (2020) and Andrade et al. (2023) highlighted the importance of measures of uncertainty in soil analysis measurements for field-specific fertilizer recommendations. As uncertainty provides a form of risk for decision makers who use boundary line outputs to make decisions, uncertainty should be included as an output of the analysis. This was also emphasized by participants in the consultation exercise in the Nairobi and Harare workshops with the majority of participants indicating that uncertainty of the outputs was important for interpretation. However, attention should not only be focused on the precise quantification of uncertainty, but should also aim to identify the sources of uncertainty and how they can be reduced (Schils et al., 2022). In this study, no measures of uncertainty were given for output of the heuristic methods because these involve subjective decisions in their implementation. Hence the uncertainty around the determined parameters will be strongly influenced by these subjective decisions, making it challenging to interpret. This mostly arises in instances where researchers have different views on reality and therefore, what to be included in a model (Walker et al., 2003).

In practice, the choice of which method to use for analysis is influenced by several factors including data availability (nature of data i.e continuous or categorical, size of data and whether it meets model assumptions among other qualities), usability of the method and the statistical competencies of the analysts (Miti et al., 2024b). For example, Smith et al. (2024) highlighted that they could not apply the cbvn method proposed by Milne et al. (2006a) on their data because the variables could not meet the assumption of normality. Currently, most published studies that have evaluated yield gaps using the boundary line methodology have used heuristic methods (Smith et al., 2024). The ease of use of the heuristic methods has been pointed out as one of the reasons for their popularity as compared to statistical methods (Miti et al., 2024b). However, the questionnaire exercise with researchers in Harare and Nairobi showed that there was no strong evidence that users found the statistical method harder to engage with, and they see value of uncertainty quantification of this approach. The cbvn provides a more robust approach to fitting boundary lines but can benefit further from improvements by extending it to include categorical variables

which currently cannot be analysed. Variables like planting date, days to weeding among others are important determinants of yield gaps and must be analysed to get a full picture of yield gaps. This was pointed out as one of the weaknesses of the cbvn by the workshop participants. Until the limitations of the cbvn method are addressed, heuristic methods will still be the preferred methods for researchers especially in agronomy where categorical variables are important factors in their studies.

The boundary line methodology is a valuable tool for extracting agronomic insights, particularly in this era of big data. However, there remains significant potential to refine this approach and broaden its applicability. Further work is needed to deepen our understanding of the agronomic implications of the boundary line methodology. While much of the current focus has been on interpretations aligned with Sprengel-Liebig Law of the Minimum (van der Ploeg et al., 1999), this perspective could be expanded to include alternative frameworks, such as the Law of the Optimum (Liebscher, 1895). Miti et al. (2024b) have outlined various ways the boundary line models can be interpreted, but more comprehensive biological exploration is required to fully understand their implications. Omission trial experiment provide a good basis to test the ability of the boundary line methodology to model the most-limiting factor. More studies with larger datasets are required to check how the detected most-limiting factor using the boundary lines compares with that from omission trials. This will enhance greater understanding on the interpretation of the boundary line methodology utilizing the Sprengel-Liebig Law of the Minimum. The experimental design of omission trials can be improved by careful thought on what variables should be measured to pick up limitations, in particular to identify limitations which a particular intervention might address (e.g. application of fertilizer). While this approach is more straightforward for less mobile nutrients, pH, and water supply (often excluded from such studies), it poses a greater challenge for dynamic variables like soil N. In omission trials, it is often assumed that adding N fertilizer addresses potential N limitations, but the complexity of soil N dynamics warrants careful consideration when interpreting results.

5. Conclusion

Our study indicates that there is consistency in the determination of the most-limiting factor amongst the different boundary line fitting methods with an overall fair to good agreement index ($\kappa > 0.4$), underscoring the coherence of these approaches for this purpose. Similarly, there is consistency in the critical soil nutrient concentration determined using different boundary line fitting methods with the exception of the binning method highlighting the impact of subjective decisions on outcomes of boundary line analysis when heuristic methods are used. These results provides a start point for the development of standard boundary line procedures for yield gap analysis to promote consistent interpretation and agronomic recommendations, and reproducible research. In contrast, our study indicates poor agreement between the boundary line methodology and both the nutrient omission experimental determination of the most-limiting factor and the RB209 established critical soil nutrient concentrations. Although there is consistency in the outputs of the different boundary line fitting methods, we recommend the use of the cbvn for determination of critical nutrient concentration as it provides more information on the uncertainty of the fitted parameters and therefore, allows for further analysis like the determination of the probability of the critical concentration falling within a given index. This is vital information for decision makers/agronomists to provide adequate solutions to close yield gaps. Furthermore, we recommend more studies with larger datasets to compare the most-limiting factors modelled by boundary line methods with those modelled experimentally to fully assess the use of boundary line methodology as a tool for determining the most-limiting factor.

CRediT authorship contribution statement

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

Declaration of competing interest

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

Acknowledgements

This work was supported by the Nottingham-Rothamsted Future Food Beacon Studentships in International Agricultural Development. Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom (BBSRC). We acknowledge support from the Growing Health Institute Strategic Programme [BB/X010953/1; BBS/E/RH/230003C]. We are thankful to AgSpace Agriculture Ltd, Dr. Lydia Wairegi and International maize and wheat improvement center (CIMMYT) for providing datasets 1, 2 and 3, respectively. We also acknowledge the input of the Research and Ethics committee at Sutton Bonington university of Nottingham campus for the assistance in getting the ethical approval for the workshops in Harare and Nairobi.

Appendix A. Exploratory analysis for datasets used the study

See Table A.1.

Appendix B. Evidence for upper boundary structure in data sets

See Table B.1.

Appendix C. Estimate of the standard deviation of measurement error

See Fig. C.1.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eja.2025.127744.

Data availability

• Dataset 1: A portion of this dataset is included in the BLA R package, freely available on CRAN at https://CRAN.R-project.org/package=BLA
• Dataset 2: Obtained from the study "Closing the cassava yield gap: An analysis from smallholder farms in East Africa", published in Field Crops Research. Access to this dataset can be requested directly from the authors of that publication.
• Dataset 3 (TAMASA Ethiopia): Available on the CIMMYT Research Data & Software Repository Network (Dataverse) at https://data.cimmyt.org/dataset.xhtml?persistentId=hdl:11529/11015, subject to terms and conditions.

Table A.1
Summary statistics for variable in the datasets used for this study.

Data	Factor	Transformation	Power	Mean	Median	sd	Skewness	O.skewness
1	P	None	–	25.96	22	14.38	1.84	0.36
1	P	Natural log	–	3.13	3.09	0.51	0.13	0.08
1	Mg	None	–	83.60	63	68.44	4.77	0.5
1	Mg	Natural log	–	4.24	4.14	0.56	0.86	0.20
1	K	None	–	198.33	183	84.52	2.36	0.21
1	K	Natural log	–	5.21	5.21	0.38	0.16	0.001
1	pH	None	–	7.56	7.74	0.65	−0.79	−0.35
1	pH	Natural log	–	2.02	2.05	0.09	−0.99	−0.40
1	pH	Box-Cox	2	28.34	29.45	4.82	−0.63	−0.31
1	pH	Yeo-Johnson	4.99	0	0.13	1.00	−0.27	−0.19
1	Yield	None	–	9.25	9.36	1.85	−0.48	−0.06
2	pH	None	–	5.82	5.9	0.46	−0.04	−0.15
2	SOC	None	–	1.72	1.61	0.63	1.11	0.32
2	SOC	Natural log	–	0.49	0.48	0.35	0.03	0.12
2	N	None	–	0.083	0.077	0.044	0.80	0.33
2	N	Natural log	–	−2.63	−2.56	0.60	−0.89	0.02
2	P	None	–	4.93	4.03	3.25	2.35	0.40
2	P	Natural log	–	1.44	1.39	0.53	0.56	0.13
2	K	None	–	0.46	0.39	0.31	1.56	0.37
2	K	Natural log	–	−0.98	−0.93	0.65	−0.03	0.01
2	Ca	None	–	2.76	1.63	2.79	0.99	0.55
2	Ca	Natural log	–	0.29	0.49	1.40	−0.43	−0.18
2	Mg	None	–	1.11	0.79	0.99	0.98	0.51
2	Mg	Natural log	–	−0.41	−0.24	1.18	−0.80	−0.19
2	Yield	None	–	13.97	13.86	6.32	0.15	0.013
3	pH	None	–	5.39	5.4	0.40	0.19	0.11
3	OC	None	–	2.17	2.16	0.46	−0.06	0.08
3	N	None	–	0.22	0.23	0.05	−0.07	−0.14
3	P	None	–	5.04	3.3	4.91	2.74	0.49
3	P	Natural log	–	1.25	1.19	0.87	−0.01	0.04
3	Ca	None	–	7.77	7.4	2.25	0.65	0.28
3	Ca	Natural log	–	2.01	2.00	0.29	−0.01	0.14
3	Mg	None	–	3.24	3.2	1.01	0.96	0.07
3	K	None	–	0.98	0.99	0.41	1.12	−0.01
3	Yield	None	–	3.34	2.81	2.09	0.82	0.29

Table B.1
The probability of observing a larger peel clustering (lower *sd* of euclidean distance of boundary points to centre) than that of a normal bivariate joint distribution on the left (l) and right (r) sections.

Dataset	n	Variable	sd_l	$\tilde{sd}_l$	p-value $_l$	sd_r	$\tilde{sd}_r$	p-value $_r$
1	6010	P	1.045	1.181	0.019	1.115	1.276	0.013
1	6010	Mg	1.309	1.153	0.997	1.087	1.250	0.008
1	6010	pH	0.952	1.064	0.035	1.244	1.155	0.877
2	110	SOC	3.610	3.643	0.484	4.077	4.115	0.482
2	110	N	2.778	3.567	0.09	3.990	4.091	0.459
2	110	pH	3.186	3.626	0.238	3.985	4.106	0.423
2	110	P	3.345	3.571	0.369	4.036	4.070	0.505
2	110	K	2.837	3.556	0.111	4.023	3.985	0.567
2	110	Mg	3.2603	3.408	0.421	3.781	3.838	0.494
2	110	Ca	3.025	3.469	0.244	4.330	3.736	0.830
3	148	N	1.169	1.358	0.169	1.541	1.417	0.742
3	148	pH	1.410	1.135	0.841	1.506	1.411	0.679
3	148	SOC	1.151	1.246	0.379	1.530	1.358	0.746
3	148	P	1.330	1.081	0.830	1.382	1.244	0.739
3	148	Ca	1.464	1.206	0.837	1.496	1.409	0.658
3	148	Mg	1.344	1.043	0.869	1.430	1.213	0.816
3	148	K	1.499	1.264	0.802	1.408	1.387	0.553

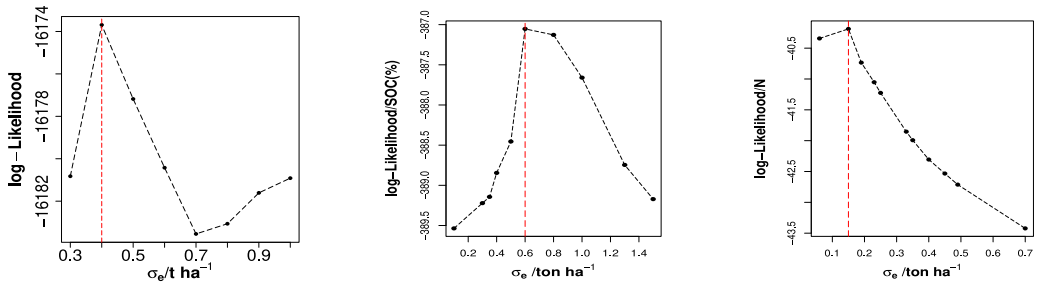


Fig. C.1. Estimate of σ_{me} using the log-likelihood profile for (a) data set 1, (b) Dataset 2 and (c) Dataset 3. The red dashed line represents the estimate of σ_{me}.

References

AHDB, 2023. Nutrient Management Guide (RB209). Section 1: Arable Crops. Agriculture and Horticulture Development Board, Coventry, UK.

Andrade, C.B., Comin, J.J., Moura-Bueno, J.M., Brunetto, G., 2023. Obtaining reference values for nutrients in vineyard soils through boundary line approach using Bayesian segmented quantile regression on commercial farm data. Eur. J. Agron. 150, 126928.

Bargagli Stoffi, F.J., Cevolani, G., Gnecco, G., 2022. Simple models in complex worlds: Occam's razor and statistical learning theory. Minds Mach. 32 (1), 13–42.

Barry, J., 2011. Doing Bayesian data analysis: A tutorial with R and BUGS. Europe' s J. Psychol. 7 (4), 778.

Baudron, F., Ndoli, A., Habarurema, I., Silva, J.V., 2019. How to increase the productivity and profitability of smallholder rainfed wheat in the Eastern African highlands? Northern Rwanda as a case study. Field Crop. Res. 236, 121–131.

Brys, G., Hubert, M., Struyf, A., 2008. Goodness-of-fit tests based on a robust measure of skewness. Comput. Statist. 23, 429–442.

Burnham, K., Anderson, D., 2004. Multimodel inference, understanding AIC and BIC in model selection. Sociol. Methods Res. 33, 261–304.

Cade, B.S., Terrell, J.W., Schroeder, R.L., 1999. Estimating effects of limiting factors with regression quantiles. Ecology 80, 311–323.

Casanova, D., Goudriaan, J., Bouma, J., Epema, G., 1999. Yield gap analysis in relation to soil properties in direct-seeded flooded rice. Geoderma 91 (3–4), 191–216.

Craufurd, P., 2017. TAMASA Ethiopia. Nutrient Omission Trial (NOT) Datasets for 2015 and 2016 Seasons. CIMMYT Research Data & Software Repository Network.

Dobson, A.J., Barnett, A.G., 2018. An Introduction to Generalized Linear Models. CRC Press.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, vol. 326, John Wiley & Sons.

Evanylo, G.K., Sumner, M.E., Taylor & Francis, 1987. Utilization of the boundary line approach in the development of soil nutrient norms for soybean production. Commun. Soil Sci. Plant Anal. 18, 1397–1401.

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., Laborte, 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 Crop. Res. 112 (1), 24–36.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378.

Fleiss, J.L., Cohen, J., Everitt, B.S., 1969. Large sample standard errors of kappa and weighted kappa. Psychol. Bull. 72 (5), 323.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical methods for rates and proportions. In: Statistical Methods for Rates and Proportions, third ed. In: Wiley series in probability and statistics, Wiley-Interscience, Hoboken, N.J.

Foley, J.A., Ramankutty, N., Brauman, K.A., Cassidy, E.S., Gerber, J.S., Johnston, M., Mueller, N.D., O'Connell, C., Ray, D.K., West, P.C., et al., 2011. Solutions for a cultivated planet. Nature 478 (7369), 337–342.

Gamer, M., Lemon, J., Gamer, M.M., Robinson, A., Kendall's, W., 2012. Package 'irr'. Var. Coefficients Interrater Reliab. Agreem. 22, 1–32.

Giller, K.E., Delaune, T., Silva, J.V., Descheemaeker, K., van de Ven, G., Schut, A.G., van Wijk, M., Hammond, J., Hochman, Z., Taulya, G., Chikowo, R., Narayanan, S., Kishore, A., Bresciani, F., Teixeira, H.M., Andersson, J.A., van Ittersum, M.K., 2021. The future of farming: Who will produce our food? Food Secur. 13, 1073–1099.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: The challenge of feeding 9 billion people. Science 327 (5967), 812–818.

Hajjarpoor, A., Soltani, A., Zeinali, E., Kashiri, H., Aynehband, A., Vadez, V., 2018. Using boundary line analysis to assess the on-farm crop yield gap of wheat. Field Crop. Res. 225, 64–73.

Harris, E.F., Smith, R.N., 2009. Accounting for measurement error: A critical but often overlooked process. Arch. Oral. Biol. 54.

Helidoniotis, F., Haddon, M., Tuck, G., Tarbath, D., 2011. The relative suitability of the von Bertalanffy, Gompertz and inverse logistic models for describing growth in blacklip abalone populations (Haliotis rubra) in Tasmania, Australia. Fish. Res. 112 (1–2), 13–21.

Koenker, R., D'Orey, V., 1987. Computing regression quantiles, applied statistics. Appl. Stat. 36, 383–393.

Kopittke, P.M., Menzies, N.W., Wang, P., McKenna, B.A., Lombi, E., 2019. Soil and the intensification of agriculture for global food security. Environ. Int. 132, 105078.

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.

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.

Lee, P.H., Yu, P.L., 2013. An R package for analyzing and modeling ranking data. BMC Med. Res. Methodol. 13, 1–11.

Liebscher, G., 1895. Untersuchungen über die Bestimmung des Düngerbedürfnisses der Ackerböden und Kulturpflanzen. J. Landwirtsch. 43, 49–125.

Makowski, D., Doré, T., Monod, H., 2007. A new method to analyse relationships between yield components with boundary lines. Agron. Sustain. Dev. 27, 119–128.

Marden, J.I., 1996. Analyzing and Modeling Rank Data. CRC Press.

Meredith, M., Kruschke, J., 2020. HDInterval: Highest (posterior) Density Intervals. R package version 0.2.

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.

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.

Miti, C., Lark, R.M., Milne, A.E., Giller, K.E., 2024a. BLA: Boundary line analysis. R package version 1.0.1.9000.

Miti, C., Milne, A., Giller, K., Lark, R., 2024b. The concepts and quantification of yield gap using boundary lines. A review. Field Crop. Res. 311, 109365.

Miti, C., Milne, A., Giller, K., Sadras, V., Lark, R., 2024c. Exploration of data for analysis using boundary line methodology. Comput. Electron. Agric. 219, 108794.

Mitscherlich, E., 1909. Des Gesetz des Minimums und das Gesetz des abnehmended Bodenertrages. Landwirsch Jahrb 3, 537–552.

Monserud, R.A., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. Ecol. Model. 62 (4), 275–293.

Myers, R.H., Myers, R.H., 1990. Classical and Modern Regression with Applications, vol. 2, Duxbury press Belmont, CA.

Nelder, J., 1961. The fitting of a generalization of the logistic curve. Biometrics 17 (1), 89–110.

Oliver, F., 1964. Methods of estimating the logistic growth function. J. R. Stat. Soc. Ser. C. Appl. Stat. 13 (2), 57–66.

PDA, 2017. Magnesium as a Nutrient for Crops and Grass. Potash Development Association, https://www.pda.org.uk/magnesium-nutrient-crops-grass/. (Accessed 29 April 2024).

Pruthviraj, N., Murali, K., Chaitanya, A., Harish, M., Karthik, A., 2024. Exploring the dynamics of nitrogen from conventional manures in the soil plant atmosphere continuum: A comprehensive review. Commun. Soil Sci. Plant Anal. 55 (11), 1690–1701.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., Ripley, M.B., 2013. Package 'MASS'. CRAN 538, 113–120.

Rousseeuw, P.J., Ruts, I., Tukey, J.W., 1999. The bagplot: a bivariate boxplot. Amer. Statist. 53 (4), 382–387.

Royall, R., 2017. Statistical Evidence: A Likelihood Paradigm. Routledge.

Sadras, V.O., 2020. On water-use efficiency, boundary functions, and yield gaps: French and Schultz insight and legacy. Crop. Sci. 60, 2187–2191.

Schils, R.L., van Voorn, G.A., Grassini, P., van Ittersum, M.K., 2022. Uncertainty is more than a number or colour: Involving experts in uncertainty assessments of yield gaps. Agricult. Sys. 195, 103311.

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

Schnug, E., Heym, J.M., Murphy, D.P.L., 1995. Boundary line determination technique (BOLIDES). In: Robert, P.C., Rust, R.H., Larson, W.E. (Eds.), Site-Specific Management for Agricultural Systems. Wiley Online Library, pp. 899–908.

Schut, A.G., Giller, K.E., 2020. Soil-based, field-specific fertilizer recommendations are a pipe-dream. Geoderma 380, 114680.

Shabani, A., Sepaskhah, A., Kamgar-Haghighi, A., Honar, T., 2018. Using double logistic equation to describe the growth of winter rapeseed. J. Agric. Sci. 156 (1), 37–45.

Shatar, T.M., McBratney, A.B., 2004. Boundary-line analysis of field-scale yield response to soil properties. J. Agric. Sci. 142, 553–560.

Smith, H.W., Ashworth, A.J., Nalley, L.L., Schmidt, A., Turmel, M.-S., Bucaro, A., Owens, P.R., 2024. Boundary line analysis and machine learning models to identify critical soil values for major crops in Guatemala. Agron. J. 116 (3), 1071–1087.

Tang, G., Shafer, S.L., Bartlein, P.J., Holman, J.O., 2009. Effects of experimental protocol on global vegetation model accuracy: a comparison of simulated and observed vegetation patterns for Asia. Ecol. Model. 220 (12), 1481–1491.

Tarkesh, M., Jetschke, G., 2012. Comparison of six correlative models in predictive vegetation mapping on a local scale. Environ. Ecol. Stat. 19 (3), 437–457.

Tukey, J.W., 1975. Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, vol. 2, Vancouver, pp. 523–531.

van der Ploeg, R.R., Böhm, W., Kirkham, M.B., 1999. On the origin of the theory of mineral nutrition of plants and the law of the minimum. Soil Sci. Am. J. 63 (5), 1055–1062.

Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer.

Walker, W.E., Harremoës, P., Rotmans, J., Van Der Sluijs, J.P., Van Asselt, M.B., Janssen, P., Krayer von Krauss, M.P., 2003. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. Integr. Assess. 4 (1), 5–17.

Wang, N., Jassogne, L., van Asten, P.J., Mukasa, D., Wanyama, I., Kagezi, G., Giller, K.E., 2015. Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. Eur. J. Agron. 63, 1–11.

Webb, R.A., 1972. Use of the boundary line in the analysis of biological data. J. Hortic. Sci. 47 (3), 309–319.

Yeo, I.-K., Johnson, R.A., 2000. A new family of power transformations to improve normality or symmetry. Biometrika 87 (4), 954–959.

Chapter 5

Extension of the Censored Normal Model to Categorical Independent Variables for Boundary Line Analysis

The chapter has been written in a Journal format and is ready for submission for publication as:

Miti. C, Milne. A.E, Giller. K.E, Lark. RM (Year). Extension of the Censored Normal Model to Categorical Independent Variable for Boundary Line Analysis.

5.1 Chapter 5 Overview

This chapter presents an extension of the censored normal boundary line fitting method to accommodate categorical independent variables. This is important because, in many agronomic studies, at least some of the independent variables are categorical (e.g., Fermont, Van Asten, et al. (2009); Hajjarpoor et al. (2018); Omondi et al. (2024)). Categorical factors can be either ordinal, where categories have a natural order or ranking such as soil fertility indices ranging from low to high or nominal, where categories represent distinct groups without inherent order such as soil classes, which reflect different soil types without an implied ranking. As the proposed censored model is based on maximum likelihood estimation, this study also evaluates the effect of data size on model fitting, particularly because many previous boundary line studies have relied on relatively small datasets. The standard deviation of measurement error (σ_{me}), a key input in the censored model, is typically estimated via likelihood profiling on a continuous variable. A central question addressed in this chapter is whether this estimate remains a useful guide when applying the model to categorical variables. To investigate this, continuous variables are discretised using both agronomic indices (e.g., RB209) and quantile-based binning, and the resulting effects on σ_{me} estimation are assessed. Using a nutrient–yield dataset, the chapter compares the performance and behaviour of categorical versus continuous boundary models, focusing on how bin size and classification method influence model outputs. The proposed method is illustrated using a dataset previously applied in yield gap analysis, demonstrating its relevance and applicability in real-world agronomic contexts. The findings offer practical guidance on when categorical boundary models are appropriate and how they can complement continuous models, particularly in data-limited or policy-focused settings. This work is intended for publication. I contributed to the study through conceptualization of the methodology, development of the `R` code, simulation design and data analysis, preparation of the original draft, and creation of the visualizations.

Extension of the Censored Normal Model to Categorical Independent Variables for Boundary Line Analysis

Miti, C.[a,b,*], Milne, A.E.[b], Giller, K.E.[c], and Lark, R.M.[a]

[a]School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, U.K.
[b]Net Zero and Resilient Farming, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, U.K
[c]Plant Production Systems, Wageningen University, P.O Box 430, 6700AK Wageningen, The Netherlands

*Correspondence: C.Miti. E-mail: chawezi.miti@nottingham.ac.uk

Abstract

The boundary line model is widely used in agronomy to identify yield-limiting factors. However, some important existing approaches such as the statistically robust censored normal model are limited to continuous covariates. We extend this model to accommodate categorical predictors, enabling the estimation of upper boundaries for discrete factor levels such as soil nutrient indices and management practices. To our knowledge, no previous study has implemented the censored normal boundary line model for categorical predictors in agronomic datasets. The model is fitted via maximum likelihood estimation, incorporating a measurement error term to account for observational uncertainty. However, the measurement error is often not available in many datasets but can be estimated using the likelihood profiling. Using a national-scale wheat dataset, we investigate how the choice of data categorisation method and sample size (n) influence the estimation of measurement error and overall model performance. Data were categorised using two approaches: the RB209 index and a quantile-based method which resulted in 5, 7, 10, and 12 binned datasets. To assess the impact of sample size, analyses were conducted on subsamples with n equal to 6110, 3055, 1222, 611, 305, 183, and 122. Results show that finer binning reduces estimated measurement error, while small sample sizes undermine the statistical support for boundary models. Comparison with continuous boundary line models confirms the agronomic plausibility of categorical boundaries derived from expert-based indices (e.g., RB209), unlike those based on arbitrary quantiles. A case study on soybean yield in Zambia further illustrates the model's utility in evaluating the limiting effects of categorical management

factors. Overall, this extension of the censored model offers a rigorous and interpretable tool for analyzing yield-limiting constraints in datasets where categorical variables dominate.

Keywords: boundary-line, categorical variables, measurement error, data size

1 Introduction

The boundary line model has been widely used to interpret biological datasets, particularly in agricultural and ecological studies where a dependent variable (e.g., crop yield or plant growth) is influenced by an independent factor with a limiting effect, such as soil fertility, water availability, or nutrient concentration (Smith et al., 2024). The observed data may exhibit a pattern where certain conditions (e.g., high nutrient availability) allow maximum biological performance, but other conditions impose constraints that prevent the response from reaching its full potential. The median regression models, which assume a continuous and symmetric relationship between predictor and response, often fail to capture biological constraints that impose upper limits on responses. This behaviour necessitates an approach that can formally describe the boundary function which is the upper envelope of the data that defines the maximum attainable response under given conditions (Webb, 1972).

A major challenge in estimating boundary line models is the proper statistical treatment of censoring, where observed responses are constrained by an upper boundary rather than following an unconstrained distribution (Milne et al., 2006b). Most approaches for boundary line estimation often rely on heuristic methods, such as selecting the upper percentiles of data within predefined bins (e.g. Casanova et al., 1999; Shatar and McBratney, 2004), which introduce arbitrary decisions regarding data inclusion and may fail to account for variability in the fitted boundary model (Lark and Milne, 2016). To overcome these limitations, Milne et al. (2006b) introduced a censored bivariate normal model, which provides a formal statistical framework for estimating boundary relationships. Their model assumes that data points arise from an underlying bivariate normal distribution, but values near the boundary are subject to right censoring, meaning that values exceeding the boundary threshold are less likely to be observed. To account for measurement imprecision and natural biological variability, the model incorporates a measurement error value in the response variable, ensuring that boundary estimates reflect the probabilistic nature of the observed data rather than being artificially constrained by arbitrary cut-off points. This approach provides a principled method for boundary estimation that is grounded in probability theory and avoids subjective data filtering.

While the censored bivariate normal model proposed by Milne et al. (2006a) effectively handles continuous independent variables, many biological and agricultural datasets involve categorical predictors that correspond to distinct groups, such as soil types, slope classification (steep, sloping, flat), soil depth (shallow, medium,

deep), and management practices (e.g., planting dates, number of weeding, pest and disease pressure) (Fermont et al., 2009; Hajjarpoor et al., 2018; Omondi et al., 2024). In these cases, each level of the categorical variable represents a different environmental condition, and the response variable may exhibit distinct means and boundary constraints across groups. A continuous boundary function as formulated by Milne et al. (2006b) is not directly applicable, as each categorical group requires an independent boundary estimate rather than a continuous curve.

In this study, we extend the censored normal framework to accommodate categorical independent variables by developing a factor-level boundary model. To our knowledge, no previous study has implemented the censored normal boundary model for categorical predictors in agronomic datasets. This approach generalizes the censored normal model by estimating a separate mean response and boundary constraint for each level of the categorical variable. The response distribution within each category is assumed to follow a censored normal distribution, where values exceeding the boundary constraint have a small likelihood of being observed. This formulation enables the estimation of boundary constraints across distinct categories while preserving the probabilistic structure of the original censored normal model. This modelling approach offers a flexible framework for analyzing biological responses constrained by factor-level limits. It is particularly relevant for applications in agronomy, ecology, and environmental science, where discrete environmental or management conditions often define the maximum achievable outcomes.

To fit the boundary model, the maximum likelihood estimation (MLE) is employed, which enables the simultaneous estimation of factor-specific means, boundary values, and residual variance. Unlike heuristic methods that depend on arbitrary decisions such as bin width, the MLE approach incorporates all available observations and explicitly accounts for censoring within the likelihood function. This provides a statistically rigorous alternative to traditional boundary estimation techniques, ensuring that the estimated boundaries reflect true upper constraints imposed by biological or environmental factors rather than subjective inclusion criteria. Given that MLE relies on the underlying data distribution to estimate model parameters accurately, a sufficiently large dataset is essential. However, most existing studies applying boundary line methodology, particularly those using heuristic methods, have relied on relatively small datasets, which may compromise model robustness and the validity of the estimated boundaries. It is acknowledged that laboratory measurements for some of the variables used in boundary line analysis can be costly to obtain, creating a trade-off between data collection costs and the level of accuracy required for the fitted model. This highlights a critical gap in current research on the boundary line methodology: determining the minimum number of data points required to fit boundary models to categorical variables with statistical confidence.

The standard deviation of measurement error (σ_{me}), as discussed earlier, is

an important parameter of the censored normal model. When σ_{me} is not available from experimental replicates, it can be estimated using geostatistical methods (Lark et al., 2020) with spatial data or, more generally, by likelihood profiling (Royall, 2017). Lark and Milne (2016), in a study investigating the effect of water-filled pore space on nitrous oxide emissions, found that the σ_{me} estimated via likelihood profiling (using water-filled pore space as a continuous variable) was consistent with that derived from duplicate measurements. In that case, likelihood profiling was applied to a continuous predictor. For continuous boundary models, the estimated parameters, including the measurement error standard deviation (σ_{me}), depend on how the data are distributed at the upper edges across the predictor's range. In contrast, when the predictor is categorical, the internal distribution within each category also influences the estimated boundary parameters. As a result, both the number and width of the categories can affect the value of σ_{me}. This raises the question: Is the σ_{me} estimated from a continuous variable a useful guide when fitting boundary line models to categorical data? One way to explore this is by discretising a continuous variable—such as converting soil nutrient concentrations into categorical fertility index scores—and then comparing the resulting boundary models. Specifically, assessing whether the σ_{me} estimated from a continuous model aligns with those obtained from categorical models based on different categorisation schemes. To our knowledge, no study has systematically evaluated how the choice of categorisation method affects the estimation of σ_{me}.

The aim of this study is to extend the censored bivariate normal model to accommodate categorical variables, evaluate how sample size affects model performance, investigate the influence of category width on σ_{me} estimates, compare agronomic interpretation of boundary line models fitted on continuous and categorical variables, and demonstrate the application of the extended model using real-world data.

2 Materials and Methods

2.1 Model structure development

To determine the response variable y under upper boundary constraints varying across levels of a categorical predictor x, we employed a censored normal model following the principles of the censored bivariate normal model presented by Milne et al. (2006b). Given that x is a categorical covariate with k levels, then the parameter set $\boldsymbol{\beta}$ can be written as:

$$\boldsymbol{\beta} = \{\bar{y}, \mu, \sigma\} \tag{1}$$

where $\bar{y}$ and μ are $k \times 1$ vectors and $\bar{y}[i]$ is the boundary value for the i^{th} level of x, $\mu[i]$ is the mean of the normal random variable for level i and σ is the standard

deviation of the normal random variable (uniform over all levels). If J is an $n \times k$ design matrix such that

$$J[l,i] = \begin{cases} 1 & \text{if observation } l \text{ out of } n \text{ corresponds to level } i \text{ of } x, \\ 0 & \text{otherwise,} \end{cases}$$

and if $j_l = J[l, 1{:}k]$ denotes the corresponding row vector for the l^{th} observation, then the boundary model is given by

$$b(x_l) = j_l^\top \bar{y}.$$

The probability density of an observation $\{x, y\}$, is assigned differently depending on their relationship to the censoring threshold, $b(x_l)$ as:

$$f(\bar{y}|x) = \begin{cases} f_N(y|j^\top \mu, \sigma) & \text{if } y < b(x_l), \\ \displaystyle\int_{b(x_l)}^{\infty} f_N(y|j^\top \mu, \sigma)\ dy & \text{if } y = b(x_l), \\ 0 & \text{if } y > b(x_l). \end{cases} \tag{2}$$

The likelihood can then be obtained from the convolution (*):

$$f(\bar{y}, x_l) = f(\bar{y}|x_l) * f_N(v|0, \sigma_{\text{me}}) \tag{3}$$

by incorporating the standard deviation of measurement error parameter σ_{me}. In practice, the parameters of the distribution are determined by minimizing the negative log-likelihood function:

$$\mathcal{L} = -\sum_{i=1}^{n} \log f(y_i|\mu_j, \sigma, b_j, \sigma_{\text{me}}), \tag{4}$$

where μ_j and b_j correspond to the estimated mean and upper bound for the j^{th} level associated with observation $\{x_i, y_i\}$. This is achieved using various numeric optimisation methods such the Nelder-Mead (Conn et al., 2009), Broyden–Fletcher–Goldfarb–Shanno (BFGS) (Nocedal and Wright, 1999), the Limited-memory Broyden–Fletcher–Goldfarb–Shanno with Box Constraints (L-BFGS-B) (Liu and Nocedal, 1989) algorithms among others commonly found in many statistical programming languages.

The standard errors of the parameter estimates b_j and μ_j are obtained from the inverse of the Hessian matrix, $\mathbf{H}$, of the log-likelihood function $\mathcal{L}(\boldsymbol{\theta})$. The Hessian matrix is given by:

$$\mathbf{H} = \left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\theta_i\, \partial\theta_j} \right].$$

The variance-covariance matrix of the maximum likelihood estimates is given as:

$$\mathbf{V} = \mathbf{H}^{-1},$$

and the standard error of the parameter estimate $\hat{\theta}_j$ (e.g., $\hat{b}_j$ or $\hat{\mu}_j$) is computed as:

$$\text{SE}(\hat{\theta}_j) = \sqrt{[\mathbf{H}^{-1}]_{jj}},$$

where $[\mathbf{H}^{-1}]_{jj}$ denotes the j^{th} diagonal element of the inverse Hessian matrix.

To assess model fit, the Akaike Information Criterion (AIC) is computed for the bounded and unbounded model. The AIC for the censored normal model (bounded) is given as:

$$\text{AIC}_{\text{bl}} = 2 \times (\mathcal{L} + 2n_j + 2) \tag{5}$$

while the standard normal unbounded model:

$$\text{AIC}_{\text{null}} = 2 \times (\mathcal{L} + n_j + 1) \tag{6}$$

where $\mathcal{L}$ represents the final negative log-likelihood value obtained from optimization and n_j is the number of categories. The term $2n_j + 2$ reflects the number of parameters in the bounded model, including category means, boundary values, standard deviation, and the standard deviation of measurement error. In contrast, the unbounded (null) model excludes the measurement error term and thus includes only $n_j + 1$ parameters. A smaller AIC value for the bounded model indicates that addition of the upper censor led to a better-fitting model. In addition, Akaike weights are computed to assess the relative support for each model given the data. For a set of candidate models, the Akaike weight quantifies the probability that a given model is the best among the set based on the AIC values. To determine the Akaike weight, the AIC differences (Δ_i) are first calculated by subtracting the minimum AIC value from each model's AIC. The relative likelihood of each model is then obtained as $\exp(-\Delta_i/2)$. Finally, the Akaike weights are derived by normalizing these relative likelihoods so that they sum to one across all models. These weights provide a convenient way to compare models and interpret model selection uncertainty (Burnham et al., 1998).

2.2 Preliminary data exploration

As the boundary line model captures the upper limit of the response variable, extreme values may misrepresent this constraint and bias the estimated censoring threshold. To address this, outliers were identified and removed within each categorical group for datasets used in this study. This was done using the boxplot method (Tukey, 1977), where a value x_i is flagged if it falls outside:

$$x_i < \text{Q}_1 - 1.5 \times \text{IQR} \quad \text{or} \quad x_i > \text{Q}_3 + 1.5 \times \text{IQR}, \tag{7}$$

where Q_1 and Q_3 are the 25^{th} and 75^{th} percentiles, and the interquartile range defined as:

$$IQR = Q_3 - Q_1. \tag{8}$$

Observations beyond these bounds were excluded to ensure the estimated boundary reflects the core data distribution. This procedure was applied separately to each categorical group to preserve within-group integrity.

2.3 Factor level width and the profile likelihood of the standard deviation of measurement error

To assess the effect of bin size on estimation of σ_{me}, a dataset (hereafter referred to as Dataset 1) compiled by AgSpace Agriculture Ltd in 2017 was used, which includes wheat yield and soil measurements (Mg, P, K) collected from management zones across farms in England (Lark et al., 2020). Soil sampling followed a standardized protocol involving 24 cores per management zone, composited into bulk samples for laboratory analysis. Nutrient extraction followed established methods: Olsen's method was used to extract P while K and Mg were extracted using 1M ammonium nitrate. The mean wheat yield per zone was recorded. For more information on the dataset see Lark et al. (2020).

As the data on the variables Mg, P and K were on a continuous scale, they could be converted into categorical data by specifying intervals on the range of values using 2 approaches, (1) the RB209 indices (AHDB, 2023) and (2) a quantile-based groupings which resulted in 5, 7, 10 and 12 binned datasets. The classification based on the RB209 index are presented in Table 1. The so-called index values provide the basis for agronomic advice to farmers in the UK. For the quantile-based groupings, the data were categorised into 5, 7, 10, or 12 groups of equal data size based on quantiles of the variable. For each case, the data were split into k groups using $k+1$ quantile-based cut points, which correspond to the sequence of quantile probabilities as:

$$\text{prob} = \{0, \tfrac{1}{k}, \tfrac{2}{k}, \ldots, \tfrac{k-1}{k}, 1\}.$$

This approach ensures that each group contains approximately the same number of observations, allowing a fair comparison of boundary estimates across different bin sizes. The categorisation was implemented in R using the `quantile()` and `cut()` functions, with the argument `include.lowest = TRUE` to retain boundary values.

The σ_{me} was estimated using log-likelihood profiling (Royall, 2017; Lark and Milne, 2016). This involved fixing σ_{me} at a series of values and, for each, estimating the remaining model parameters by maximum log-likelihood, conditional on the fixed σ_{me}. The resulting sequence of maximized log-likelihoods form a likelihood profile. The value of σ_{me} corresponding to the maximum of this profile was taken

Table 1: Classification of soil P, K and Mg analysis results into Indices

Index	Phosphorus (P) mg L^{-1}	Potassium (K) mg L^{-1}	Magnesium (Mg) mg L^{-1}
0	0–9	0–60	0–25
1	10–15	61–120	26–50
		121–180 (2-)	
2	16–25	181–240 (2+)	51–100
3	26–45	241–400	101–175
4	46–70	401–600	176–250
5	71–100	601–900	251–350

as its estimate. For each of the categorised data, the log-likelihood profile was determined for σ_{me} values ranging from 0-0.7 kg ha^{-1}. In each case, the σ_{me} value that maximised the log-likelihood value was considered the most plausible σ_{me} for that categorical dataset. In a similar way, the log-likelihood profiling was done on a continuous scale for Mg, P and K for σ_{me} values ranging from 0-0.7 kg ha^{-1} in order to compare the results. Prior to conducting the log-likelihood profiling on continuous scale, outliers for a combination of yield and each of the three independent variable were identified using two methods (a) the bagplot (Rousseeuw et al., 1999) and (b) the boxplot methods (Tukey, 1977), and these were removed from the data. The σ_{me} was estimated using the resulting two datasets.

2.4 Effect of data size on fitted boundary line model

The AgSpace dataset (described in section 2.3), comprising 6110 data points, was used to examine the effect of data size on the fitted boundary models. Smaller subsets were created by randomly selecting 50%, 20%, 10%, 5%, and 3% of the original data, corresponding to sample sizes of 3055, 1222, 611, 305 and 122, respectively. This was achieved using the `sample()` function in base `R`. For each of these datasets, the categorical boundary model for Mg, P and K was fitted. The resulting AIC_{bl} and AIC_{null} values for each model were used to check whether the boundary model was appropriate at each data size. The Akaike weights, ω_{bl} and ω_{null}, were used to quantify the appropriateness of the boundary model when no model was clearly superior based on the AIC values.

2.5 Comparison of categorical and continuous boundary model

The fitted categorical (RB209 and quantile-based) boundary models for Mg, P and K from the AgSpace dataset were compared to the boundary line model fitted on a continuous scale to check for consistency of the agronomic interpretation consistency. The continuous boundary line models were fitted following the procedure outlined by Lark and Milne (2016). Prior to fitting the continuous boundary line model, an exploratory data analysis was conducted to examine the summary statistics of the yield, Mg, P, and K using the `summastat()` function of the `BLA R` package.

This step aimed to assess the plausibility of the (censored) normality assumption. The assumption appeared reasonable for yield, while soil Mg, P, and K required log transformation to better satisfy normality. As done in section 2.3, outliers were identified using the bagplot and boxplot methods. The boundary line models were then fitted to the resulting two datasets using the `cbvn()` of the `BLA R` package. A `trapezium` boundary model was fitted for all the three variables based on the structure of the data at the upper edges. The trapezium model is given as:

$$y = \min(\beta_0,\ \beta_1 + \beta_2 \mathrm{x},\ \beta_3 + \beta_4 \mathrm{x}) \tag{9}$$

where β_0 is the plateau response, β_1 and β_3 are the y-intercepts for the two linear components, and β_2 and β_4 are the slopes with positive and negative values respectively. The initial start values for the optimization process for the boundary line models were determined using the `startValue()` function of the `BLA R` package with argument set to `model="trapezium"`. The σ_{me} value was determined by log-likelihood profiling as described in Section 2.3.

2.6 Case Study: Fitting the Censored Boundary Model to Categorical Variables

We illustrate the fitting of a boundary line model to categorical data using a censored normal model, based on a dataset (hereafter referred to as Dataset 2) collected by the International Institute for Tropical Agriculture (IITA) through crop cut surveys conducted in Malawi, Zambia, and Mozambique, as part of a study aimed at estimating soybean yield gaps, identifying their causes, and prioritizing interventions. For this illustration, we use only the data from Zambia, where 180 soybean-growing households in Katete and Sinda districts participated in the survey. Trained enumerators administered standardized questionnaires covering farmer information, field size, field slope, seed rate, management practices (e.g., weeding, pest and disease control), crop condition, and yield measurement, following the crop cut survey protocol described in Kindie et al. (2023).

The field area was measured using a handheld Global Positioning System device by tracing the field perimeter on foot and capturing geospatial coordinates. Disease and pest damage in soybean fields were assessed by enumerators using a four-point scale: 1 indicated no visible signs, 2 for limited signs, 3 for moderate, and 4 for widespread damage. Weed management was evaluated on a three-point scale, where 1 represented well-managed fields with no weed infestation, 2 for moderate management with some weed presence, and 3 for poorly managed fields with more than half the area infested. Crop condition was scored as follows: 1 denoted good crop stand across the field, 2 sparse stand, 3 patchy distribution, and 4 poor overall performance. Field slope was categorized into five classes based on gradient: 1 for flat ($< 4\%$), 2 for gentle (4–9%), 3 for moderate (10–15%), 4 for steep (16–30%),

and 5 for extremely steep ($> 30\%$).

Soybean yields were measured using standardized crop cut protocols. Three $2\,\text{m} \times 2\,\text{m}$ subplots were laid out diagonally across the field, and all plants within each subplot were harvested. Grains were threshed, dried to constant weight under ambient conditions, and weighed. Dry grain weights (DGW) were adjusted to 12% moisture content using the formula:

$$\text{DGW} = \text{Field grain weight} \times \frac{100 - \text{moisture content (\%)}}{88}. \tag{10}$$

Yields were then averaged across the three subplots to obtain the final yield estimate per field. For full details on crop cut surveys for yield gap decomposition see Kindie et al. (2023).

The categorical boundary model was fitted to yield as a function of management practices (disease damage, pest damage, and weed management), crop condition, and field slope. For each variable, outliers were identified for each level separately and removed from the dataset. The censored model was fitted following the procedure described in section 2.1. The σ_{me} was estimated using the log-likelihood profiling approach described in section 2.3 using the continuous variable seed-rate.

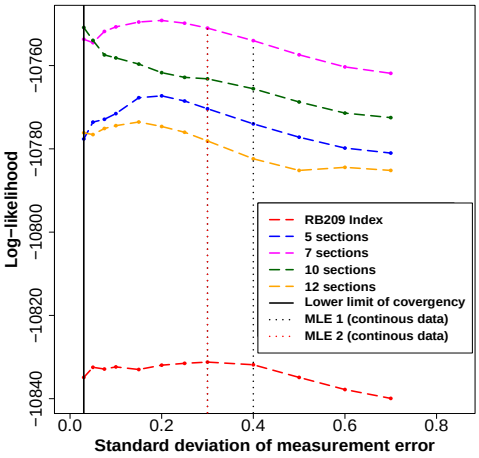
3 Results

3.1 Standard deviation of measurement error determination

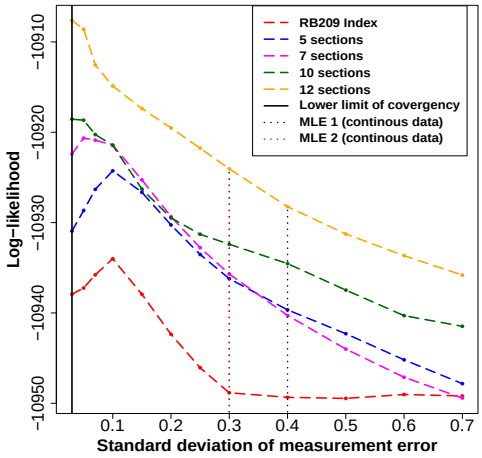
The σ_{me} for Mg, P, and K in Dataset 1 was estimated using different data categorisation approaches. Figure 1 presents the log-likelihood profiles of the estimated σ_{me} based on two grouping methods, the RB209 index and four quantile-based groupings which resulted in 5, 7, 10 and 12 binned datasets. For Mg, the log-likelihood was maximized at 0.3 (RB209), 0.2 (5 sections), 0.2 (7), < 0.02 (10), and 0.15 (12). For P, the corresponding values were 0.1 (RB209), 0.1 (5), 0.05 (7), 0.05 (10), and $<$ 0.02 (12). For K, the maxima were < 0.02 (RB209), 0.05 (5), 0.02 (7), < 0.02 (10), and < 0.02 (12). In general, the estimated σ_{me} decreased as the number of sections increased (i.e reducing bin sizes). For the continuous variables Mg, P, and K, the σ_{me} was consistently estimated at 0.4 t ha^{-1} when outliers were identified using the bagplot method. When outliers were identified using the boxplot method, σ_{me} was maximized at 0.3 t ha^{-1}. Attempts to estimate values below 0.02 failed to converge, indicating a lower limit for reliable optimization.

3.2 Effects of data size boundary line fitting

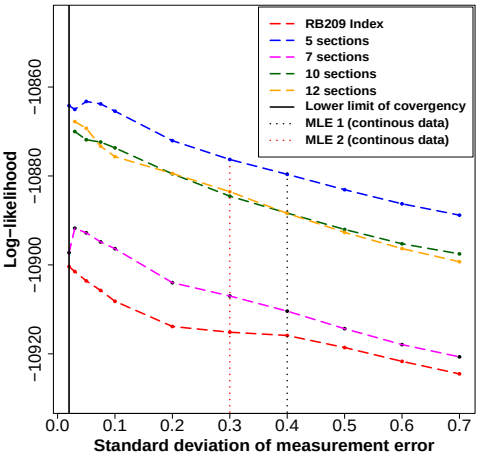
Figure 2 and Appendices A.1 and A.2 present the fitted categorical boundary models for Mg, P, and K in Dataset 1, respectively, applied to datasets of varying sizes 6110, 3055, 1222, 611, 305, 183, and 122. Table 3.2 presents the AIC values for both the fitted boundary (AIC_{bl}) and the unbounded (AIC_{null}) models across varying data



(a)



(b)



(c)

Figure. 1: Log-likelihood profiles of measurement error values for (a) Magnesium, (b) Phosphorus, and (c) Potassium, based on different approaches for dividing the data into categories. MLE 1 and MLE 2 refer to the measurement error estimates obtained from data where outliers were identified using the bagplot and boxplot methods, respectively.

sizes for Mg, P, and K. The corresponding Akaike weights are also included. For Mg, AIC_{null} exceeds AIC_{bl} when the data size is 611 or greater. For both P and K, AIC_{null} exceeded AIC_{bl} in all cases, with the corresponding Akaike weights for AIC_{bl} exceeding 0.97. However, for smaller data sizes (183 and 611), the Akaike weights dropped to 0.74 and 0.87 for P and K, respectively.

Table 2: AIC values for bounded (AIC_{bl}) and unbounded (AIC_{null}) models fitted to datasets of varying sizes

Boundary model	Dataset size	Λ_{bl}	Λ_{null}	ω_{bl}
Magnesium-Yield	6110	**22393.34**	22442.64	1.00
Magnesium-Yield	3055	**11200.50**	11225.12	0.99
Magnesium-Yield	1222	**4327.31**	4335.82	0.98
Magnesium-Yield	611	**2075.61**	2075.83	0.53
Magnesium-Yield	305	1039.78	**1031.91**	0.02
Magnesium-Yield	183	631.02	**622.89**	0.02
Magnesium-Yield	122	408.22	**404.47**	0.13
Phosphorus-Yield	6110	**22148.15**	22212.81	1.00
Phosphorus-Yield	3055	**10939.34**	10959.76	0.99
Phosphorus-Yield	1222	**4432.43**	4443.63	0.99
Phosphorus-Yield	611	**2146.75**	2162.17	0.99
Phosphorus-Yield	305	**1019.33**	1026.42	0.97
Phosphorus-Yield	183	**654.69**	656.77	0.74
Phosphorus-Yield	122	**360.08**	369.12	0.99
Potassium-Yield	6110	**22087.54**	22154.58	1.00
Potassium-Yield	3055	**11014.84**	11034.17	0.99
Potassium-Yield	1222	**4381.74**	4402.00	0.99
Potassium-Yield	611	**2125.93**	2129.54	0.86
Potassium-Yield	305	**999.57**	1010.92	0.99
Potassium-Yield	183	**600.85**	611.68	0.99
Potassium-Yield	122	**387.64**	394.80	0.97

Λ_{bl} and Λ_{null} are AIC values for the censored and uncensored normal models respectively; ω_{bl} is the Aikaike weight for the censored normal models.
Bold values of Λ_{bl} and Λ_{null} indicate the smaller of the two values.

3.3 Comparison of categorical and continuous boundary model

The Figures 3, 4 and 5 presents the continuous boundary line models for Mg, P and K superimposed on the categorical boundary models For Dataset 1. The parameters of the continuous boundary models for Mg, P and K are presented in Tables 3 and 4, based on data in which outliers were identified using the bagplot and boxplot methods, respectively. The categories of the Mg, P and K are based on the RB209 index and equal-sized data sections (5, 7, 10 and 12). The categorical boundary points derived from RB209 index categories closely mirror the structure of the continuous boundary models for Mg, P, and K. Consistent with agronomic expectations, yield increases with nutrient concentrations in indices 0 and 1, reaching a plateau at index 2 for Mg and P, and at index 2+ for K. Applying alternative approaches to cate-

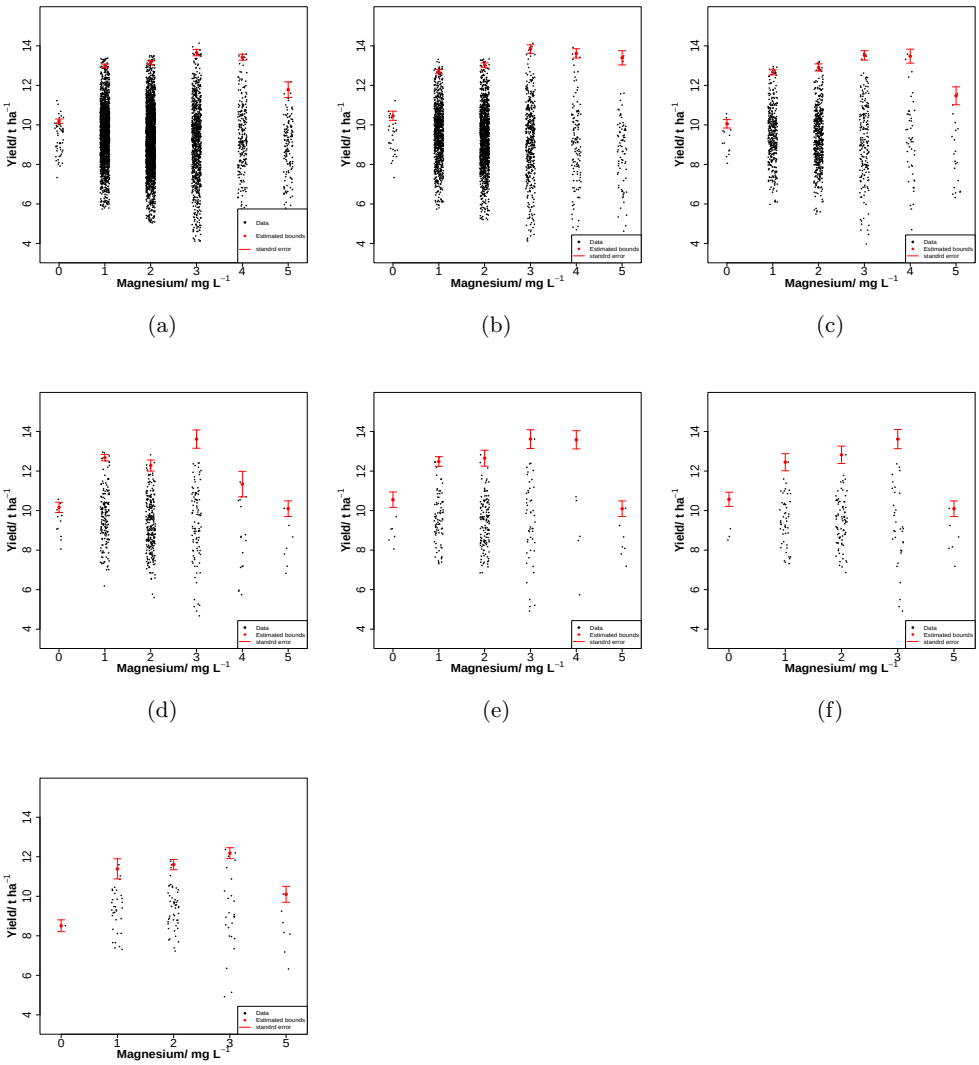


Figure. 2: Categorical boundary models fitted to Magnesium (Mg) data (based on the RB209 index) using varying sample sizes: (a) 6110, (b) 3055, (c) 1222, (d) 611, (e) 305, (f) 183, and (g) 122 observations.

gorising the data resulted in boundary points that did not align with the boundary line fitted on a continuous scale. In all cases, the boundary points were identical across categories, resulting in a flat upper boundary.

Table 3: Parameters of fitted continuous boundary line models for Mg, P and K with standard error of parameters given in brackets. Outlier removal using the bagplot method

Parameter	Phosphorus (P)	Magnesium (Mg)	Potassium (K)
Model	trapezium	trapezium	trapezium
β_1	5.75 (0.88)	−15.33 (2.61)	−21.88 (0.04)
β_2	2.82 (0.37)	8.13 (0.84)	8.12 (0.16)
β_0	13.43 (0.09)	13.20 (0.19)	13.35 (0.19)
β_3	82.57 (12.66)	40.62 (1.90)	46.23 (7.03)
β_4	−15.77 (2.71)	−5.15 (0.31)	−5.67 (1.16)
Λ_{bl}	**32376.46**	**33145.73**	**28464.52**
Λ_{null}	32429.55	33189.39	28476.13
ω_{bl}	1.0	1.0	0.99
ω_{null}	0.0	0.0	0.01

Λ_{bl} and Λ_{null}, AIC values for the censored and uncensored bivariate normal models; ω, Aikaike weight for the fitted model.
Bold values of Λ_{bl} and Λ_{null} indicate the smaller of the two values.

Table 4: Parameters of fitted continuous boundary line models for Mg, P and K with standard error of parameters given in brackets. Outlier removal using the boxplot method

Parameter	Phosphorus (P)	Magnesium (Mg)	Potassium (K)
Model	trapezium	trapezium	trapezium
β_1	3.61(0.86)	−20.87 (2.76)	−7.54 (2.84)
β_2	3.80 (0.41)	10.12 (0.92)	4.72 (0.71)
β_0	13.24 (0.08)	13.35 (0.02)	13.35 (0.09)
β_3	31.77 (10.59)	28.95 (0.01)	32.58 (0.00)
β_4	−4.49 (2.28)	−2.85(0.02)	−3.05 (0.03)
Λ_{bl}	**30980.18**	**32557.91**	**27815.31**
Λ_{null}	31021.72	32582.17	27832.25
ω_{bl}	1.0	0.99	0.99
ω_{null}	0.0	0.01	0.01

Λ_{bl} and Λ_{null}, AIC values for the censored and uncensored bivariate normal models; ω, Aikaike weight for the fitted model.
Bold values of Λ_{bl} and Λ_{null} indicate the smaller of the two values.

3.4 Case Study Results: Boundary Estimates for Categorical Variables

This section presents the results for boundary line models fitted to variables in Dataset 2. Using log-likelihood profiling on the continuous seed-rate variable, the σ_{me} was estimated to be 0.13 t ha^{-1}. This was used as an input in fitting the categorical boundary models for Disease damage, pest damage, weed management,

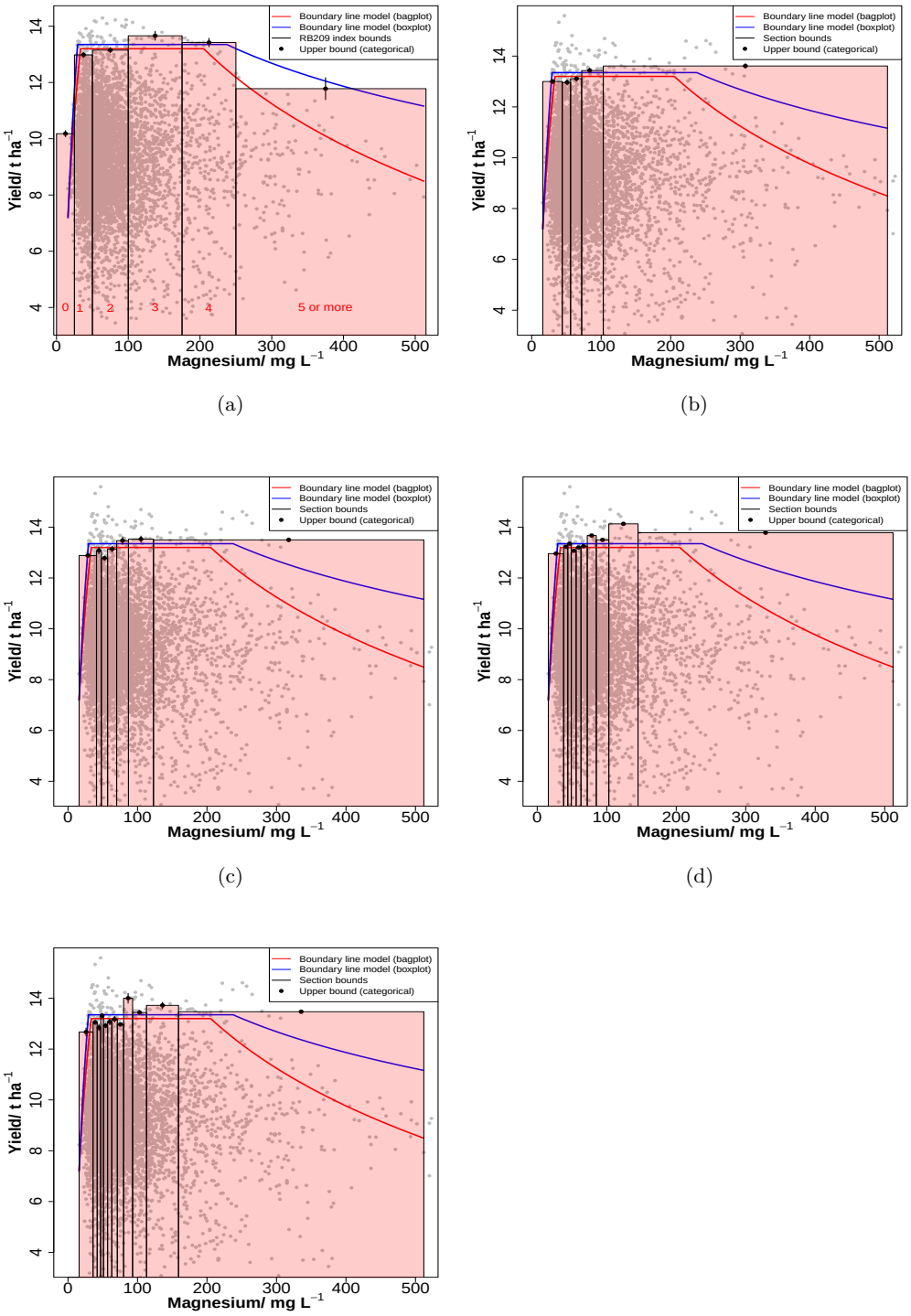


Figure. 3: Continuous boundary line models for Magnesium (Mg) as a function of yield, compared to boundaries determined using categorical classifications. Mg categories were defined based on: (a) RB209 indices, and equal-sized data sections of (b) 5, (c) 7, (d) 10, and (e) 12 categories. The continuous boundary models were fitted after outlier removal using the bagplot (red) and boxplot (blue) methods.

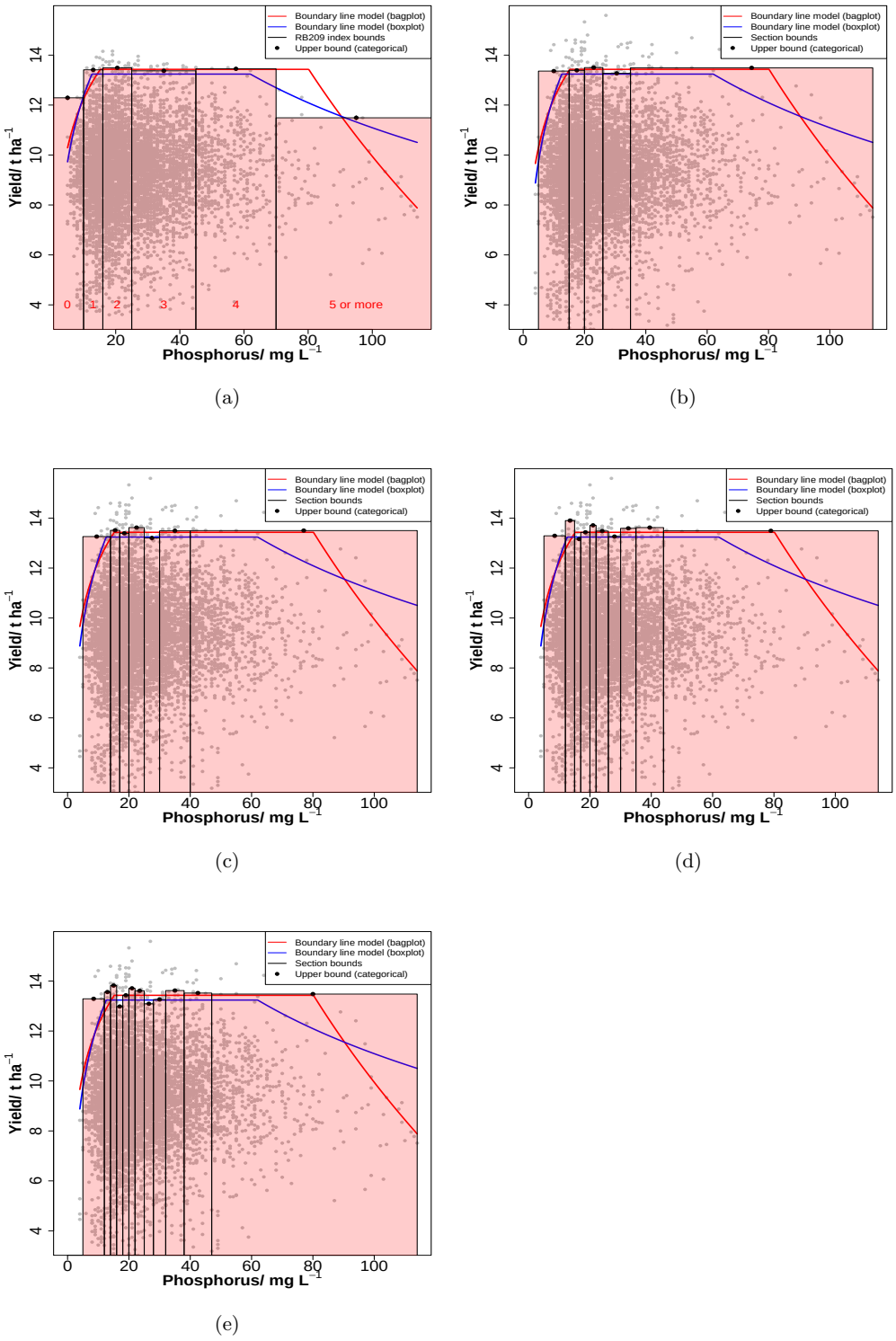


Figure. 4: Continuous boundary line models for Phosphorus (P) as a function of yield, compared to boundaries determined using categorical classifications. P categories were defined based on: (a) RB209 indices, and equal-sized data sections of (b) 5, (c) 7, (d) 10, and (e) 12 categories. The continuous boundary models were fitted after outlier removal using the bagplot (red) and boxplot (blue) methods.

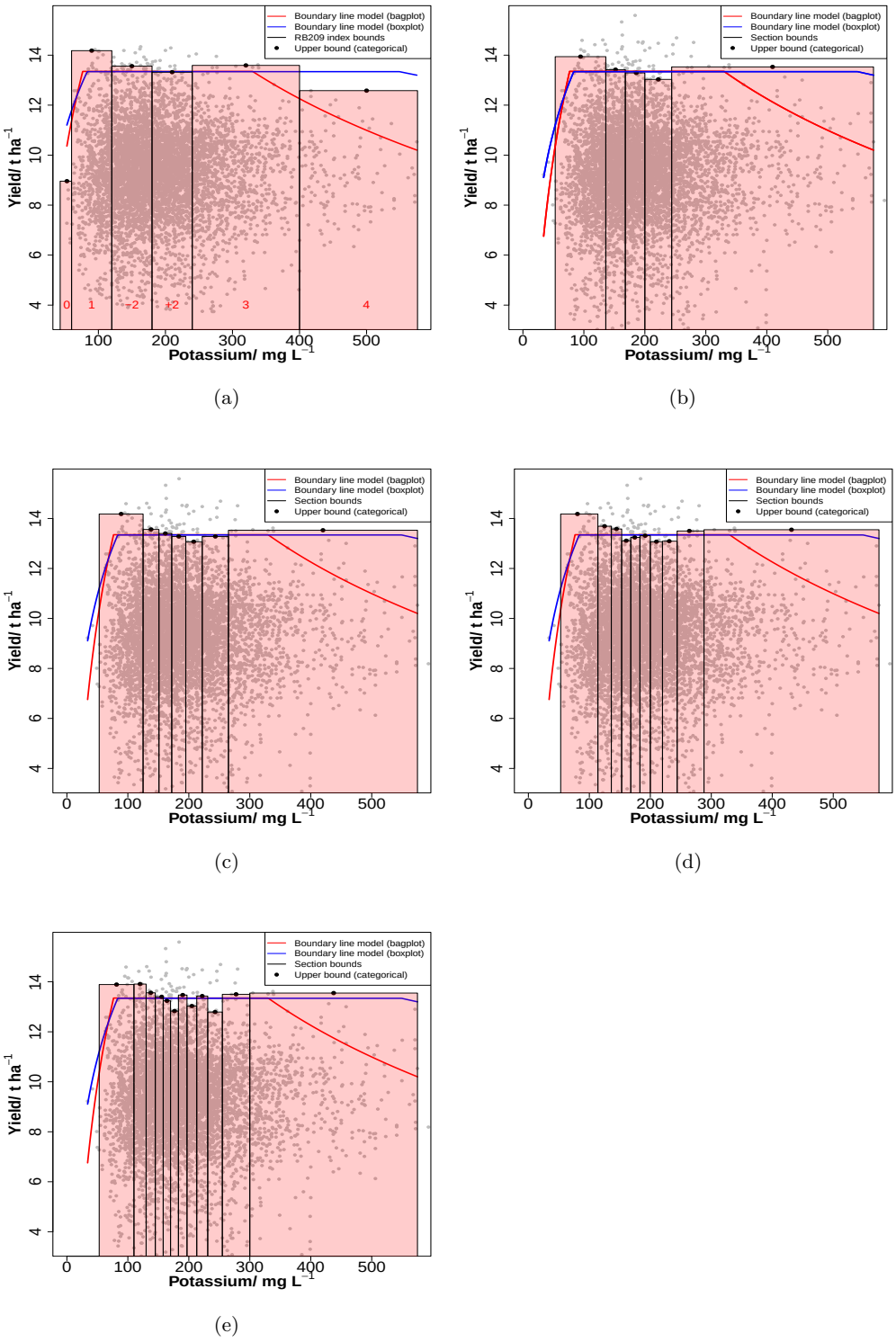


Figure. 5: Continuous boundary line models for Potassium (K) as a function of yield, compared to boundaries determined using categorical classifications. K categories were defined based on: (a) RB209 indices, and equal-sized data sections of (b) 5, (c) 7, (d) 10, and (e) 12 categories. The continuous boundary models were fitted after outlier removal using the bagplot (red) and boxplot (blue) methods.

Table 5: Boundary estimates for categories of the various factors and corresponding AIC and Akaike weight values. Standard error given in brackets.

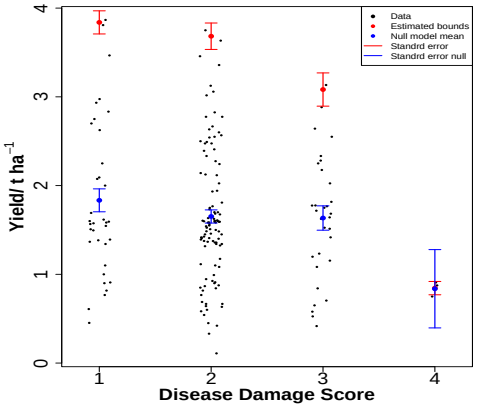
Factor	Boundary value /Levels				Λ_{bl}	Λ_{null}	ω_{bl}
	1	2	3	4			
Disease damage	3.84(0.13)	3.68(0.14)	3.08(0.19)	0.84(0.08)	**408.87**	412.46	0.86
Pest damage	3.35(0.18)	3.86(0.10)	2.57(0.10)	2.11(0.15)	409.61	**407.11**	0.22
Weeding mgt	4.45(0.18)	3.77(0.10)	1.58(0.06)		**417.65**	423.23	0.94
Crop condition	3.42(0.13)	3.58(0.17)	2.50(0.19)	2.05(0.18)	389.62	**384.75**	0.08
Field slope	2.58(0.12)	2.59(0.07)	3.98(0.17)	2.95(0.12)	381.25	**380.53**	0.41

Λ_{bl} and Λ_{null}, AIC values for the censored and uncensored normal models; ω, Aikaike weight for the fitted model.
Bold values of Λ_{bl} and Λ_{null} indicate the smaller of the two values.

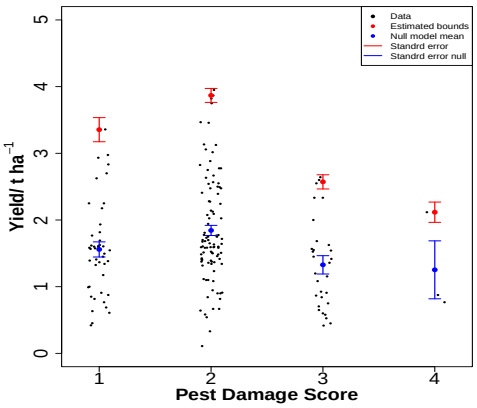
crop condition and field slope. Figure 6 presents the fitted boundary models, and Table 5 provides the corresponding boundary estimates for each category along with the AIC and Akaike weights for Dataset 2.

The boundary estimates align with agronomic expectations for most factors. For disease damage, a clear decreasing trend in boundary yield was observed with increasing severity from 3.84 t ha^{-1} in category 1 (no visible damage) to just 0.84 t ha^{-1} in category 4 (severe damage). The boundary model showed strong support, with an AIC of 408.87 compared to 412.46 for the null model, and an Akaike weight of 0.86, indicating an 86% probability that the boundary model provides a better explanation of the data. A similar trend was observed for crop condition, where the boundary yield declined from 3.58 to 2.05 t ha^{-1} across worsening categories. However, in this case, the null model had better support ($AIC_{null} = 384.75$ vs. AIC_{bl} =389.62), and the Akaike weight for the boundary model was only 0.08, suggesting very weak support for the inclusion of a boundary. For pest damage, the trend was less consistent, with the largest boundary yield observed in category 2 (3.86 t ha^{-1}) rather than category 1. The boundary model had a larger AIC (409.61) than the null (407.11), and a low Akaike weight of 0.22, indicating limited evidence in favour of the boundary model.

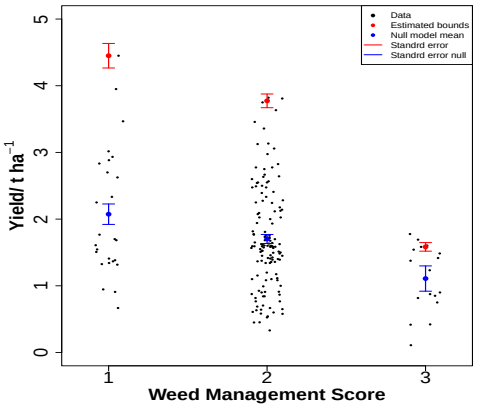
In the case of weeding management, boundary yields decreased sharply from 4.45 t ha^{-1} in category 1 to 1.58 t ha^{-1} in category 3. The boundary model received strong support, with an AIC of 417.65 compared to 423.23 for the null model and a corresponding Akaike weight of 0.94, reflecting a high likelihood that the boundary model is the better fit. For field slope, no clear trend was observed. The boundary values remained relatively flat (around 2.58–2.95 t ha^{-1}) across most categories, with a larger value of 3.98 t ha^{-1} in category 3. The AIC values for the boundary and null models were close (381.25 and 380.53), and the Akaike weights (0.41 and 0.59, respectively) indicated nearly equal support for both models, suggesting weak evidence in either direction.



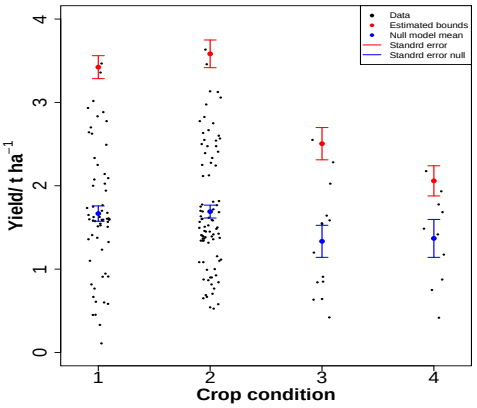
(a)



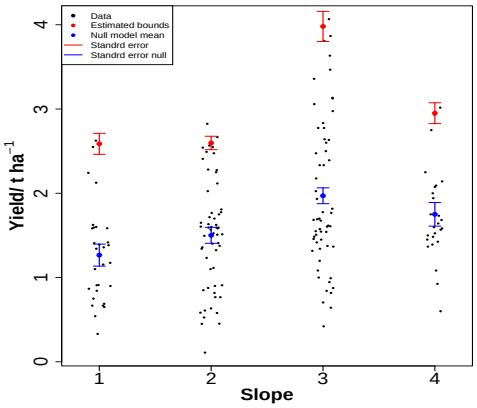
(b)



(c)



(d)



(e)

Figure. 6: Fitted categorical boundary models for (a) Disease damage score, (b) pest damage score, (c) weed management score, (d) crop condition and (e) field slope

4 Discussion

This study demonstrated the feasibility of extending the censored bivariate normal model to categorical independent variables. Unlike continuous models that assume smooth transitions across predictor values, the categorical model estimates separate upper boundaries for each category without requiring assumptions about equal spacing or linear progression. This approach is particularly valuable for agronomic and ecological datasets, where independent variables often represent discrete management practices, environmental classifications, or ordinal indicators that do not conform to continuous scales. By allowing flexible upper boundary estimation within these defined groups, the model accommodates the structure and constraints inherent in real-world field data.

In addition, the framework provides a statistical basis for fitting boundary line models, enabling the evaluation of model performance and goodness of fit. We assessed each fitted boundary model against a corresponding null model without boundaries (an unbounded model) using the AIC and Akaike weights. These comparisons are essential in statistical modelling because they allow for the quantification of model support while penalizing for model complexity, thereby balancing goodness of fit with parsimony. This helps avoid over-fitting and ensures that improvements in model performance are not simply due to increased flexibility (Burnham and Anderson, 2004). In this context, demonstrating that the boundary model outperforms the null model provides empirical justification for the inclusion of boundary constraints, supporting the relevance and utility of the boundary line approach in explaining yield-limiting/-reducing factors.

Most previous studies that applied the boundary line model to categorical data have relied on heuristic fitting approaches, typically by estimating boundary points for each category and subsequently fitting a continuous boundary model to these points (e.g., Omondi et al., 2024; Hajjarpoor et al., 2018; Wang et al., 2015; Fermont et al., 2009). When the categorical variable is ordinal or represents progressive changes such that the assumption of equal spacing between categories is reasonable, fitting a continuous model to the estimated boundary points may yield additional insights. This approach facilitates interpretation and enables prediction of boundary yields for intermediate or missing levels in the data. In cases where the categories are nominal, the imposition of a continuous structure may not be appropriate. However, a logical or expert-defined ordering can be employed to ensure that any inference drawn from the boundary estimates remains valid and meaningful.

The estimation of the σ_{me} using log-likelihood profiling revealed that finer binning (i.e., more categories) tended to reduce the estimated σ_{me}. This finding aligns with expectations that larger bins incorporate more within-group variability, which adds on to measurement error estimate, inflating its value. In contrast, finer categorisation better captures homogeneity within bins, reducing the σ_{me} estimate. While categorisation simplifies complex data structures, it risks conflating measure-

ment error with intra-group variability, particularly if the number of groups is too small or unevenly spaced. This can also explain why the σ_{me} value determined using continuous data was always larger than that determined from categorical data as it uses the whole data as one group to determine σ_{me}. Nevertheless, Lark and Milne (2016) found that the σ_{me} determined from the log-likelihood profiling on a continuous scale was similar to that determined from repeated measurements and therefore, in the absence of replicated measurements, σ_{me} can be approximated from profiling on continuous covariates where available. The σ_{me} values estimated from the two datasets resulting from different outlier detection methods were similar, at 0.3 and 0.4 t ha^{-1}, respectively.

Our investigation into the effect of data size confirmed that larger datasets significantly enhance the stability and reliability of parameter estimates in the censored normal model. As sample size decreased, the standard errors associated with boundary estimates increased substantially (as shown by increase in error bars), reflecting heightened uncertainty and reduced confidence in the fitted model. This sensitivity to data size is a well-recognized limitation of maximum likelihood estimation, which relies on sufficient information in the data to produce precise and consistent parameter estimates (Harrell, 2015). This exercises however, enabled an assessment of the minimum data size required for stable model estimation.

Applying the AIC allowed us to systematically assess how sample size affects model performance. The bounded (censored) model was consistently preferred over the null (unbounded) model when the sample size exceeded 600 observations for Mg, and required even fewer observations for P and K. In contrast, at very small sample sizes (e.g., 122 or 183 observations), the unbounded model often outperformed the boundary model for Mg. Even in some cases where the boundary model was favoured for P and K, the Akaike weights were below 0.86, indicating a diminished probability of it being the best model among the candidates. A simulation study by Miti et al. (2024) found that a minimum of 400 observations is required to reliably detect the presence of a boundary in a dataset on a continuous scale, when such a boundary truly exists.

These findings highlight a critical implication for modelling with the categorical censored model. Adequate sample size is essential to obtain robust estimates and to avoid model misspecification (Burnham et al., 1998). This was evident for Dataset 2 (n = 180), where for three out of five variables, the unbounded model was preferred based on smaller AIC values. In such cases, interpreting results through the lens of boundary line theory may be misleading, potentially overstating the presence or strength of limiting factors. Importantly, the censored model framework offers a safeguard by making such limitations explicit through likelihood-based metrics, encouraging caution in interpreting results from smaller datasets. This property enhances the model's utility in practice, as it helps avoid overconfidence in boundary-driven inferences when data support is insufficient.

The comparison of the categorical boundary model with a continuous boundary line model revealed that RB209-based categories aligned well with the fitted continuous boundary models, reinforcing their agronomic relevance. The boundary line models fitted to data with outliers identified using both the bagplot and boxplot methods were similar across all variables, except at larger values where data were sparse. Under optimal growing conditions, increasing the concentration of Mg, P, or K in the soil is expected to result in larger yields at RB209 indices 0 and 1. At index 2, the crop response begins to diminish, and by index 3, yield gains typically plateau. In index 4, excessive nutrient levels may lead to nutrient imbalances, potentially reducing yield (PDA, 2017). This pattern was reflected by both the continuous boundary model and the categorical boundary model based on the RB209 classification. In contrast, arbitrary quantile-based groupings often failed to reproduce the underlying boundary structure and produced flat upper bounds, which are inconsistent with the expected agronomic response. This result highlights the importance of selecting categorisation schemes grounded in domain knowledge. However, it is necessary that there is considerable amounts of data in each category to avoid bias. When categorisation is arbitrary or driven purely by statistical quantiles, the resulting boundary model may obscure meaningful biological patterns. Agronomically informed indices, such as RB209, provide more interpretable and reliable boundaries that reflect true biological limits.

The categorical boundary models fitted to the management practice index in Dataset 2 underscored the need to account for critical factors when interpreting boundary line model results. Although the fitted models captured agronomically consistent trends such as lower yields associated with higher disease and pest incidence, poor crop conditions, and inadequate weed control, there were not sufficient data to support the boundary line model, as shown by the larger AIC values than for the unbounded null model, in most cases. Thus, while the data may be useful for identifying general yield trends, using the upper bounds to estimate attainable yields may not be appropriate. This highlights a key advantage of the censored model, which allows for formal model assessment using likelihood-based criteria, an element often absent in heuristic approaches commonly found in the literature. Most previous studies have relied on such heuristic methods without statistical evaluation of model fit, which may lead to misleading conclusions and inappropriate agronomic recommendations. Additional care should be taken when interpreting the results of the boundary line model, since the boundary constraints must be understood in relation to the specific meaning of each categorical variable. For instance, a category defined by increased weeding frequency may indicate either improved management or, conversely, higher weed pressure serving as a proxy for an unmeasured constraint. Therefore, the categorical model should not be interpreted solely based on numerical outputs; instead, it must be evaluated alongside agronomic expertise and contextual understanding.

5 Conclusion and Recommendation

In this study, we introduced a statistical framework for fitting boundary line models to categorical data, addressing a key gap in the application of boundary line methodology when explanatory variables are not measured continuously. Our results demonstrate that while the categorical approach provides a practical solution in contexts where data are aggregated or discretised (e.g., through agronomic guidelines like RB209), it is inherently more sensitive to the choice of categorisation scheme. Specifically, the estimation of the standard deviation of measurement error (σ_{me}) is strongly influenced by how the data are binned. We also show that the robustness of categorical models declines with smaller sample sizes, highlighting the need for cautious interpretation in data-limited scenarios. Overall, this extension enhances the flexibility of boundary line analysis and expands its applicability to a broader range of agronomic datasets. Moreover, the results obtained when the data was transformed from continuos to categorical for dataset 1, the results were similar.

6 Acknowledgements

This work was supported by the Nottingham-Rothamsted Future Food Beacon Studentships in International Agricultural Development. Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom (BBSRC). We acknowledge support from the Growing Health Institute Strategic Programme [BB/X010953/1; BBS/E/RH/230003C].

References

AHDB (2023). *Nutrient Management Guide (RB209). Section 1: Arable Crops*. Agriculture and Horticulture Development Board, Coventry, UK.

Burnham, K. and Anderson, D. (2004). Multimodel inference, understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33:261–304.

Burnham, K. P., Anderson, D. R., Burnham, K. P., and Anderson, D. R. (1998). *Practical use of the information-theoretic approach*. Springer.

Casanova, D., Goudriaan, J., Bouma, J., and Epema, G. (1999). Yield gap analysis in relation to soil properties in direct-seeded flooded rice. *Geoderma*, 91(3-4):191–216.

Conn, A. R., Scheinberg, K., and Vicente, L. N. (2009). *Introduction to derivative-free optimization*. SIAM.

Fermont, A. M., Van Asten, P. J., Tittonell, P., Van Wijk, M. T., and Giller, K. E. (2009). Closing the cassava yield gap: An analysis from smallholder farms in East Africa. *Field Crops Research*, 112(1):24–36.

Hajjarpoor, A., Soltani, A., Zeinali, E., Kashiri, H., Aynehband, A., and Vadez, V. (2018). Using boundary line analysis to assess the on-farm crop yield gap of wheat. *Field Crops Research*, 225:64–73.

Harrell, F. E. (2015). *Regression Modeling Strategies*. Springer Series in Statistics. Springer, New York, 2 edition.

Kindie, T., Tesfaye, S., Lulseged, T., Gizaw, D., and Wuletawu, A. (2023). Crop cut surveys yield gap decomposition add-on. *Excell. Agron. Initiat. CGIAR*.

Lark, R. M., Gillingham, V., Langton, D., and Marchant, B. P. (2020). Boundary line models for soil nutrient concentrations and wheat yield in national-scale datasets. *European Journal of Soil Science*, 71:334–351.

Lark, R. M. and Milne, A. E. (2016). Boundary line analysis of the effect of water-filled pore space on nitrous oxide emission from cores of arable soil. *European Journal of Soil Science*, 67:148–159.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528.

Milne, A. E., Ferguson, R. B., and Lark, R. M. (2006a). Estimating a boundary line model for a biological response by maximum likelihood. *Annals of Applied Biology*, 149:223–234.

Milne, A. E., Wheeler, H. C., and Lark, R. M. (2006b). On testing biological data for the presence of a boundary. *Annals of Applied Biology*, 149:213–222.

Miti, C., Milne, A., Giller, K., Sadras, V., and Lark, R. (2024). Exploration of data for analysis using boundary line methodology. *Computers and Electronics in Agriculture*, 219:108794.

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*. Springer.

Omondi, J. O., Chiduwa, M. S., Kyei-Boahen, S., Masikati, P., and Nyagumbo, I. (2024). Yield gap decomposition: Quantifying factors limiting soybean yield in Southern Africa. *npj Sustainable Agriculture*, 2(1):32.

PDA (2017). Magnesium as a nutrient for crops and grass. Potash Development Association. https://www.pda.org.uk/magnesium-nutrient-crops-grass/. Accessed: 29/04/2024.

Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387.

Royall, R. (2017). *Statistical Evidence: A Likelihood Paradigm*. Routledge.

Shatar, T. M. and McBratney, A. B. (2004). Boundary-line analysis of field-scale yield response to soil properties. *Journal of Agricultural Science*, 142:553–560.

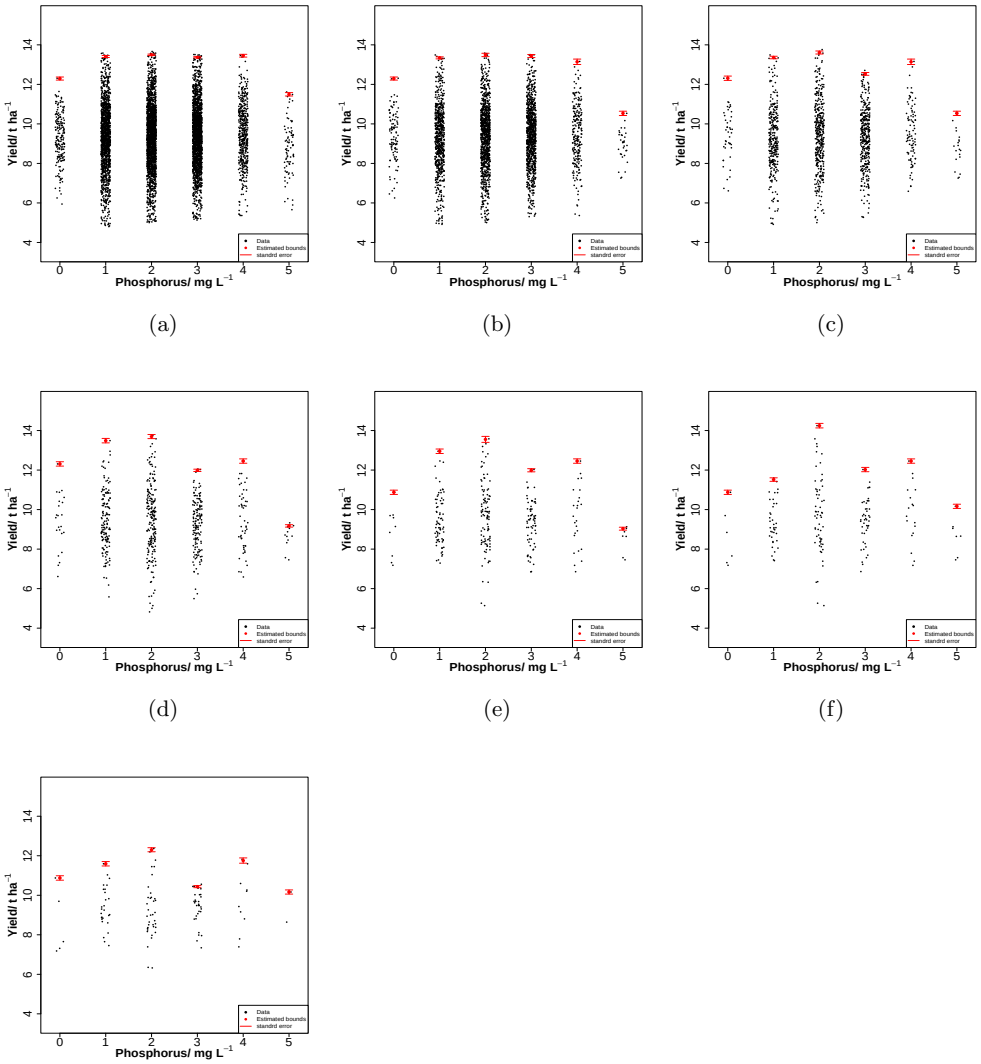
Smith, H. W., Ashworth, A. J., Nalley, L. L., Schmidt, A., Turmel, M.-S., Bucaro, A., and Owens, P. R. (2024). Boundary line analysis and machine learning models to identify critical soil values for major crops in Guatemala. *Agronomy Journal*, 116(3):1071–1087.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Wang, N., Jassogne, L., van Asten, P. J., Mukasa, D., Wanyama, I., Kagezi, G., and Giller, K. E. (2015). Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. *European Journal of Agronomy*, 63:1–11.

Webb, R. A. (1972). Use of the boundary line in the analysis of biological data. *Journal of Horticultural Science*, 47(3):309–319.

A Appendix 1 : Categorical boundary models for Phosphorus and Potassium

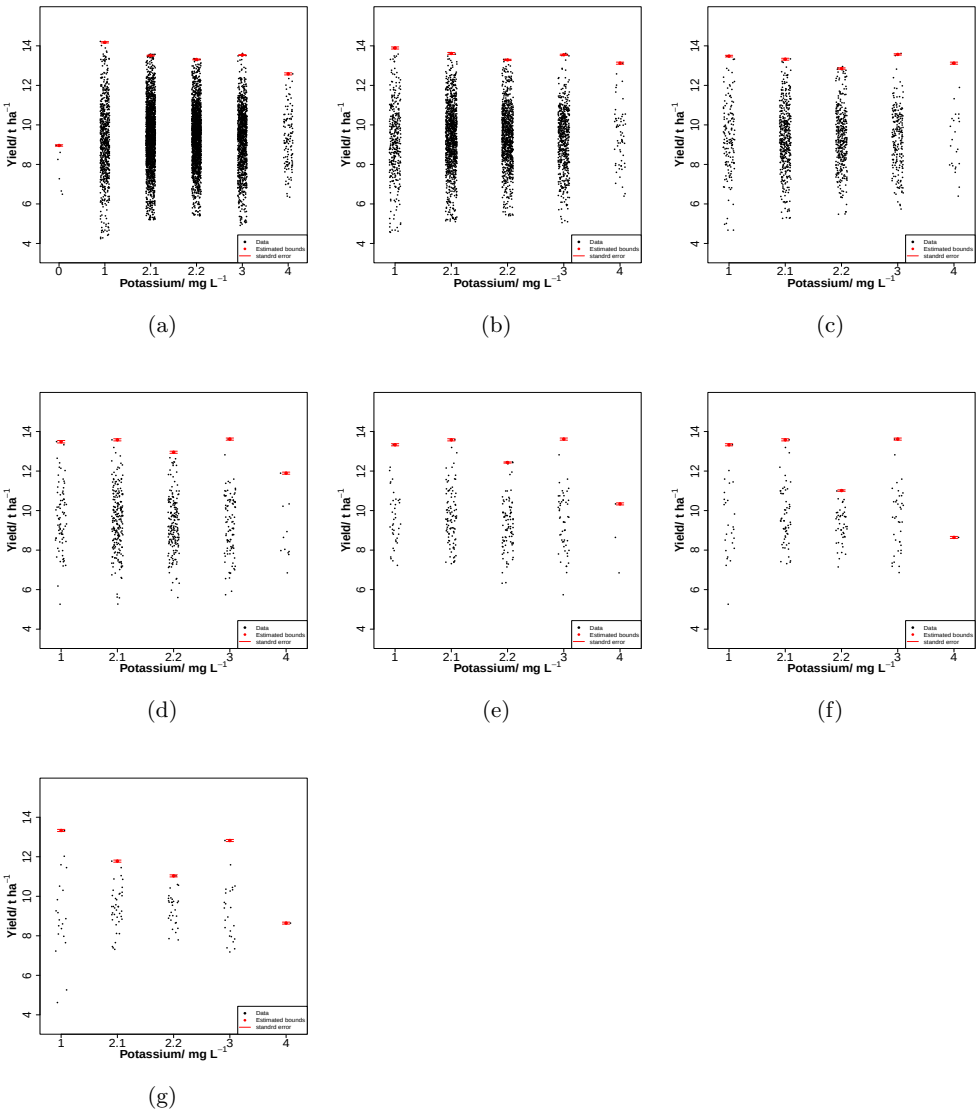


(a) (b) (c)

(d) (e) (f)

(g)

Figure. A.1: Categorical boundary models fitted to Phosphorus (P) data (based on the RB209 index) using varying sample sizes: (a) 6110, (b) 3055, (c) 1222, (d) 611, (e) 305, (f) 183, and (g) 122 observations.



(a) (b) (c)

(d) (e) (f)

(g)

Figure. A.2: Categorical boundary models fitted to Potassium (K) data (based on the RB209 index) using varying sample sizes: (a) 6110, (b) 3055, (c) 1222, (d) 611, (e) 305, (f) 183, and (g) 122 observations.

Chapter 6

BLA: An R Package for Boundary Line Analysis

This chapter is based on an `R` package for carrying out boundary line analysis, `BLA`, published on CRAN as:

Miti, C., Lark, R. M., Milne, A. E., & Giller, K. E. (2024). `BLA`: Boundary Line Analysis [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/BLA/index.html (R package version 1.0.1).

A manuscript has been developed and is ready for submission to the *Journal of Statistical Software* as:

Miti, C., Lark, R. M., Milne, A. E., & Giller, K. E. (2025).BLA: An R Package for Carrying Out Boundary Line Analysis for Biological Responses

6.1 Chapter 6 Overview

This chapter introduces the `BLA R` package, the first open-source software tool specifically designed to implement boundary line analysis (BLA) for assessing biological responses. The package consolidates both heuristic and statistical methods for boundary line fitting, including tools for exploratory data analysis, model selection, parameter estimation, and interpretation of results. Key features include functions to test for boundary structure in data, assist with parameter initialization, and identify most-limiting factors using the Law of the Minimum. The chapter demonstrates the application of the `BLA R` package using a large-scale dataset of wheat yield and soil properties from England, showcasing its ability to derive critical nutrient thresholds and identify yield-limiting factors. The `BLA R` package has been published on the Comprehensive `R` Archive Network (CRAN) platform and can be installed directly in R software. This chapter is due for submission to the ***Journal of Statistical Software*** for publication. I contributed to the development of the `BLA R` package by review of available boundary line fitting methods, developing data exploratory methods, coding the methods into `R` functions, compiling the functions into an `R` package, submitting the package to CRAN and maintaining the package. For this manuscript, I contributed to data analysis, visualization of results and manuscript preparation.

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

BLA: An R Package for Carrying Out Boundary Line Analysis for Biological Responses

Chawezi Miti
University of Nottingham

Alice E. Milne
Rothamsted Research

Richard. M. Lark
University of Nottingham

Abstract

The boundary line has been proposed as a model to describe a biological response as a function of a factor(s) of interest in uncontrolled environments. Under this approach, the upper boundary of the plot of some biological response variable (e.g. crop yield) against a potentially limiting factor represents the limiting response to that factor. This approach has been applied in many biological studies. Various boundary line fitting methods are available which include the bolides algorithm, binning procedure, quantile regression, Markowski quantile regression and the censored bivariate normal model among others. Exploratory methods have also been developed to test the strength of evidence of bounded behaviour in a dataset as well as methods to determine initial starting values for fitting boundary line models. A challenge for researchers with limited programming and statistical capacity is that there is no specialised open-source package dedicated to boundary line analysis. This makes these methods inaccessible and hinders reproducibility of research. We have developed an R package called **BLA** that brings together the principal methods of boundary line analysis, and the new approaches to data exploration and inference. Here we present the theoretical basis and structure of the different functions available in the **BLA** package. We illustrate the use of the package by applying it to a dataset from England which comprises wheat yield and soil properties, P and pH, collected by AgSpace Agriculture Ltd. The boundary line methodology was applied to this data to extract information on the possible limitations to increased productivity and provide target values for soil properties to avoid limitations on yield. The results indicate that the critical soil P concentration is 15.29 mg l^{-1} with a 95% confidence interval of 13.46 – 17.79 mg l^{-1} while that of pH is 6.06 with a confidence interval of 5.93 – 6.26. Both these values are consistent guidance values used for decision-making on P fertilizer and soil liming in England and Wales. Soil P was found to be the most-limiting factor on 14.82% of the sampled points while soil pH only accounted for less than 1.89%. As only two factors were evaluated, the yield-limiting factor was unidentified for most observations (83.29%).

Keywords: R package, boundary line, numeric optimization, most-limiting factor.

1. Introduction: Boundary line analysis

Research into how biological systems respond to different factors is essential for understanding how to improve their efficiency. This research is often based on designed experiments where the response to the presence or absence of one or more factors of interest is observed in controlled conditions. However, controlled experiments are not always possible, and biological response data may be measured in uncontrolled environments. Theories to model the relationship of biological response as a function of a factor(s) of interest in such settings have been developed. One such theory is the boundary line model proposed by Webb (1972). This assumes that when a biological response, y, is observed as function of a single factor, x, in uncontrolled environment i.e. when the influence of factors other than x is not controlled, a scatter of randomly distributed points results (e.g. conforming to a bivariate normal distribution). If the dataset includes sufficient points where the factors other than x are not limiting then the upper boundary of the data cloud describes the relationship between x as a limiting factor and the response variable (Smith, Ashworth, Nalley, Schmidt, Turmel, Bucaro, and Owens 2024). This function, the maximum response of y given x, is what Webb (1972) called the boundary line. Any data points below this maximum response result from the limiting effect of other, uncontrolled, factors.

The boundary line model has been widely applied in different biological studies including plant physiology (e.g. Houshmandfar, Fitzgerald, Armstrong, Macabuhay, and Tausz 2015; Smith and Hardie 2023), microbial processes in soil (e.g. Elliott and de Jong 1993; Lark and Milne 2016; Schmidt, Thöni, and Kaupenjohann 2000), dendroecology (e.g. Black and Abrams 2003), yield gap analysis (e.g. Andrade, Comin, Moura-Bueno, and Brunetto 2023; Casanova, Goudriaan, Bouma, and Epema 1999; Fermont, van Asten, Tittonell, van Wijk, and Giller 2009; Hajjarpoor, Soltani, Zeinali, Kashiri, Aynehband, and Vadez 2018; Shatar and McBratney 2004; Smith *et al.* 2024; van Vugt and Franke 2018; Wairegi, van Asten, Tenywa, and Bekunda 2010), crop water-use efficiency (e.g. Baral, Bhandari, Kumar, and Min 2022; Chambers, Hinckley, Cox, Metcalf, and Aslin 1985; Fink, Grassini, Rocateli, Bastos, Kastens, Ryan, Lin, Patrignani, and Lollato 2022; French and Schultz 1984; Sadras 2020) and crop modelling (e.g. Montaner, Maddonni, and DiNapoli 1997) among other studies to evaluate the influence of various factors on the biological responses.

The original boundary line method of Webb (1972) involves the visual selection of data points at the upper edges of the data cloud. The boundary line model is then fitted to the selected data points by least squares. Improvements to the boundary line fitting methodology have been suggested over the years. Some of the notable methods include binning (Casanova *et al.* 1999) and the boundary line determination (bolides) algorithm (Schnug, Heym, and Murphy 1995) which select boundary points according to a repeatable heuristic rule, quantile regression methods(Andrade *et al.* 2023), the censored bivariate normal model (Milne, Ferguson, and Lark 2006a; Lark and Milne 2016) which is estimated by maximum likelihood, and the Makowski quantile regression (Makowski, Doré, and Monod 2007) which uses the quantile regression procedure in a Bayesian setting. In addition, data exploratory methods have been developed for testing for evidence of bounded variation in a dataset to justify the use of boundary line analysis on data (Miti, Milne, Giller, Sadras, and Lark 2024c; Milne, Wheeler, and Lark 2006b). The concept behind these exploratory methods is that if a data set exhibits a boundary in which values of y are constrained by an upper limit conditional on x, it is expected that the upper edges of x/y plot would be more densely occupied than is the fringe

of a random bivariate distribution. The clustering arises because the response can not go beyond the maximum possible biological limit (boundary line). Such a distribution supports the biological interpretation of a boundary model.

Most studies which apply boundary line analysis now use computational methods rather than the selection of boundary points 'by eye'. However, an open source library of code for boundary line analysis is lacking. A single package with a range of boundary line analysis functions would offer users an easily accessible tool for reproducible complete analysis from data exploration to boundary line post-hoc analysis especially those with limited statistical and programming capabilities.

The R platform (R Core Team 2024), an open-source program for statistical analysis, is gaining popularity over the past decade because of its high flexibility and accessibility for statistical analysis. The development of an R library dedicated to boundary line analysis has the potential to enhance reproducible application/research of boundary line analysis. To this effect, we have developed and published an R package on the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/web/packages/) called **BLA** (boundary line analysis) (Miti, Lark, Milne, and Giller 2024a) that contains a series of functions supporting boundary line analysis. It includes functions to identify outliers, check for distribution characteristics of variables, test evidence of boundary behaviour in a dataset, determine initial start values for model optimization, fit the boundary line model and do post-hoc analysis. Here, we present the theoretical background of the boundary line analysis procedures, the structure of **BLA** functions and illustrate the application of the **BLA** R package on a real dataset. The current version of **BLA** is freely available as an R package from the CRAN (https://cran.r-project.org/web/packages/BLA/index.html) and the user manual and vignettes can be found at https://chawezimiti.github.io/BLA/.

2. Theory

2.1. Data exploration

There are three condition that need to be met for an interpretable boundary line analysis to be implemented. First, the dataset used should be large enough to cover a wide range of possible conditions (Smith *et al.* 2024). In terms of Lark, Gillingham, Langton, and Marchant (2020) the data set must *exhibit* the boundary. In addition, when the censored bivariate normal model is used, the variables x and y should plausibly be drawn from an underlying bivariate normal distribution (perhaps after transformation) albeit with some censoring of the y variable including circumstances where the response of interest y, is limited by the factor of interest, x, and where other factors are limiting. The assumption of normality can be checked by accessing summary statistics indices like the mean, median, skewness and octile skewness. The octile skewness index is a measure of skewness more robust than the standard skewness coefficient based on moments, as it minimizes the influence of outliers by using stable percentile values to measure distribution asymmetry (Brys, Hubert, and Struyf 2003). As a rule of thumb transformation is unlikely to be needed if octile skewness is in [-0.2, 0.2]. Additionally, summary plots including histograms, box-plots and qq-plots can be used for visual interpretation to decide whether a transformation is required.

The second condition is that the data set should not contain bivariate outliers i.e. data

points that deviate significantly from the overall distribution. This is important as fitting of the boundary line is heavily influenced by points at the edges of the scatter. There are various repeatable methods to identify bivariate outliers (Smiti 2020). The simplest method is based on the Mahalanobis distance which is the distance between a point and the mean of the distribution. This is compared to a chi-squared distribution with degrees of freedom equal to the number of variables. Data points with a Mahalanobis distance exceeding a critical value (from the chi-squared table) are considered outliers. More robust methods like the bagplot, based on half space location depth (i.e. the smallest number of data points contained in any closed halfplane with the line passing through a bivariate point) (Rousseeuw, Ruts, and Tukey 1999), and methods based on isolation forest (Liu, Ting, and Zhou 2012) or k-nearest neighbours (D'Orazio 2023) are also used. These identify outliers considering the structure of the data scatter. For consistency, the bagplot method is recommended for operations using the **BLA** package as it has used within the `expl_boundary()` function to check for outliers in the simulation step when checking for evidence of boundary structure (see below).

The third condition implicit in any interpretation, is that data points at the upper edges of a data cloud represent a biological limit for a given value of x. Miti *et al.* (2024c) proposed a method based on convex-hull peels to assess this assumption. If there is an upper bound on the biological response to x then it is expected that observations will show some clustering in the initial peels. This is determined by checking how close points in the upper edges (peels) of the data are to each other compared to a realization of a bivariate normal random variable. For a data cloud of y against x, data points in n consecutive peels are identified as the upper points of n convex hulls (Skiena 1998) with peeling. This is achieved using the steps below as described by Miti *et al.* (2024c).

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

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

$\mathbf{v}_l$ denotes the vertex such that:

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

$\mathbf{v}_l$ is 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 $l \leq i \leq m$.

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 $\mathbf{v}_i \in \hat{V}$,

$$\mathbf{v}_i \in \hat{V}_\mathrm{l} \iff x_i \leq \dot{x}, \tag{6}$$

$$\in \hat{V}_\mathrm{r} \text{ otherwise.} \tag{7}$$

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 dataset, then our analysis was based on the combined subsets of sections from the first ten peels:

$$\hat{\mathcal{V}}_l = \bigcup_{m=1,\ldots 10} \hat{V}_\mathrm{l}^m. \tag{8}$$

and

$$\hat{\mathcal{V}}_r = \bigcup_{m=1,\ldots 10} \hat{V}_\mathrm{r}^m. \tag{9}$$

To determine evidence of the presence of a boundary limiting response, the dispersion of the vertices in the two sections of the data cloud is evaluated and compared with the same statistic for a bivariate normal random variate of the same size with the same parameters simulated j number of times. This is achieved using the standard deviation (sd) of the Euclidean distances between vertices in the left and right subsets and the centroid of the full dataset, $\mathbf{m} = [\bar{x}, \bar{y}]^\mathrm{T}$, where $\bar{x}$ and $\bar{y}$ are the average values of the x and y variables over all observations respectively. The Euclidean distances between a vertex $\mathbf{v}_i$ and the centroid is given by

$$d_i = \left\{ (\mathbf{v}_i - \mathbf{m})^\mathrm{T} (\mathbf{v}_i - \mathbf{m}) \right\}^{\frac{1}{2}}. \tag{10}$$

The sd of the values d_i for all $\mathbf{v}_i \in \hat{\mathcal{V}}_l$ and $\mathbf{v}_i \in \hat{\mathcal{V}}_r$ can be calculated. A smaller sd than the corresponding sd for the peels of a bivariate normal random variate with same distribution properties indicates evidence of a greater concentration of vertices in the upper bound (left or right section) of the dataset. The strength of this evidence is determined by Monte Carlo method to obtain a distribution of the sd for the case of the multivariate normal null distribution (Mecklin and Mundfrom 2005) which is simulated m number of times. The p-value is determined as the probability of obtaining a value of sd which provides evidence as strong or stronger against the multivariate normal null hypothesis, if it were true, as the evidence from our actual data.

2.2. Boundary line models

There is a variety of models that can be used to describe biological responses. Table 1 shows the equations for several commonly-used models. For example, the logistic model described by Nelder (1961) has been widely used for boundary lines in agronomic studies (Fermont *et al.* 2009; Ndabamenye, Asten, Blomme, Vanlauwe, Uzayisenga, Annandale, and Barnard 2013; Kintché, Hauser, Mahungu, Ndonda, Lukombo, Nhamo, Uzokwe, Yomeni, Ngamitshara, Ekoko, Mbala, Akem, Pypers, Matungulu, Kehbila, and Vanlauwe 2017; Wairegi *et al.* 2010; Wang, Jassogne, van Asten, Mukasa, Wanyama, Kagezi, and Giller 2015). For the equations described in Table 1, β_0 represents the asymptote or the potential maximum of the response variable, β_2 is the biological constant, β_1 is the parameter governing the rate at which the

response variable approaches its potential maximum except for the Schmidt model were it represents the x value when y is maximum, and β_3 is the allometric constant. The models in Table 1 are all built-in the **BLA** package. However, **BLA** allows the user to select custom functions in addition to the above.

Table 1: Biological response model to describe the boundary line

Model	Equation	**BLA** code	Reference
Linear	$y = \beta_1 + \beta_2 x$	`"blm"`	Dobson and Barnett (2018)
Linear-plateau	$y = \min(\beta_0, \beta_1 + \beta_2 \mathrm{x})$	`"lp"`	Makowski *et al.* (2007)
Logistic	$y = \frac{\beta_0}{1+\exp(\beta_2(\beta_1 - x))}$	`"logistic"`	Hosmer (2000)
Logistic	$y = \frac{\beta_0}{1+\beta_2 \exp(-\beta_1 x)}$	`"logisticND"`	Nelder (1961)
Inverse logistic	$y = \beta_0 - \frac{\beta_0}{1+\exp(\beta_2(\beta_1 - x))}$	`"inv-logistic`	Helidoniotis, Haddon, Tuck, and Tarbath (2011)
Double logistic	$y = \frac{\beta_0}{1+\exp(\beta_2(\beta_1 - x))} - \frac{\beta_0}{1+\exp(\beta_4(\beta_3 - x))}$	`"double-logistic"`	Shabani, Sepaskhah, Kamgar-Haghighi, and Honar (2018)
Mitscherlic	$y = \beta_0 - (\beta_0 - \beta_2)\exp(-\beta_1 x)$	`"mit"`	Mitscherlich (1909)
Schmidt	$y = \beta_0 - \beta_1(x - \beta_2)^2$	`"schmidt"`	Schmidt *et al.* (2000)
Custom	-	`"other"`	-

2.3. Boundary points selection and model fitting methods

The **BLA** package offers several heuristic and statistical methods for fitting a boundary line (Miti, Milne, Giller, and Lark 2024b). These are described in detail below.

Heuristic boundary fitting methods

The heuristic methods for fitting boundary lines comprise the visual, binning, bolides and quantile regression. The visual method is the simplest method for fitting boundary line. Boundary points are simply identified by eye and parameters of the fitted line are obtained by the least square method. This is not repeatable, and cannot be used to make statistical inferences. We recommend that it is used only to choose initial values of boundary line parameters when using other methods.

The binning method encompasses various approaches that divide the range of the independent variable, x, into equal intervals. The intervals can be created using either a static window with a fixed width or a moving, overlapping window, as described by Shatar and McBratney (2004). In each interval, a boundary response value is selected based on a set criterion, such as the 90^{th}, 95^{th}, 99^{th} or 99.7^{th} percentile (Casanova *et al.* 1999; Kintché *et al.* 2017; Schmidt *et al.* 2000; van Vugt and Franke 2018). The chosen percentile should not be too high (like the 100^{th} percentile) to avoid the influence of outlier boundary points, nor too low, to ensure that larger response values are not excluded (Schmidt *et al.* 2000). The boundary line is then fitted to the boundary response variable.

For bolides, boundary points to which a boundary model is fitted are selected by a stepwise procedure (Schnug *et al.* 1995). First, the minimum (x_{min}), maximum (x_{max}), and the maximum response point (x_{ymax}) values for the factor, x_i, in the dataset are identified. Starting with the largest response value at x_{min}, the next boundary data point is the data point along the x-axis with a larger response value than the previously selected boundary point. If multiple points exist, the largest is selected. Further boundary data points following the x-axis are identified until the boundary point equal to x_{ymax} is reached. The boundary points between the x_{max} and x_{ymax} are identified in a similar way starting from x_{max} but moving in the opposite direction.

The least squares fit is used to find the best-fitting curve to a set of boundary data points identified using the bolides and binning methods by minimizing the sum of squares of the deviations (residuals) of the boundary data points to the curve (Dobson and Barnett 2018). Given a set of data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, the least squares fit is determined by finding the parameters of a function $f(x)$ that minimizes the sum of the squares of the differences between the observed y_i and the predicted values, $f(x_i)$, for each x_i. Mathematically, for a function of the form:

$$y_i = f(x_i|\boldsymbol{\beta}) + \epsilon, \tag{11}$$

where $\boldsymbol{\beta}$ is a vector of parameters $\beta_1, \beta_2, ..., \beta_m$ of the basis function $f(x)$ and ϵ is the error (noise), the least squares fit aims to minimize the sum of squares of the residual function:

$$S = \sum \epsilon_i^2 = \sum [y_i - f(x_i|\boldsymbol{\beta})]^2. \tag{12}$$

The optimal estimates $\hat{\boldsymbol{\beta}}$ are found by differentiating S with respect to each parameter β_j of $\boldsymbol{\beta}$ and solving the simultaneous equation:

$$\frac{\partial S}{\partial \beta_j} = 0, \text{ for } j = 1, 2, ..., m. \tag{13}$$

These equations can be solved using numerical minimization algorithms approaches namely the first-order derivative (e.g. gradient descent, steepest descent) and second-order derivative methods (e.g. Newton methods). The first-order derivative methods rely on following the derivative (or gradient) downhill/uphill to find the optimal solution for the objective function while the second-order derivative methods are based on the derivative of the derivative stored in matrix known as the Hessian matrix, **H**. The **H** is a square matrix of second-order partial derivatives of a scalar-valued function. Given a scalar-valued function $f(x)$ of n variables $x_1, x_2, ..., x_n$, **H** is defined as:

$$\mathbf{H}_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}, \tag{14}$$

The $(i, j)^{\text{th}}$ entry of the **H** is the second partial derivative of f with respect to x_i and x_j. The **H** therefore captures the curvature of a multi-variable function and provides information about the local behavior of the function near a critical point. In optimization algorithms like Newton's method, **H** is used to determine the direction and step size for updating the current solution.

The second-order derivative methods like the Newton's method are more efficient in estimating the minima of the objective functions because the **H** gives both the direction towards the minima as well as the required step size. However, the limitation of Newton's method is that it requires that **H** is available and is a positive definite. The determination of **H** is a computationally expensive operation and may not be stable depending on the properties of the objective function. The solution to these challenges is the use of Quasi-Newton methods which are second-order optimization algorithms that approximate the **H** using the gradient, hence no need to calculate it precisely for each step of the algorithm. The Quasi-Newton algorithm can be written as a quadratic approximation as:

$$g(x) = \frac{1}{2}(x - q)^T \mathbf{B}(x - q) + (x - q)^T \nabla f(q) + f(q), \tag{15}$$

where $g(x)$ is quadratic approximation of the objective function at the point x near q using information from the function at q, $f(q)$ is the value of the function at point q, $(x - q)$ represents the displacement vector from the point q to point x and **B** is the approximation of **H** determined with the constraints that (1) it is symmetric, (2) the quadratic approximation at the new point, $\nabla g(x)$, has the same gradient as the last point, $\nabla f(q)$, and (3) **B** at new point is close to that at the previous point (Boyd and Vandenberghe 2004).

Some of the commonly used second-order optimization algorithms for numerical optimization which are available in the **BLA** package are the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm and its extension the Limited Memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm. Other optimization methods like the Nelder–Mead method also known as downhill simplex or amoeba, a heuristic search method often applied to optimization problems for which derivatives may not be known are also available. In the Nelder–Mead method the **H** can be approximated using information from the calculated curvature of the surface in the neighbourhood of the minimum using the values at the midpoints of the edges of the final simplex.

Numerical optimization algorithms to obtain parameter estimates require starting values of the model parameters which can be challenging, particularly for non-linear models (Draper and Smith 1998). A function has been built in **BLA** to help determine of starting values for the inbuilt models. Wrong or inappropriate starting values result in longer iteration, greater

execution time, non-convergence of the iteration, and possibly convergence a local minimum. To avoid this, several starting values can be used for the optimization and the output with the smallest value is chosen.

The third heuristic method is the quantile regression method which models the relationship of y as a function of x conditional of quantile τ $(0 < \tau < 1)$ (Davino, Furno, and Vistocco 2014) as:

$$Q_Y(\tau|\mathbf{X}) = \mathbf{X}\beta_\tau, \tag{16}$$

where $\mathbf{Q_Y}(\tau|\mathbf{X})$ is a n x 1 vector of dependent variable Y conditional on τ, β is a $p \times 1$ vector of regression parameters, $\mathbf{X}$ is an $n \times p$ matrix of predictors. Unlike the other heuristic methods that fit the boundary line using the least squares methods, estimates of the coefficients, $\mathbf{b}_\tau$, of β_τ for quantile regression are obtained by minimizing the weighted absolute residual values

$$\sum_{i=1}^{n} \rho_\tau(y_i - x_i\boldsymbol{\beta}^T), \tag{17}$$

where the residuals are given by $y_i - x_i\boldsymbol{\beta}^T$ and $\rho_\tau(u)$ is the quantile loss function defined as:

$$\rho_\tau(u) = \begin{cases} \tau \cdot u & \text{if } u \geq 0, \\ (1-\tau) \cdot u & \text{if } u < 0. \end{cases} \tag{18}$$

The negative residuals are given weight equal to $1-\tau$ while the positive residuals are given weights equal to τ. Similar to the least squares method, these equations are solved using various numerical minimization algorithms. The selected τ value, as with the binning method, is arbitrary.

Statistical method

The censored bivariate normal model procedure is a statistical method based on a joint bivariate normal distribution of factor x and response y with an upper boundary which y can not exceed. Lark and Milne (2016) described this distribution as:

$$f(y, x) = \phi(\boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{C}), \tag{19}$$

where ϕ is the bivariate normal density function, $\boldsymbol{\mu}$ is a vector of means, $\boldsymbol{C}$ is the covariance matrix and $\boldsymbol{Z}$ is a vector that represents the boundary function with parameters, $\boldsymbol{\beta}$. For a boundary line model given as, $b(x) = \bar{y}$, any variate {y,x} for which y is greater than the boundary value, $\bar{y}$, is replaced by $\{\bar{y}, x\}$.

The notation so far describes the underlying model. Our observations, $\tilde{y}$ are made with error:

$$\tilde{y} \;=\; \bar{y} + \varepsilon, \quad \varepsilon \;\sim N(0, \sigma_\varepsilon), \tag{20}$$

where σ_ε denotes the standard deviation of the measurement error. With this, the censored bivariate normal model for our data can be written as:

$$f(\tilde{y}, x|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{C}, \sigma_e). \tag{21}$$

For simplicity, we can omit the parameters from the density function and, following the properties of conditional densities, the function $f(\tilde{y}, x)$ can be written as:

$$f(\tilde{y}, x) = f(\tilde{y}|x)f(x), \tag{22}$$

where $f(x)$ is the probability density function of x. Given the assumptions made about measurements error in Equation 20 the conditional density in Equation 22, can be written as a convolution $(f * g)$ of the two functions as:

$$f(\tilde{y}, x) = f(\bar{y}|x) * f_N(v|0, \sigma_e). \tag{23}$$

The conditional density $f(\bar{y}|x)$ in Equation 23 is the censoring of the conditional density $f(y|x)$ and can therefore be written as:

$$f_N(y|\mu_{y|x}, \sigma_{y|x}), \tag{24}$$

where $\mu_{y|x}$ and $\sigma_{y|x}$ are the conditional mean and standard deviation, respectively, of y given as:

$$\mu_{y|x} = \mu_y + (x - \mu_x)\frac{\text{Cov}\{x, y\}}{\sigma_x^2}, \tag{25}$$

and

$$\sigma_{y|x} = \sigma_y\sqrt{1 - \rho^2}, \tag{26}$$

where μ_y and μ_x are elements of $\boldsymbol{\mu}$ in Equation 19 and the correlation, ρ and (co)variances, Cov, are from the covariance matrix $\mathbf{C}$ in Equation 19. The censored conditional density for an upper-bounded distribution (upper boundary), can be broken down as:

$$f(\bar{y}|x) = \begin{cases} f_N(y|\mu_{y|x}, \sigma_{y|x}) & \text{if } y < b(x), \\ \int_{b(x)}^{\infty} f_N(y|\mu_{y|x}, \sigma_{y|x})\ dy & \text{if } y = b(x), \\ 0 & \text{if } y > b(x). \end{cases} \tag{27}$$

Given a variate, $\{\bar{y}, x\}$, in a data set with parameters $\boldsymbol{\mu}$ and $\boldsymbol{C}$ bounded by a model with parameters $\boldsymbol{\beta}$ and measurement error σ_e, the likelihood can be computed by substituting $f(\tilde{y}, x)$ from Equation 23 into Equation 22. The likelihood of the censored bivariate normal distribution for the dataset is calculated as the product of the individual likelihoods. To avoid numerical underflow when dealing with a large number of small likelihoods, the sum of the log likelihoods is computed instead of multiplying the individual likelihoods.

The standard deviation of the measurement error (σ_e), which accounts for the error in boundary line, is a parameter of the censored bivariate normal model. In principle it can be estimated with the others (Lark and Milne 2016) but in practice it is fixed at a value chosen separately. Independent estimates of measurement error might be available (e.g Kosmowski, Chamberlin, Ayalew, Sida, Abay, and Craufurd (2021)). Lark and Milne (2016) showed that estimates of measurement error for N_2O emission from soil closely matched their maximum likelihood estimates in the boundary line model. Lark *et al.* (2020) used the nugget variance from a spatial model of crop yields as an estimate of measurement error. An alternative is to use the likelihood profiling procedure (Royall 2017). In this case, σ_e is fixed at each of a set of values in turn, and the remaining parameters, θ, are estimated conditional on σ_e by maximum likelihood. The maximized likelihoods for the sequence of values constitute a likelihood profile. The value of σ_e is selected where the profile likelihood is maximized. The **BLA** package provides a function for likelihood profiling.

The parameters of the boundary line and the bivariate normal distribution are determined by minimizing a negative log-likelihood objective function following Equation 23. The negative

log-likelihood is used because the optimization methods in this package are minimization algorithms, hence the log-likelihood objective function is multiplied by negative one to achieve this. As explained in Section 2.3.1, the minimization of objective functions is achieved using various numerical minimization algorithms described in 2.3.1. The uncertainty (standard error) of the parameters is determined using the Hessian matrix $\mathbf{H}$ following the procedure described by Dobson and Barnett (2018). The standard errors for each estimated parameter in $\boldsymbol{\beta}$ can be computed by taking the inverse of $\mathbf{H}$ and then taking the square root of the corresponding terms on the main diagonal as:

$$SE(\beta_i) = \sqrt{\mathbf{H}^{-1}[i,i]}, \tag{28}$$

where $SE(\beta_i)$ is the standard error of a i^{th} parameter in $\boldsymbol{\beta}_i$. In cases where $\mathbf{H}$ is not a positive definite matrix and hence not invertible, a near positive definite matrix of $\mathbf{H}$ is determined following the procedure of Tanaka and Nakata (2014).

The Akaike Information Criterion (AIC) is used to compare the censored bivariate normal model with an alternative bivariate normal model with no censoring boundary on a particular data set. The AIC criterion penalizes the more complex model, and so provides evidence that the inclusion of the boundary is justified. The AIC values for these two scenarios, Λ_{bl} and Λ_{bvn} respectively, are calculated as:

$$\Lambda = 2p + 2\ell, \tag{29}$$

where Λ is the AIC value, p is the number of model parameters and ℓ is the log-likelihood value. The scenario with the smallest Λ is deemed the better model. If Λ_{bl} is smaller Λ_{bvn}, this implies that the additional complexity of the of boundary line model by comparison to the bivariate normal model is justified by the magnitude of the improved likelihood. In general if one selects the model with smaller AIC this minimizes the expected information loss through the selection decision (Verbeke and Molenberghs 2000).

2.4. The Most-limiting factor

One approach to interpretation of the boundary line uses the principle of the most limiting factor, as first formulated by Liebig (1840), to determine the factor, out of those observed, that most likely limits the biological response following, the so-called Law of the Minimum:

$$y = \min\left\{f_i(x_1), f_i(x_2), \ldots, f_i(x_n)\right\}, \tag{30}$$

where $f_i(x_i)$ is the boundary line function for factor i, where $i = 1, 2, ...n$. In practice, the most-limiting factor using the boundary line methodology for a given point, is the factor, x_i, that predicts the smallest biological response from the fitted boundary lines, $f_i(x_i)$ (Shatar and McBratney 2004). In the case were all the factors predict the maximum possible biological response, the limiting factor is classified as unknown.

3. The BLA library functions

The theoretical framework and procedures outlined in Section 2 have been operationalized through a series of functions that together form the **BLA** package, designed to facilitate practical implementation of boundary line analysis. These include (i) data exploration functions,

(ii) boundary fitting functions, and (iii) post-hoc boundary line fitting functions. The package relies on several core R dependencies, including **data.table** (Dowle and Srinivasan 2023), **MASS** (Venables and Ripley 2002), **mvtnorm** (Genz, Bretz, Miwa, Mi, Leisch, Scheipl, and Hothorn 2021), **numDeriv** (Gilbert 2019), and the base **stats** package (R Core Team 2025). This section describes the structure and functionality of these components in detail.

3.1. Data exploration functions

The following functions are used to carry out exploratory analysis prior to fitting the boundary line model to a dataset.

Summary statistics

The `summastat()` function obtains summary statistics of the individual variables to be analysed. Its output comprises the mean, median, first and third quartile, variance, skewness, octile skewness (Brys *et al.* 2003), kurtosis and the number of univariate outliers as judged by the outer fences of the data (Tukey 1977). These, along with summary plots (see below) can be used to assess the plausibility that the variables are drawn from a normal, or censored normal random variable. A skewness value outside the range of $[-1, 1]$ or a more robust measure of skewness, octile skewness, outside the range of $[-0.2, 0.2]$ are commonly treated as evidence that the data require transformation, but should be interpreted alongside exploratory plots, the histogram, boxplot and qq-plot (Webster and Lark 2019). The application of the `summastat()` function is shown in the code block below.

```
R> library(BLA)
R> x <- soil$P

R> summastat(x)
```

Outlier detection

The **BLA** package does not have a function to detect bivariate outliers. We recommend the `bagplot()` function of the **aplpack** package for the identification of bivariate outliers. The usage of the `bagplot()` function is demonstrated in the chunk of code below.

```
R> library (aplpack)

R> x <- log(soil$P)
R> y <- soil$yield
R> dat <- cbind(x,y)

R> bagplot(dat)
```

Boundary exploration

The boundary exploration function, `exp_boundary()`, is designed to check for evidence of boundary structure in a data cloud as described in Section 2.1 especially when heuristic

methods are to be used to fit the boundary line model. It inputs the `x` and `y` variables, the number of `shells` (peels) to be extracted from the upper edge of the data cloud and the number of `simulations` for the normal multivariate distributions used to test evidence. Its output is a dataframe containing the *sd* of euclidean distances, mean *sd* and the *p*-values in the left and right sections of the data cloud. Additionally, an $\{x, y\}$-scatter plot showing the boundary points in the left and right sections is produced together with two histograms for the distribution of the *sd* in the left and right sections of data from the Monte Carlo simulations. The *sd* from actual data is also shown on the histograms. The script snippet below is an illustration of the `expl_boundary()` function.

```
R> x <- log(soil$P)
R> y <- soil$yield

R> expl_boundary(x,y, shells=10, simulations=1000)
```

Start values for optimization algorithms

The `startValues()` function helps identify initial parameter values for boundary functions when these are estimated by least squares or maximum likelihood as described in Sections 2.3.1 and 2.3.2. This function applies only for models that are built in **BLA** package, Table 1. It entails (i) plotting an $\{x,y\}$-scatter plot, (ii) running the `startValues()` function with the argument `model` set to the desired model e.g. `model="lp"` for linear-plateau, and (iii) use the mouse to click on the $\{x,y\}$-scatter plot the points at the upper edges of the data cloud that make up the structure of the chosen model in a clockwise movement. The number of points to click depend on the model selected. This number of clicks is displayed in the console when the function is run. For example `"lp"` model requires clicking two boundary points i.e. any point on the sloping side of the model and the likely inflection point (visual method described in section 2.3.1). Other models like the `"logistic"` require clicking multiple boundary points as shown in Figure 1. The output is `list` that includes the model form and the suggested starting values. Note that for the `startValues()` functions to work effectively, the appearance zoom in the global options of R and the zoom in the computer display settings should be equal i.e. both should be set to 100%. A demonstration of the `startValues()` function is provided in the code block below.

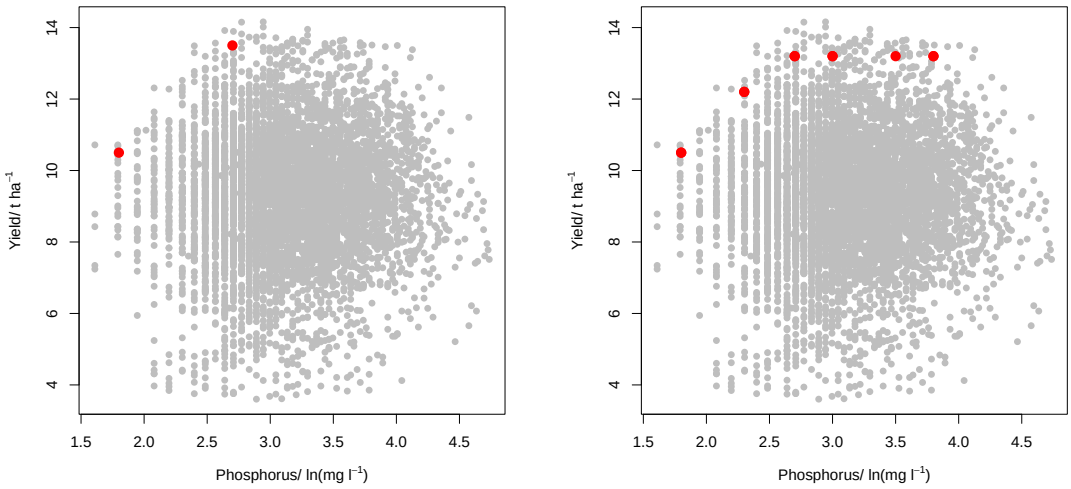
```
R> x <- log(soil$P)
R> y <- soil$yield
R> plot(x,y)

R> startValues(model="lp")
```

3.2. Boundary fitting functions

The following functions are used to determine the parameters of the boundary line models.

Binning method

The function `blbin()` determines the parameters of the suggested boundary line model using the boundary points selected using the binning method described in Section 2.3.1. It inputs



(a) (b)

Figure 1: Example of the points (red) to click on the scatter-plot of wheat yield against soil P for (a) `startValues(model="lp")` and (b) `startValues(model="logistic",6)`.

the the `x` and `y` variables, a vector describing the size of the `bins`, the percentile `tau` considered as the boundary value for the bins , the preferred boundary line `model`, the start values `start` (described in section 3) and `optim.method` for the optimization method to be used. When `model="other"`, then the argument `equation` should define the custom equation using the `function()` function of base R. The `optim.method` includes the options `"Nelder-Mead"`, `"BFGS"`, `"CG"`, `"SANN"` and `"Brent"`. The output of the function is a `list` consisting of the model name, model equation form, estimated parameters of the boundary line (no uncertainty) and the residual mean square value. The implementation of the `blbin()` function is presented the script snippet below.

```
R> x <- evapotranspiration$`ET(mm)`
R> y <- evapotranspiration$`yield(t/ha)`
R> bins <- c(100,350,25)

R> blbin(x,y, bins=bins ,start=c(0.5,0.02), model = "blm",
+        optim.method="Nelder-Mead")
```

Boundary line determination technique

The function `bolides()` estimated the parameters of the suggested boundary line model using the boundary points selected with the bolides algorithm (Schnug *et al.* 1995) described in Section 2.3.1. It inputs the the `x` and `y` variables, the preferred boundary line `model`, the start values `start` (described in section 3) and `optim.method` for the optimization method

to be used. The optimization options are similar to what was described in section 3.2.1. The output is a `list` consisting of the model name, model equation form, estimated parameters of the boundary line (no uncertainty) and the residual mean square value. Below is the script snippet of the `bolides()` function.

```
R> x <- evapotranspiration$`ET(mm)`
R> y <- evapotranspiration$`yield(t/ha)`

R> bolides(x,y, start = c(0.5,0.02), model= "blm",
+                optim.method="Nelder-Mead")
```

Quantile regression

The function `blqr()` determines the parameters of the suggested boundary line model using the principle of quantile regression (Cade and Noon 2003) described in Section 2.3.1. It inputs the the `x` and `y` variables, the preferred boundary line `model`, the quantile value `tau` to consider as the boundary, the start values, `start` (described in section 3) and `optim.method` for the optimization method to be used. The optimization options are similar to what was described in section 3.2.1. The output is a `list` consisting of the model name, model equation form, estimated parameters of the boundary line and the weighted residual sum squares value. Below is a practical example of applying the `blqr()` function.

```
R> x <- evapotranspiration$`ET(mm)`
R> y <- evapotranspiration$`yield(t/ha)`

R> blqr(x,y, start = c(0.5,0.02), model= "blm",tau=0.95,
+        optim.method="Nelder-Mead")
```

Censored bivariate normal model

The function `cbvn()` estimates the parameters of a censored bivariate normal model (Milne *et al.* 2006b; Lark and Milne 2016) described in Section 2.3.2. The inputs of this function include a dataframe `data` containing the x and y variables, the boundary line `model` to be fitted, the standard deviation of the measurement error value `sigh` (see next section), the initial start values `start` (described in section 3) of model together with the bivariate distribution properties (mean of `x` and `y`, standard deviation of `x` and `y`, and the correlation of `x` and `y`) and `optim.method` for the optimization method to be used. The optimization options are similar to what was described in section 3.2.1. The output is a `list` consisting of the model name, model equation form, estimated parameters of the boundary line and bivariate distribution with their respective standard error values, the AIC values for the uncensored bivariate normal model and censored bivariate normal model respectively, and a hessian matrix. The operation of the `cbvn()` function is illustrated below.

```
R> x <- evapotranspiration$`ET(mm)`
R> y <- evapotranspiration$`yield(t/ha)`
R> dat <- data.frame(x,y)
```

```
R> start <- c(0.5,0.02,mean(x),mean(y),sd(x),sd(y),cor(x,y))

R> cbvn(data=dat,start = start, sigh = 0.4, model= "blm",
+        optim.method="BFGS")
```

Likelihood profile for standard deviation of measurement error estimation

When an independent estimate of the standard deviation of measurement error is not available, the function `ble_profile()` can be used to develop a likelihood profile for suggested standard deviation of measurement error values. The inputs of this function include a dataframe, `data`, containing the `x` and `y` variables, a suggested boundary line `model`, the initial start values, `start` (described in section 3) of boundary model together with the bivariate distribution properties (mean of `x` and `y`, standard deviation of `x` and `y`, and the correlation of `x` and `y`), a vector of length n with suggested standard deviation of measurement error values `sigh` and the optimization method `optim.method` to be used. The model and optimization options are similar to what was described in section 3.2.4. The output is a `list` of length 2 containing log-likelihood profile values and the corresponding standard deviations of measurement error. A likelihood profile plot is also produced. Below, the `ble_profile()` function is demonstrated in use.

```
R> x<-evapotranspiration$`ET(mm)`
R> y<-evapotranspiration$`yield(t/ha)`
R> dat <- data.frame(x,y)
R> sigh <- c(0.08,0.1,0.11,0.13,0.15,0.2,0.3,0.4,0.5)
R> start <- c(0.5,0.02,mean(x),mean(y),sd(x),sd(y),cor(x,y))

R> ble_profile(dat,sigh=sigh,start=start,model="blm",
+            optim.method="BFGS")
```

3.3. Functions for interpretation of boundary models

The following functions are used to carry out further analysis after fitting the boundary line model.

Prediction of boundary value

The `predictBL()` function is used to predict the boundary value given x using the boundary line parameters obtained using `blbin()`, `bolides()`, `blqr()` and `cbvn()` functions. The inputs for this function are the `x` value(s) for which prediction is to be made and a model object from the results using the `blbin()`, `bolides()`, `blqr()` and `cbvn()` functions. This results in a predicted value(s) of y given that no other factor is limiting. The usage of the `predictBL()` function is demonstrated below.

```
R> x <- evapotranspiration$`ET(mm)`
R> y <- evapotranspiration$`yield(t/ha)`
R> z <- bolides(x,y, start = c(0.5,0.02), model= "blm", xmax = 350,
+            optim.method="Nelder-Mead")
```

```
R> predictBL(z,x)
```

Most-limiting factor

When several boundary line models of different factors of interests on y are obtained, the function `limfactor()` can be used to determine the most-limiting factor according to the Liebig (1840) law of the minimum. The inputs of the function are boundary line objects (for different factors of interest) obtained using either of the boundary line fitting function `blbin()`, `bolides()`, `blqr()` and `cbvn()`. The output is a `vector` indicating the factor that predicted the smallest response for each entity. Below is an illustration of the `limfactor()` function in action. We Assume that the vectors N, P and K are predicted responses using the fitted BL.

```
R> N <- rnorm(10,50,5)
R> K <- rnorm(10,50,4)
R> P<- rnorm(10,50,6)

R> limfactor(N,K,P)
```

4. Illustrative Example

4.1. Background

AgSpace Agriculture Ltd conducts soil sampling to provide fertilizer recommendations for its customers using pre-identified management zones within each field. The recommendations are based on the comparisons of soil variables with standard values which are usually established. The boundary line methodology has been recognised as means to derive standard values from which fertilizer recommendations can be made (Smith *et al.* 2024). Here, we apply the boundary line methodology to a dataset collected by AgSpace Agriculture Ltd to extract information on the possible limitations to increased productivity and provide target values that maximise production in order to increase productivity. The dataset used in this illustration is from a survey of 6110 wheat farms across England in 2016. Wheat yields (t ha^{-1}) and associated soil properties including soil pH and soil P concentration (mg l^{-1}) were collected. This dataset is available in the **BLA** package. The aims of this illustration are to (1) modelling the boundary line for the wheat yield as a function of soil P and soil pH using the censored bivariate normal model, (2) determining their critical values of soil P and soil pH, and (3) determining if soil P or pH is the most-limiting factor in each field.

4.2. Methods and Results

The boundary line analysis was conducted in three general steps. These were data exploratory analysis, fitting the boundary line model and conducting post hoc analysis. The R code for this analysis can be found at https://github.com/chawezimiti/BLA_Package_illustration.

In the exploratory step, summary statistics of the variables yield, soil P and soil pH were obtained with the `summastat()` function to assess the plausibility of the assumption of (censored)

normality. The assumption was plausible for yield (Table 2). Soil P required a transformation to logs. The Yeo-Johnson transformation (Yeo and Johnson 2000) was applied to data on soil pH.

Table 2: Summary statistics for the original and transformed variables soil P and pH

Variable	Transformation	Power (λ)	Mean	median	sd	skewness	O.Skewness
soil P	None	-	25.96	22	14.38	1.84	0.36
Soil P	log	0	3.13	3.09	0.50	0.13	0.08
Soil pH	None	-	7.56	7.74	0.65	-0.79	-0.36
Soil pH	log	0	2.02	2.05	0.09	-0.99	-0.40
Soil pH	box-cox	2	28.34	29.45	4.82	-0.63	-0.31
Soil pH	Yeo-Johnson	4.99	0	0.13	1.00	-0.27	-0.19
Wheat Yield	None	-	9.25	9.36	1.86	-0.48	-0.06

The bivariate outliers for a combination of yield and each of the two independent variable were identified using the `bagplot()` function and these were removed from the data. The code block below shows the execution of the analysis and results for data on Soil P and yield (code block for soil pH not shown here). A dataframe, `dat`, containing the variables soil P and yield was constructed and used as input for the `bagplot()` function. From the resulting object, `out`, bivariate outliers were excluded by creating a new dataset, `vals`, which included only the data points located within the bag and loop regions (see also Figure 2).

```
R> dat <- data.frame(x=log(soil$P), y=soil$yield)
R> out <- bagplot(dat,show.whiskers = F)
R> legend("bottomright", legend = c("Bag","Loop","Depth median", "outlier"),
+       pch = c(15,15,8,3), col = c(adjustcolor( "blue", alpha.f = 0.7),
+               "lightblue","red","red"))

R> vals <- rbind(out$pxy.bag,out$pxy.outer)
R> head(vals)

            x        y
[1,] 3.091042  10.86890
[2,] 2.639057  10.07150
[3,] 3.135494  10.57600
[4,] 2.995732  10.58160
[5,] 3.044522  10.75790
[6,] 3.637586   9.02014
```

The model to describe the relationship of yield as a function of soil P concentration and soil pH were selected based on the structure of the boundary points and agronomic principles. For soil P the trapezium model was used as indicated by the structure of the boundary points and strong evidence of bounding structure in both sections of the data. This is an agronomically plausible model as yield is expected to increase with soil P concentration until a maximum biological limit is reached. The soil P concentration at this point is known as the critical concentration, and any increase in soil P beyond this point results in 'luxury uptake' of P, or

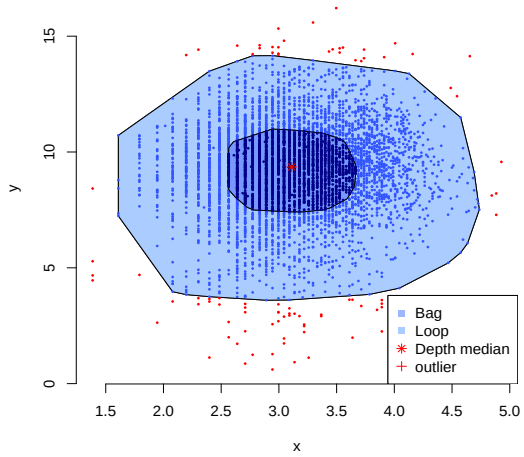


Figure 2: Bagplot showing data points in the bag and loop, the depth median and the outliers for soil P and Yield data

accumulation in the soil, with no yield gain. However, excessive soil P can induce zinc and iron deficiencies and disrupt mycorrhizal associations (Khan, Siddique, Shabala, Zhou, and Zhao 2023). As a result, at some soil P concentration beyond the critical soil P concentration, yield reduction is expected. For soil pH a linear-plateau model was used as indicated by the structure of the boundary points. It is expected that an increase in soil pH from more acidic conditions will result in increased wheat yield until the maximum yield is achieved in the optimum pH range (Holland, White, Glendining, Goulding, and McGrath 2019). A reduction in yield may be expected at larger pH values but the data structure does not indicate that.

The parameters of the boundary line were determined using the `cbvn()` function. The initial start values for the optimization process for the boundary line models for soil P and soil pH were determined using the `startValue()` function with argument set to `model="trapezium"` and `model="lp"` respectively. The obtained initial start values for soil P and soil pH are given in Table 3. As there was no directly measured value of the measurement error for yield, the likelihood profile was used to determine the standard deviation of the measurement error (σ_e) using the `ble_profile()` function. The tested possible σ_e values ranged from 0.3 to 1 t ha^{-1}. From the likelihood profile, a value of 0.4 was selected to represent σ_e (Figure 3).

Table 3: Initial start values for the boundary models of soil P and soil pH

Factor	β_1	β_2	β_0	β_3	β_4	μ_x	μ_y	s_x	s_y	ρ_{xy}
Soil P	4.30	3.40	13.80	32.8	−4.9	3.13	9.29	0.50	1.73	0.03
Soil P	2.50	4.10	13.42	32.0	−4.8	3.13	9.29	0.50	1.73	0.03
Soil P	3.50	3.70	13.35	47.70	−8.4	3.13	9.29	0.50	1.73	0.03
Soil P	2.83	4.11	13.70	32.00	−4.60	3.13	9.29	0.50	1.73	0.03
Soil P	4.10	3.40	13.60	29.00	−4.10	3.13	9.29	0.50	1.73	0.03
Soil pH	20.00	4.30	13.57			0.003	9.28	0.99	1.76	0.12
Soil pH	21.30	5.40	13.44			0.003	9.28	0.99	1.76	0.12
Soil pH	20.90	5.20	13.57			0.003	9.28	0.99	1.76	0.12
Soil pH	18.47	3.54	13.87			0.003	9.28	0.99	1.76	0.12
Soil pH	20.70	4.50	13.67			0.003	9.28	0.99	1.76	0.12

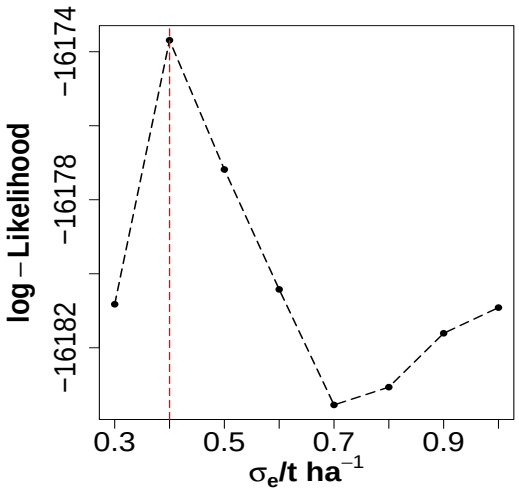


Figure 3: Estimate of σ_{me} using the log-likelihood profiling. The red dashed line represents the estimate of σ_{me}

The `cbvn()` function was used to fit the boundary line models. This was run using a wide range of initial starting values to avoid sticking at a locally optimum solution. The run with the smallest negative likelihood was taken as the best model fit. The results of the fitted models and their standard error values (indicated in brackets), and the AIC values for the boundary (Λ_{bl}) and the bivariate normal (Λ_{mvn}) models are shown in Table 4 and Figure 4. The Λ values for both the soil P and soil pH models indicate that the boundary line models were appropriate as the Λ values were smaller than the corresponding bivariate normal models. Looking at the two boundary line models (Figure 4), we interpret the rising section of the boundary model as showing the P-limited/pH-limited response of yield, and the soil P/soil pH value at which this intersects the horizontal section as a critical value. In the case of soil P, the falling section of the line is likely to reflect the accumulation of P in the soil in areas with other consistent large limitations on crop yield (e.g. soil compaction in

headlands observed during the sampling process by AgSpace), or some effects of excess P on the crop as described above.

Table 4: Parameters of fitted boundary line models for soil P and pH

Factor	β_1	β_2	β_0	β_3	β_4	Λ_{bl}	Λ_{mvn}
P	5.75(0.88)	2.82(0.37)	13.44(0.09)	82.57(12.66)	−15.77(2.71)	32376.46	32429.55
pH	26.93(1.88)	6.99(0.82)	13.61(0.10)			41073.00	41093.24

From the model parameters in Table 4, the critical values of soil P and pH were determined using the formula:

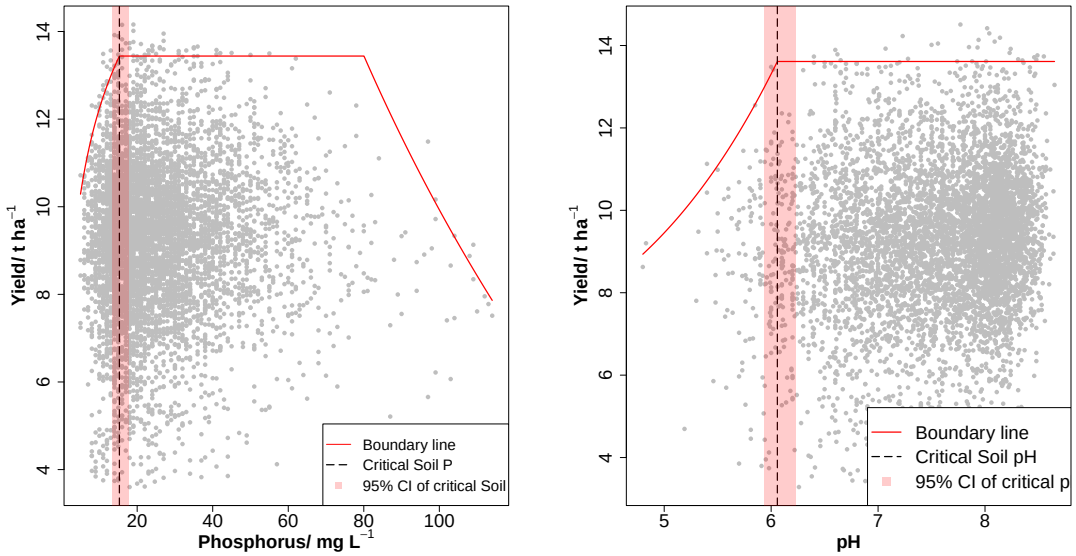
$$x_{\text{crit}} = \frac{\beta_0 - \beta_2}{\beta_1} \tag{31}$$

where β_0 is the plateau yield, β_1 is the intercept value and β_2 is the rate of change in yield with increase in factor x_i. A critical value of a factor indicates the values beyond which increase in the factor does not result in any increase in yield. The determination of critical values is vital as it helps to determine fertilization strategies preventing economic losses and environmental pollution (Andrade *et al.* 2023). The critical values determined using the model parameters shown in Table 4 were back transformed to the original scale for easy interpretation. The critical value for soil P was found to be 15.29 mg l^{-1} with a 95% confidence interval (determined as the highest density interval (Meredith and Kruschke 2020)) of 13.46 – 17.79 mg l^{-1} while that of pH was found to be 6.06 with a confidence interval of 5.93 – 6.26. The critical values obtained here are in agreement with the standard values derived from experimental data by RB209 (AHDB 2023) for soils of the UK. They give a range of the critical value for soil P to be 15 – 25 mg l^{-1} and a pH of 6.5 (just outside the confidence interval).

The model parameters of the two models were used to predict the largest yield given a value of soil P and pH. These were used to determine the most-limiting factor using the `limfactor()` function. The results as shown in Figure 5 indicate that soil P was found to be limiting on 14.82% of the sampled points while soil pH only accounted for 1.89%. As only two factors were evaluated, most of the yield limitation cause was unidentified (83.29%). This is because there are many other factors affecting yield which we did not consider in this case as data were not available.

4.3. Final Remarks

In this illustration, the boundary line methodology has been used to extract some useful information from survey data that can be used to enhance agricultural productivity. The determination of the most-limiting factor helps farmers prioritize which factor to address first in a farming system while the determination of critical values for nutrients helps in the management of plant nutrition as a whole and can be highly useful in site specific agricultural management.



(a) (b)

Figure 4: Boundary line models fitted to (a) soil P and (b) pH on the original scale

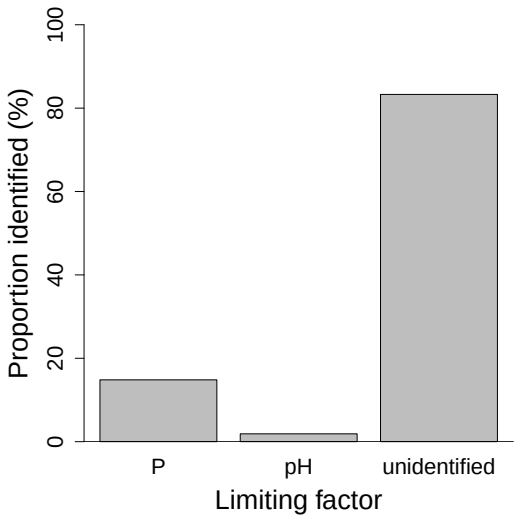


Figure 5: Proportions of the identified limiting factor in the management units

5. Conclusion

We introduced the **BLA** R package, the first package specifically designed for implementing boundary line analysis. It simplifies the entire process of fitting the boundary line into user-friendly functions. The example provided in this paper demonstrates the usefulness of the **BLA** R package for addressing practical ecological problems, such as agriculture crop production, by making the process of fitting the boundary line model and extracting information from it easier and reproducible.

6. Acknowledgements

We 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. We acknowledge support from the Growing Health Institute Strategic Programme [BB/X010953/1; BBS/E/RH/230003C]. We further acknowledge AgSpace Agriculture limited for allowing us to use the datasets shown here and included in the **BLA** package.

References

AHDB (2023). *Nutrient Management Guide (RB209). Section 1: Arable Crops*. Agriculture and Horticulture Development Board, Coventry, UK. URL https://ahdb.org.uk/documents/RB209/RB209_Section1_WEB_2017-12-01.pdf.

Andrade CB, Comin JJ, Moura-Bueno JM, Brunetto G (2023). "Obtaining Reference Values for Nutrients in Vineyard Soils through Boundary Line Approach Using Bayesian Segmented Quantile Regression on Commercial Farm Data." *European Journal of Agronomy*, **150**, 126928.

Baral R, Bhandari K, Kumar R, Min D (2022). "Yield Gap Analysis of Alfalfa Grown under Rainfed Condition in Kansas." *Agronomy*, **12**(9), 2190.

Black BA, Abrams MD (2003). "Use of Boundary-line Growth Patterns as a Basis for Dendroecological Release Criteria." *Ecological Applications*, **13**(6), 1733–1749.

Boyd SP, Vandenberghe L (2004). *Convex Optimization*. Cambridge university press.

Brys G, Hubert M, Struyf A (2003). "A Comparison of Some New Measures of Skewness." In *Developments in Robust Statistics: International Conference on Robust Statistics 2001*, pp. 98–113. Springer.

Cade BS, Noon BR (2003). "A Gentle Introduction to Quantile Regression for Ecologists." *Front Ecol Environ*, **1**, 412–420. URL www.frontiersinecology.org.

Casanova D, Goudriaan J, Bouma J, Epema G (1999). "Yield Gap Analysis in Relation to Soil Properties in Direct-seeded Flooded Rice." *Geoderma*, **91**(3-4), 191–216. doi:https://doi.org/10.1016/S0016-7061(99)00005-1.

Chambers J, Hinckley T, Cox G, Metcalf C, Aslin R (1985). "Boundary-Line Analysis and Models of Leaf Conductance for Four Oak-Hickory Forest Species." *Forest Science*, **31**, 437–450.

Davino C, Furno M, Vistocco D (2014). *Quantile Regression: Theory and Applications*, volume 988. John Wiley and Sons.

Dobson AJ, Barnett AG (2018). *An Introduction to Generalized Linear Models*. CRC press.

Dowle M, Srinivasan A (2023). *data.table: Extension of 'data.frame'*. R package version 1.15.4, URL https://CRAN.R-project.org/package=data.table.

Draper NR, Smith H (1998). *Applied Regression Analysis*, volume 326. John Wiley & Sons.

D'Orazio M (2023). "Some Approaches to Outliers' Detection in R." *Romanian Statistical Review*, (1).

Elliott JA, de Jong E (1993). "Prediction of Field Denitrification Rates: A Boundary-Line Approach." *Soil Science Society of America Journal*, **57**, 82–87. ISSN 1435-0661. doi:https://doi.org/10.2136/sssaj1993.03615995005700010016x.

Fermont AM, van Asten PJ, Tittonell P, van Wijk MT, Giller KE (2009). "Closing the Cassava Yield Gap: An Analysis from Smallholder Farms in East Africa." *Field Crops Research*, **112**, 24–36. ISSN 03784290. doi:https://doi.org/10.1016/j.fcr.2009.01.009. Https://doi.org/10.1016/j.fcr.2009.01.009.

Fink KP, Grassini P, Rocateli A, Bastos LM, Kastens J, Ryan LP, Lin X, Patrignani A, Lollato RP (2022). "Alfalfa Water Productivity and Yield Gaps in the US Central Great Plains." *Field Crops Research*, **289**, 108728.

French RJ, Schultz JE (1984). "Water Use Efficiency of Wheat in a Mediterranean-type Environment 2. Some Limitations to Efficiency." *Aust. J. Agric. Res*, **35**, 765–75.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-3, URL https://CRAN.R-project.org/package=mvtnorm.

Gilbert P (2019). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1, URL https://CRAN.R-project.org/package=numDeriv.

Hajjarpoor A, Soltani A, Zeinali E, Kashiri H, Aynehband A, Vadez V (2018). "Using Boundary Line Analysis to Assess the On-Farm Crop Yield Gap of Wheat." *Field Crops Research*, **225**, 64–73. ISSN 03784290. doi:https://doi.org/10.1016/j.fcr.2018.06.003.

Helidoniotis F, Haddon M, Tuck G, Tarbath D (2011). "The Relative Suitability of the von Bertalanffy, Gompertz and Inverse Logistic Models for Describing Growth in Blacklip Abalone Populations (Haliotis rubra) in Tasmania, Australia." *Fisheries Research*, **112**(1-2), 13–21.

Holland J, White P, Glendining M, Goulding K, McGrath S (2019). "Yield Responses of Arable Crops to Liming–An Evaluation of Relationships Between Yields and Soil pH from a Long-term Liming Experiment." *European Journal of Agronomy*, **105**, 176–188.

Hosmer DW (2000). *Applied Logistic Regression, 2*. John Wiley and Sons.

Houshmandfar A, Fitzgerald GJ, Armstrong R, Macabuhay AA, Tausz M (2015). "Modelling Stomatal Conductance of Wheat: An Assessment of Response Relationships Under Elevated CO_2." *Agricultural and Forest Meteorology*, **214-215**, 117–123. URL https://www.sciencedirect.com/science/article/pii/S0168192315006814.

Khan F, Siddique AB, Shabala S, Zhou M, Zhao C (2023). "Phosphorus Plays Key Roles in Regulating Plants' Physiological Responses to Abiotic Stresses." *Plants*, **12**(15), 2861.

Kintché K, Hauser S, Mahungu NM, Ndonda A, Lukombo S, Nhamo N, Uzokwe VN, Yomeni M, Ngamitshara J, Ekoko B, Mbala M, Akem C, Pypers P, Matungulu KP, 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." *European Journal of Agronomy*, **89**, 107–123. ISSN 11610301. doi:https://doi.org/10.1016/j.eja.2017.06.011.

Kosmowski F, Chamberlin J, Ayalew H, Sida T, Abay K, Craufurd P (2021). "How Accurate are Yield Estimates from Crop Cuts? Evidence from Smallholder Maize Farms in Ethiopia." *Food Policy*, **102**, 102122.

Lark RM, Gillingham V, Langton D, Marchant BP (2020). "Boundary Line Models for Soil Nutrient Concentrations and Wheat Yield in National-Scale Datasets." *European Journal of Soil Science*, **71**, 334–351. ISSN 13652389. doi:https://doi.org/10.1111/ejss.12891.

Lark RM, Milne AE (2016). "Boundary Line Analysis of the Effect of Water-filled Pore Space on Nitrous Oxide Emission from Cores of Arable Soil." *European Journal of Soil Science*, **67**, 148–159. ISSN 13652389. doi:https://doi.org/10.1111/ejss.12318.

Liebig J (1840). "Organic Chemestry in its Application to Agriculture and Physiology." *Friedrich Vieweg und Sohn Publ. Co., Braunschweig, Germany*.

Liu FT, Ting KM, Zhou ZH (2012). "Isolation-based Anomaly Detection." *ACM Transactions on Knowledge Discovery from Data*, **6**(1), 1–39.

Makowski D, Doré T, Monod H (2007). "A New Method to Analyse Relationships Between Yield Components with Boundary Lines." *Agronomy for Sustainable Development*, **27**, 119–128. ISSN 17740746. doi:https://doi.org/10.1051/agro:2006029.

Mecklin CJ, Mundfrom DJ (2005). "A Monte Carlo Comparison of the Type I and Type II Error Rates of Tests of Multivariate Normality." *Journal of Statistical Computation and Simulation*, **75**(2), 93–107.

Meredith M, Kruschke J (2020). *HDInterval: Highest (Posterior) Density Intervals*. R package version 0.2.

Milne AE, Ferguson RB, Lark RM (2006a). "Estimating a Boundary Line Model for a Biological Response by Maximum Likelihood." *Annals of Applied Biology*, **149**, 223–234. ISSN 00034746. doi:https://doi.org/10.1111/j.1744-7348.2006.00086.x.

Milne AE, Wheeler HC, Lark RM (2006b). "On Testing Biological Data for the Presence of a Boundary." *Annals of Applied Biology*, **149**, 213–222. ISSN 00034746. doi:https://doi.org/10.1111/j.1744-7348.2006.00085.x.

Miti C, Lark RM, Milne AE, Giller KE (2024a). *BLA: Boundary Line Analysis*. R package version 1.0.1.9000, URL https://chawezimiti.github.io/BLA/.

Miti C, Milne A, Giller K, Lark R (2024b). "The Concepts and Quantification of Yield Gap Using Boundary Lines. A Review." *Field Crops Research*, **311**, 109365.

Miti C, Milne A, Giller K, Sadras V, Lark R (2024c). "Exploration of Data for Analysis using Boundary Line Methodology." *Computers and Electronics in Agriculture*, **219**, 108794.

Mitscherlich E (1909). "Des Gesetz des Minimums und das Gesetz des Abnehmended Bodenertrages." *Landwirsch Jahrb*, **3**, 537–552.

Montaner JG, Maddonni G, DiNapoli M (1997). "Modeling Grain Yield and Grain Yield Response to Nitrogen in Spring Wheat Crops in the Argentinean Southern Pampa." *Field Crops Research*, **51**(3), 241–252.

Ndabamenye T, Asten PJV, Blomme G, Vanlauwe B, Uzayisenga B, Annandale JG, Barnard RO (2013). "Nutrient Imbalance and Yield Limiting Factors of Low Input East African Highland Banana (Musa spp. AAA-EA) Cropping Systems." *Field Crops Research*, **147**, 68–78. ISSN 03784290. doi:https://doi.org/10.1016/j.fcr.2013.04.001.

Nelder J (1961). "The Fitting of a Generalization of the Logistic Curve." *Biometrics*, **17**(1), 89–110.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R Core Team (2025). *stats: The R Stats Package*. R Foundation for Statistical Computing, Vienna, Austria. Included with base R, URL https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html.

Rousseeuw PJ, Ruts I, Tukey JW (1999). "The Bagplot: A Bivariate Boxplot." *The American Statistician*, **53**(4), 382–387. doi:https://doi.org/10.1080/00031305.1999.10474494.

Royall R (2017). *Statistical Evidence: A Likelihood Paradigm*. Routledge.

Sadras VO (2020). "On Water-use Efficiency, Boundary Functions, and Yield Gaps: French and Schultz insight and legacy." *Crop Science*, **60**, 2187–2191. ISSN 14350653. doi:https://doi.org/10.1002/csc2.20188.

Schmidt U, Thöni H, Kaupenjohann M (2000). "Using a Boundary Line Approach to Analyze N_2O Flux Data from Agricultural Soils." *Nutrient Cycling in Agroecosystems*, **57**, 119–129. doi:https://doi.org/10.1023/A:1009854220769.

Schnug E, Heym JM, Murphy DPL (1995). "Boundary Line Determination Technique (BOLIDES)." In PC Robert, RH Rust, WE Larson (eds.), *Site Specific Management for Agricultural Systems*, pp. 899–908. Wiley Online Library. doi:https://doi.org/10.2134/1995.site-specificmanagement.c66.

Shabani A, Sepaskhah A, Kamgar-Haghighi A, Honar T (2018). "Using Double Logistic Equation to Describe the Growth of Winter Rapeseed." *The Journal of Agricultural Science*, **156**(1), 37–45.

Shatar TM, McBratney AB (2004). "Boundary Line Analysis of Field-scale Yield Response to Soil Properties." *Journal of Agricultural Science*, **142**, 553–560. ISSN 00218596. doi:https://doi.org/10.1017/S0021859604004642.

Skiena SS (1998). *The Algorithm Design Manual*, volume 2. Springer.

Smith HW, Ashworth AJ, Nalley LL, Schmidt A, Turmel MS, Bucaro A, Owens PR (2024). "Boundary Line Analysis and Machine Learning Models to Identify Critical Soil Values for Major Crops in Guatemala." *Agronomy Journal*, **116**(3), 1071–1087.

Smith JF, Hardie AG (2023). "Determination of Optimal Soil pH and Nutrient Concentrations for Cultivated Rooibos Tea using the Boundary Line Approach." *South African Journal of Plant and Soil*, **40**(4-5), 208–215. URL https://doi.org/10.1080/02571862.2023.2259860.

Smiti A (2020). "A Critical Overview of Outlier Detection Methods." *Computer Science Review*, **38**, 100306.

Tanaka M, Nakata K (2014). "Positive Definite Matrix Approximation with Condition Number Constraint." *Optimization Letters*, **8**(3), 939–947.

Tukey J (1977). *Exploratory Data Analysis*. Addison Wesley.

van Vugt D, Franke AC (2018). "Exploring the Yield Gap of Orange-fleshed Sweet Potato Varieties on Smallholder Farmers' Fields in Malawi." *Field Crops Research*, **221**, 245–256. ISSN 03784290. doi:https://doi.org/10.1016/j.fcr.2017.11.028.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Fourth edition. Springer, New York. R package version 7.3-60, URL https://CRAN.R-project.org/package=MASS.

Verbeke G, Molenberghs G (2000). *Linear Mixed Models for Longitudinal Data*. Springer.

Wairegi LW, van Asten PJ, Tenywa MM, Bekunda MA (2010). "Abiotic Constraints Override Biotic Constraints in East African Highland Banana Systems." *Field Crops Research*, **117**, 146–153. ISSN 03784290. doi:https://doi.org/10.1016/j.fcr.2010.02.010.

Wang N, Jassogne L, van Asten PJ, Mukasa D, Wanyama I, Kagezi G, Giller KE (2015). "Evaluating Coffee Yield Gaps and Important Biotic, Abiotic, and Management Factors Limiting Coffee Production in Uganda." *European Journal of Agronomy*, **63**, 1–11. ISSN 11610301. doi:https://doi.org/10.1016/j.eja.2014.11.003.

Webb RA (1972). "Use of the Boundary Line in the Analysis of Biological Data." *Journal of Horticultural Science*, **47**(3), 309–319. doi:https://doi.org/10.1080/00221589.1972.11514472.

Webster R, Lark RM (2019). "Analysis of Variance in Soil Research: Examining the Assumptions." *European Journal of Soil Science*, **70**, 990–1000. doi:https://doi.org/10.1111/ejss.12804.

Yeo IK, Johnson RA (2000). "A New Family of Power Transformations to Improve Normality or Symmetry." *Biometrika*, **87**(4), 954–959.

Affiliation:

Chawezi Miti
School of Biosciences
University of Nottingham
Leicestershire LE12 5RD, U.K.
E-mail: chawezi.miti@nottingham.ac.uk

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd

Chapter 7

General Discussion

The overall aim of this thesis was to advance the application of the boundary line methodology for yield gap analysis through a critical review of the methods and interpretations, methodological innovations, comparative evaluation of fitting techniques, and the development of an `R` package to facilitate broader accessibility. The findings presented across the five chapters of this thesis contribute to addressing key challenges in the implementation of boundary line analysis and provide insights into best practices for future agronomic research and decision-making.

7.1 Key Findings and Reflections

A fundamental contribution of this thesis is the critical review of the use of boundary line models for yield gap analysis (Chapter 2) which laid a foundation for the improvements to the boundary line methodology. The review identified inconsistencies in the application of boundary line methods, particularly due to the lack of standardized fitting techniques and statistical inference (Hajjarpoor et al., 2018). One of the key components that is often omitted in boundary line analysis is assessing whether the data are suitable for the analysis. The review emphasized the need for rigorous exploratory data analysis to justify the use of boundary line models, given that many studies apply boundary line analysis without testing whether the data exhibit a boundary-limited response (Milne et al., 2006). It also showed that heuristic approaches for fitting boundary models, which are commonly used, lack a rigorous statistical foundation and thereby raising concerns about the reproducibility and robustness of results (Milne et al., 2006).

The review further underscored the agronomic significance of boundary line model interpretations. Boundary line models can be interpreted agronomically using the Law of the Minimum or Law of the Optimum. The application of the Law of the Minimum in boundary line analysis allows for the estimation of the most-limiting factor, whereas

the Law of the Optimum provides a more realistic perspective by considering interactions between factors. Under the law of the minimum, if a data point is located exactly on the boundary line of a given factor, x_i, it implies that all other factors are not limiting and therefore, the yield response will be affected by the increase or reduction in the factor x_i. This assumption of independence among limiting factors in the Law of the Minimum has been challenged, and this thesis argues for a more holistic approach that accounts for variable interactions, possibly through the use of multivariate reaction surfaces (Myers et al., 2016) in boundary line modelling. Additionally, the review highlighted a growing need for integrating modern computational approaches, such as Bayesian inference, to improve model reliability and predictive capabilities (Andrade et al., 2023). By identifying these gaps and opportunities, the critical review established a clear direction for the methodological improvements and innovations developed in this thesis.

Data availability and quality are common challenges in empirical research, making exploratory data analysis an essential first step in the analytical process. Exploratory data analysis has been emphasized by Tukey et al. (1977) and Chatfield (1985) as a vital step in the data analytical process, guiding the selection of appropriate statistical methods and ensuring valid interpretations. Tukey et al. (1977) highlighted the value of visualizing data to uncover unexpected patterns, outliers, and hidden structures that may not conform to formal model assumptions. Similarly, Chatfield (1985) stressed the importance of preliminary data examination, noting that failure to do so can result in misapplied models and misleading conclusions. Both authors underscore that exploratory data analysis is not merely a preliminary task but a fundamental component of sound statistical practice. However, there is a lack of data exploratory analysis procedures for boundary line analysis. To address the lack of data exploratory tools for boundary line analysis, Chapter 3 of this thesis introduced a novel approach to exploratory data analysis using principles of the convex hull peels to assess whether a given dataset exhibits a boundary-limited response. This method addresses a critical gap in boundary line analysis applications by providing an objective criterion to justify the use of boundary line models. The study demonstrated that datasets exhibiting clustering in the upper convex hull peels are more likely to conform to a boundary-limited structure, supporting the validity of applying boundary line analysis (Lark & Milne, 2016; Casanova et al., 1999). This innovation improves the reliability of yield gap estimations by ensuring that boundary lines are not erroneously applied to data that lack a true limiting response thereby reducing the risk of misinterpretation as emphasized by Chatfield (1985).

In the method presented in Chapter 3 for establishing evidence of boundary structures, the standard deviation of the Euclidean distance from the data centroid was used to quantify clustering in the upper peels of convex hulls. However, alternative indices can also provide insight into data concentration. For instance, the area enclosed by data

points in the upper peels estimated using a concave hull can serve as a complementary measure of clustering. A smaller area suggests a tighter grouping of points, indicating a possible limiting boundary. Area-based measures can account for irregular shapes and may capture clustering in non-uniform data distributions. This approach could therefore offer a useful extension to the peel density method, providing an additional line of evidence when assessing the plausibility of boundary structures in biological data.

The developed exploratory method was illustrated through case studies, demonstrating its practical application to agricultural datasets. The findings reinforce the importance of integrating exploratory data analysis as a standard step in boundary line analysis. Moreover, this method enable researchers to assess whether additional data collection is needed to enhance confidence in boundary model estimations. This innovation is a key contribution to standardizing the boundary line methodology and should be considered the first step in the analysis. Conducting exploratory analysis to justify the use of a methodology is essential for instilling confidence not only in the results and their interpretation but also among users of the outputs. However, some studies that have applied the boundary line methodology, such as Omondi et al. (2024), have omitted this crucial step. In such cases, the visual plots of the fitted boundary models often appear inappropriate (e.g. Fig 2a of Omondi et al., 2024), raising questions about the validity of the model interpretations. This lack of methodological justification can undermine the credibility of the findings and makes it difficult for other researchers to use these results as a foundation for future studies or for informing policy decisions.

The need for standardization of the boundary line fitting method has been heavily highlighted throughout this thesis and some steps to achieve this have been given including the use objective data exploratory methods and the provision of an open source `BLA R` package for its implementation. However, the concern should not only be on the consistency of procedures to fit the boundary line model but this should be extended to the interpretation of outputs for agronomic recommendations. Chapter 4 of this thesis presented a comparative evaluation of multiple boundary line fitting methods, including binning, bolides, quantile regression, and the censored bivariate normal model for agronomic recommendations. The results indicated that while all methods broadly identified the same most-limiting factors, there were notable inconsistencies in determination of critical values e.g critical soil nutrient concentrations. While the censored bivariate normal model and quantile regression ($\tau = 0.99$) methods produced estimates of critical nutrient concentrations (x_{crit}) that aligned with agronomic guidelines by the RB209 (AHDB, 2023) for soils of England, other methods, particularly binning, exhibited greater variability. These discrepancies underscore the impact of subjective methodological choices, including the selection of τ values and bin size, on agronomic interpretations and recommendations. This finding highlights the need for a common procedure for boundary line

model fitting.

The stakeholder feedback gathered through workshops in Nairobi and Harare as part of this study indicated no strong preference for any single boundary line fitting method, further reinforcing the need for standardized guidelines for fitting boundary lines. Based on the findings of this study, the censored bivariate normal model is recommended due to its objective nature and ability to quantify uncertainty in boundary line parameters. Including uncertainty information on parameters helps decision-makers assess risks, which is crucial for stakeholders investing in efforts to narrow yield gaps. This information should also stimulate the questions, Where is the uncertainty coming from? How bad is it? and, Can we do something about it?. Addressing these questions help in reducing uncertainty and thereby increasing confidence in the recommended actions. Further studies to incorporate Bayesian approaches to further refine uncertainty quantification and model robustness are needed.

While most stakeholders agreed that the censored bivariate normal model was statistically robust and provided for a more comprehensive interpretation of the boundary line model, they highlighted the inability of the function `cbvn()` within the `BLA R` package to incorporate categorical variables in its current form. Chapter 5 of this thesis addresses this concern through the development of a framework to incorporate categorical variables into the censored normal model. This modeling approach offers a flexible framework for analyzing biological responses constrained by factor-level limits. It is particularly relevant for applications in agronomy, ecology, and environmental science, where discrete environmental or management conditions often define the maximum achievable outcomes.

The stakeholder feedback workshop further highlighted a common issue in datasets, the frequent absence of one key input for the censored bivariate normal model which is the standard deviation of measurement error (σ_{me}). To prevent such gaps in datasets, it is crucial to consult a statistician during the study's planning phase. This ensures that study requirements, including data collection and analysis strategies, are properly defined, reducing the risk of missing essential data needed to estimate σ_{me} directly. As Fisher (1938) famously noted,

> *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."*

Nevertheless, if direct measurements of σ_{me} are unavailable, the `BLA R` package offers a σ_{me} maximum likelihood profiling-based estimation tool. Alternatively, if location data is available, geostatistical methods can be used, specifically by estimating the variogram and using the nugget effect as an upper bound approximation for σ_{me}.

The need for consultations with statisticians in the planning phase of the study was further reinforced from the results obtained from the comparison of boundary line

outputs and those determined experimentally in Chapter 4. The study revealed a misalignment between experimentally determined limiting factors (via omission trials) and those identified using boundary line analysis. This discrepancy underscores the importance of proper experimental design, particularly in selecting appropriate indicators of nutrient availability which are related to crop yield. From the results of this study, it can be argued that certain measures of soil N may not reliably indicate its availability to plants due to its dynamic nature. Consequently, the boundary line model of soil N may not always be sufficient. In addition, the small dataset size may have contributed to the misalignment between experimentally determined limiting factors and those identified through boundary line analysis. The dataset used in this study comprised only 148 observations. However, simulation results from Miti et al. (2024) suggest that at least 400 data points are needed to reliably fit a boundary model. Nevertheless, boundary line analysis remains a valuable diagnostic tool, though its interpretations should be supported by experimental validation whenever possible.

The lack of open source statistical tools for boundary line analysis was identified as one of the possible drivers of the inconsistencies in boundary line methodology. To address this, Chapter 6 detailed the development of an open-source `BLA R` package, which integrates the various boundary line fitting methods and data exploratory tools into an accessible platform for researchers. By consolidating these methods into a single package, this contribution enhances reproducibility and usability, particularly for researchers with limited statistical and programming backgrounds. A key challenge in statistical methods for fitting boundary models (e.g. censored bivariate normal model) is selecting appropriate initial starting values for the optimization process used to estimate model parameters. The `BLA R` package addresses this by providing a dedicated interactive tool for this process, representing one of several important contributions that could enhance the adoption of statistical methods for boundary line modeling and the application of boundary lines generally in agronomic research.

The development of the `BLA R` package for statistical boundary line analysis was grounded in an iterative and participatory process that prioritized user interaction and feedback. During the development phase, I conducted two stakeholder workshops in Kenya and Zimbabwe at IITA and CIMMYT respectively aimed at introducing the package, demonstrating its capabilities, and crucially, eliciting users' experiences, challenges, and expectations. Participants included agronomists, soil scientists, and data analysts actively engaged in yield gap studies, whose diverse perspectives provided practical insights that could not be gathered from technical development alone. These interactions uncovered usability challenges, identified gaps in documentation, and prompted refinements in functionality and user interface. Several suggestions from workshop participants were directly incorporated into the package, such as addition of functions for determination

of initial start values for optimization processes in boundary line model fitting, thereby making the tool more accessible and effective for a broader audience. The value of this engagement was further affirmed when I was invited to deliver a two-day postgraduate training course on yield gap analysis using the `BLA R` package at Wageningen University in the Netherlands and at the University of Zambia. These sessions fostered deeper discussions with researchers, helping to disseminate the methodological relevance of the tool and sparking ideas for future enhancements. Key questions that emerged included the sample size requirements for implementing boundary line analysis and whether the approach could be extended to capture lower boundaries within a dataset. The question on sample size requirement has been addressed in Chapters 3 and 5 of this thesis. A lower boundary response is a plausible occurrence in biological data. For example, the relationship between soil organic matter and clay content may exhibit a lower bound, as clay stabilizes organic matter and prevents its degradation below a certain threshold (Hassink, 1997). Such relationships highlight the need for boundary line analysis methods to accommodate lower boundaries, particularly in situations where the response variable is constrained from below by the predictor. While the boundary line fitting functions in the `BLA` R package are designed to fit upper boundaries, lower boundaries can still be analyzed by inverting the data, specifically, by multiplying the response variable by -1 and fitting the upper boundary. The resulting model can then be interpreted as representing a lower boundary in the original data space.

This experience from these workshops and postgraduate training courses highlighted that meaningful interaction with end users in the development of statistical procedures is essential not only for creating technically sound tools but also for ensuring their real-world applicability. Statistical methods, particularly those designed for domain-specific applications like agronomic yield gap analysis, must evolve through dialogue between developers and practitioners. Such collaboration helps to bridge the gap between theoretical innovation and practical utility, ensuring that methods are intuitive, transparent, and fit for purpose. As data-driven decision-making becomes increasingly central to agricultural and environmental research, fostering these feedback loops will be vital in designing robust, user-oriented statistical tools that can adapt to the complex realities of applied science.

The `BLA R` package has already gained significant traction, as evidenced by its download statistics from CRAN and ongoing engagements by users on the `BLA` GitHub repository's Issues page. Since its publication on 28$^{\text{th}}$ May, 2024, the `BLA R` package stands at over 6000 downloads as of 30$^{\text{th}}$ March, 2025 (`https://www.datasciencemeta.com/rpackages`, see also Appendix C.1). Following the publication of the `BLA R` package on CRAN, there has been engagement from users interested in its methods and applications. Notably, one user reached out to discuss the application of statistical boundary line methods from the package for application on agricultural data from Rwanda, reflecting

both interest in and the practical relevance of the approach. Such interactions highlight the package's potential for facilitating broader adoption of statistical methods for fitting boundary lines and underscore the importance of clear documentation and user support in the dissemination of research software.

The importance of open-source software in agriculture cannot be overstated especially for tools like the `BLA R` package. Open-source platforms promote transparency, reproducibility, and cost-effectiveness, empowering researchers and practitioners to tailor solutions to local contexts. Additionally, the collaborative nature of open-source development drives continuous improvement, ensuring that tools like the `BLA R` package remain at the forefront of agricultural innovation. This collaborative aspect is evident in the active engagement of users, who provide feedback and suggest improvements. For instance, some `BLA R` package users have raised issues and improvement suggestions with the functions `startValue()` and `ble_profile()` respectively via the Issues page on GitHub (see https://github.com/chawezimiti/BLA/issues), which have since been resolved.

7.2 Further Reflections on Data Quality, Metadata, and Open-access in Agronomic Studies

As with any statistical modelling approach, the accuracy and reliability of results from boundary line analysis for agronomic research are inherently dependent on both the quality and quantity of the data utilized. Sufficiently large datasets are necessary to capture the inherent variability in agronomic properties across different spatial and temporal scales. From a simulation exercise in Chapter 3, it showed that for reliable boundary line analysis results, at least 400 data points are required. Additionally, comprehensive metadata is essential for enhancing the interpretability of research findings. However, metadata are often neglected in agronomic studies (Kool et al., 2020), limiting the ability to derive meaningful insights from boundary line analysis results.

A common limitation in agronomic research is the selective reporting of metadata. While aspects such as nutrient and weed management, both of which are foundational to boundary line studies, are typically well-documented, other crucial factors such as land preparation practices, field history, and pest and water management are often omitted (Kool et al., 2020). These missing elements are significant as they influence yield-defining, yield-limiting, or yield-reducing factors. For instance, field history provides essential insights into the initial soil fertility status, as previous cropping patterns may either enhance or deplete soil nutrients, subsequently influencing the yield potential of the current crop (Falconnier et al., 2016). Similarly, land preparation details, including tillage methods, can affect soil structure, water retention, and aeration factors that play a crucial role in

early-season crop establishment and growth (Li et al., 2024). Moreover, while nutrient, water, pest, and weed management are frequently recorded in agronomic studies, they primarily influence the crop during the growing season, which is also the period most boundary line analyses focus on. However, without the context provided by historical and preparatory management practices, the interpretation of boundary line models remains incomplete. A comprehensive dataset that integrates all these factors is essential for identifying the underlying causes of yield gaps and for formulating precise and actionable agronomic recommendations.

Another critical but often overlooked component of agronomic metadata is the geo-location of study sites. Including spatial coordinates (longitude and latitude) in research data enhances the validity and applicability of findings by enabling spatial interpolation, facilitating site-specific analyses, and improving the reproducibility of studies (Walters et al., 2021). In boundary line studies, geo-location data can be particularly valuable for improving the estimation of measurement error of response variable, a crucial component of the censored bivariate normal model, using geostatistical techniques, especially in cases where direct error measurement is not available. Incorporating geo-spatial methodologies can strengthen the robustness of boundary line analyses and enhance their utility e.g. in precision agriculture and regional yield forecasting (Casanova et al., 1999).

The availability and accessibility of open source agronomic data remain a significant challenge in scientific research (da Cruz & do Nascimento, 2019). There is an urgent need for researchers to prioritize the publication of datasets on open-access platforms as open data facilitates knowledge sharing, enabling other researchers to build upon previous findings, conduct meta-analyses, and explore alternative interpretations that may not have been considered in the original studies (Yoong et al., 2022; Walters et al., 2021). For example, in our study to develop exploratory methods for boundary line analysis (Chapter 3) i.e. the Peel Cluster Method, extensive datasets were required to rigorously test the methodology. Similarly, our comparative study of different boundary line fitting techniques would have greatly benefited from access to diverse open-source datasets of varying sizes and conditions. The lack of publicly available data in boundary line research is a major constraint and this hinders methodological advancements.

Open-access datasets not only promote reproducibility and transparency in scientific research but also facilitate interdisciplinary collaborations, leading to novel insights and broader applications of agronomic findings (Ali & Dahlhaus, 2022; Allemang & Teegarden, 2017). A notable example is the ISRIC Soil Information System, which provides harmonized, gridded soil data at a global scale, enabling researchers to address critical data gaps and strengthen agronomic analyses (Hengl et al., 2017). Another example of an open-access data platform is the Centre for Agriculture and Bioscience International (CABI) Data Sharing Toolkit, developed in collaboration with the Open Data Institute

to improve access to information on soil health, agronomy, and fertilizers thereby contributing to enhanced food security in Sub-Saharan Africa and South Asia (CABI, 2021). The toolkit is built on the FAIR principles, ensuring that data are Findable, Accessible, Interoperable, and Reusable, and it aims to strengthen understanding of sound data-sharing practices and their potential benefits, including greater food security. While such open-access initiatives mark important progress toward data transparency and accessibility, the availability of open datasets for boundary line research remains limited. Despite the significant number of boundary line studies conducted to date, the scarcity of openly available data remains a major bottleneck for future research. Addressing this issue requires concerted efforts from the scientific community, funding agencies, and policymakers to establish data-sharing frameworks and incentives that encourage researchers to make their datasets publicly accessible.

Enhancing agronomic research requires improvements in data quality, comprehensive metadata reporting, and a shift toward open-access data sharing. By addressing these challenges, we can significantly improve the reliability of agronomic models like the boundary line models, facilitate knowledge transfer, and ultimately support the development of more effective and sustainable agricultural practices and analytical tools.

7.3 Thesis Limitations

Like any empirical study, this research faced several limitations. Although the overall objectives of the thesis were successfully achieved, certain methodological and data-related constraints inevitably influenced the scope and depth of the analyses. Recognising these limitations is important, as they highlight promising directions for future research aimed at strengthening boundary-line modelling and its broader applications.

One of the main challenges encountered was data availability, particularly the limited range of dataset sizes and diversity across crops and environments. Ideally, the study would have benefited from datasets of varying sizes to enable a more systematic evaluation of how sample size influences boundary line model performance and parameter stability. For instance, the analysis comparing the determination of the most limiting factor using boundary line and experimental approaches relied on a relatively small dataset, which may have constrained the robustness and generalisability of the findings. Future work incorporating datasets of different sizes could provide deeper insights into model sensitivity to data quantity and improve the reliability of estimates.

Although the datasets used in the other analyses were adequate for their intended purposes, the overall study would have benefited from greater diversity in terms of crop types, soil and environmental conditions, management practices, and geographic coverage. Such diversity would enhance the representativeness of the boundary line relationships

and broaden the applicability of the findings across agroecological contexts.

Another limitation relates to the planned extension of boundary line modelling within a Bayesian statistical framework. One of the initial objectives of this research was to integrate prior knowledge such as previously estimated parameter distributions or expert-informed constraints into the modelling process. This would have provided a coherent framework for uncertainty quantification, improved parameter estimation in the presence of limited or noisy data, and allowed the incorporation of prior agronomic understanding. However, achieving this objective was constrained by the limited availability of advanced training opportunities in Bayesian statistics at the University of Nottingham and through external programmes during the study period. Consequently, the Bayesian extension of the boundary line model could not be fully implemented within the scope of this thesis. Future research should therefore focus on developing and applying Bayesian boundary line models, exploring how prior information can be systematically incorporated to improve model robustness, interpretability, and predictive performance across diverse datasets and agroecological contexts.

In summary, the limitations identified in this thesis primarily relate to data availability, diversity, and the unrealised potential of a Bayesian extension. Nevertheless, these challenges provide a valuable foundation for future methodological advancement. Addressing them through larger and more diverse datasets, enhanced statistical training, and the integration of Bayesian approaches will help strengthen the empirical and theoretical underpinnings of boundary line modelling, ultimately improving its utility in agronomic research and yield-gap analysis.

Chapter 8

Conclusions and Future Direction

This thesis has made significant contributions to both the theoretical development and practical application of boundary line analysis in yield gap assessment. By addressing key methodological challenges, refining exploratory data analysis techniques, and developing the open-source `BLA R` package, this work advances the use of boundary line models in agronomic research. The new exploratory method introduced helps determine when boundary line analysis is appropriate, thereby improving the robustness of its application and encouraging critical reflection on its suitability case by case. Moreover, the extension of the censored normal model to accommodate categorical predictors expands its applicability to real-world agronomic datasets where such variables are common.

While the various boundary line fitting methods produced comparable results, the censored normal model is recommended due to its ability to support model-based inference, assess goodness of fit, and quantify uncertainty. These enhancements promote methodological standardization and increase reproducibility, strengthening the role of boundary line analysis in sustainable food production.

Future research should explore Bayesian extensions to the boundary line framework to improve parameter estimation and uncertainty quantification. Additionally, interdisciplinary collaborations especially among agronomists, data scientists, and policymakers are essential to ensure that the insights gained from boundary line analysis translate into actionable, context-sensitive agricultural recommendations. The work presented in this thesis thus provides a robust foundation for future advances in boundary line modeling, contributing to global efforts to close yield gaps and enhance food security through scientifically sound and data-driven approaches.

References

Abravan, P., Soltani, A., Majidian, M., & Mohsenabadi, G. (2016). Factors limiting canola yield and determining their optimum range by boundary line analysis. *Iioab Journal*, *7*(8), 161–167. Retrieved from https://www.iioab.org/articles/IIOABJ_7.8_161-167.pdf

AHDB. (2023). Nutrient Management Guide (RB209). Section 1: Arable Crops [Computer software manual]. Coventry, UK. Retrieved from https://ahdb.org.uk/documents/RB209/RB209_Section1_WEB_2017-12-01.pdf

Alexandratos, N., & Bruinsma, J. (2012). World agriculture towards 2030/2050: The 2012 revision ((food and agriculture organization of the united nations).

Ali, B., & Dahlhaus, P. (2022). The role of fair data towards sustainable agricultural performance: a systematic literature review. *Agriculture*, *12*(2), 309.

Allemang, D., & Teegarden, B. (2017). A global data ecosystem for agriculture and food. *F1000Research*, *6*(1844), 1844.

Andrade, C. B., Comin, J. J., Moura-Bueno, J. M., & Brunetto, G. (2023). Obtaining reference values for nutrients in vineyard soils through boundary line approach using Bayesian segmented quantile regression on commercial farm data. *European Journal of Agronomy*, *150*, 126928.

Baudron, F., Ndoli, A., Habarurema, I., & Silva, J. V. (2019). How to increase the productivity and profitability of smallholder rainfed wheat in the Eastern African highlands? Northern Rwanda as a case study. *Field Crops Research*, *236*, 121-131. doi: https://doi.org/10.1016/j.fcr.2019.03.023

Boote, K. J., Jones, J. W., & Pickering, N. B. (1996). Ceres-maize: A simulation model of maize growth. *Agricultural Systems*, *51*(3), 229–260.

CABI. (2021, March 25). *Data sharing toolkit could contribute to unlocking greater food security in sub-saharan africa and south asia.* https://www.cabi.org/news-article/data-sharing-toolkit-could-contribute-to-unlocking-greater

-food-security-in-sub-saharan-africa-and-south-asia/. (Accessed: 2025-10-13)

Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Front Ecol Environ*, *1*, 412-420. Retrieved from www.frontiersinecology.org

Casanova, D., Goudriaan, J., Bouma, J., & Epema, G. (1999). Yield gap analysis in relation to soil properties in direct-seeded flooded rice. *Geoderma*, *91*(3-4), 191–216. doi: https://doi.org/10.1016/S0016-7061(99)00005-1

Cassman, K. G., Dobermann, A., Walters, D. T., & Yang, H. (2003). Meeting cereal demand while protecting natural resources and improving environmental quality. *Annual Review of Environment and Resources*, *28*(1), 315–358.

Chatfield, C. (1985). The initial examination of data. *Journal of the Royal Statistical Society: Series A (General)*, *148*(3), 214–231.

Chen, G., Cao, H., Liang, J., Ma, W., Guo, L., Zhang, S., ... Zhang, F. (2018). Factors affecting nitrogen use efficiency and grain yield of summer maize on smallholder farms in the north china plain. *Sustainability*, *10*(2), 363.

da Cruz, S. M. S., & do Nascimento, J. A. P. (2019). Towards integration of data-driven agronomic experiments with data provenance. *Computers and Electronics in Agriculture*, *161*, 14–28.

Falconnier, G. N., Descheemaeker, K., Van Mourik, T. A., & Giller, K. E. (2016). Unravelling the causes of variability in crop yields and treatment responses for better tailoring of options for sustainable intensification in southern mali. *Field Crops Research*, *187*, 113–126.

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 Report*, *41*. Retrieved from https://digitalcommons.unl.edu/wffdocs

Fayazi, H., Zeinali, E., Soltani, A., & Torabi, B. (2022). Lentil genotypes selection for freezing tolerance by fluorescence chlorophyllestimating the potential and yield gap of forage maize in iran based on the global yield gap atlas protocol. *Iranian Journal of Field Crop Science*, *53*(3), 233–247.

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 Research*, *112*, 24-36. doi: https://doi.org/10.1016/j.fcr.2009.01.009

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 Research*, *112*(1), 24–36.

Fisher, R. A. (1938). Presidential Address to the First Indian Statistical Congress. *Sankhya*, *4*(14), e7.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., ... others (2005). Global consequences of land use. *Science*, *309*(5734), 570–574.

Gerber, J. S., Ray, D. K., Makowski, D., Butler, E. E., Mueller, N. D., West, P. C., ... others (2024). Global spatially explicit yield gap time trends reveal regions at risk of future crop yield stagnation. *Nature Food*, *5*(2), 125–135.

Grassini, P., Cassman, K. G., & Ittersum, M. v. (2017). Exploring maize intensification with the Global Yield Gap Atlas. *Better Crops with Plant Food*.

Guilpart, N., Grassini, P., Sadras, V. O., Timsina, J., & Cassman, K. G. (2017). Estimating yield gaps at the cropping system level. *Field Crops Research*, *206*, 21–32.

Hajjarpoor, A., Soltani, A., Zeinali, E., Kashiri, H., Aynehband, A., & Vadez, V. (2018). Using boundary line analysis to assess the on-farm crop yield gap of wheat. *Field Crops Research*, *225*, 64-73. doi: https://doi.org/10.1016/j.fcr.2018.06.003

Hassink, J. (1997). The capacity of soils to preserve organic c and n by their association with clay and silt particles. *Plant and Soil*, *191*, 77–87.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., ... others (2017). Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one*, *12*(2), e0169748.

Hochman, Z., Gobbett, D., Horan, H., & Garcia, J. N. (2016). Data rich yield gap analysis of wheat in australia. *Field Crops Research*, *197*, 97–106.

Ishikawa, S., Nakashima, T., Iizumi, T., & Hare, M. (2020). Evaluating irrigated rice yields in japan within the climate zonation scheme of the global yield gap atlas. *The Journal of Agricultural Science*, *158*(8-9), 718–729.

Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., et al. (2003). The dssat cropping system model. *European Journal of Agronomy*, *18*(3-4), 235–265.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., ... Hochman, Z. (2003). An overview of apsim, a model designed for farming systems simulation. *European Journal of Agronomy*, *18*, 267–288.

Kool, H., Andersson, J. A., & Giller, K. E. (2020). Reproducibility and external validity of on-farm experimental research in africa. *Experimental Agriculture*, *56*(4), 587–607. doi: https://doi.org/10.1017/S0014479720000174

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. *European Journal of Soil Science*, *71*, 334-351. doi: https://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. *European Journal of Soil Science*, *67*, 148-159. doi: https://doi.org/10.1111/ejss.12318

Lee, P. H., & Yu, P. L. (2013). An R package for analyzing and modeling ranking data. *BMC Medical Research Methodology*, *13*, 1–11.

Li, J., Yan, K., Duan, Q., Li, J., & Chen, Z. (2024). Effects of tillage practices on water storage and soil conservation in red soil slope farmland in southern china. *Scientific Reports*, *14*(1), 28781.

Makowski, D., Doré, T., & Monod, H. (2007). A new method to analyse relationships between yield components with boundary lines. *Agronomy for Sustainable Development*, *27*, 119-128. doi: https://doi.org/10.1051/agro:2006029

Milne, A. E., Wheeler, H. C., & Lark, R. M. (2006). On testing biological data for the presence of a boundary. *Annals of Applied Biology*, *149*, 213-222. doi: https://doi.org/10.1111/j.1744-7348.2006.00085.x

Miti, C., Milne, A., Giller, K., Sadras, V., & Lark, R. (2024). Exploration of data for analysis using boundary line methodology. *Computers and Electronics in Agriculture*, *219*, 108794.

Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., & Foley, J. A. (2012). Closing yield gaps through nutrient and water management. *Nature*, *490*(7419), 254–257.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons.

Nehbandani, A., Soltani, A., Hajjarpoor, A., Dadrasi, A., & Nourbakhsh, F. (2020). Comprehensive yield gap analysis and optimizing agronomy practices of soybean in Iran. *Journal of Agricultural Science*, *158*, 739-747. doi: https://doi.org/10.1017/S0021859621000241

Nezamzade, E., Soltani, A., Dastan, S., & Ajamnoroozi, H. (2020). View of Factors causing yield gap in rape seed production in the eastern of Mazandaran province, Iran. *Italian Journal of Agronomy*, *15*, 10-19. doi: https://doi.org/10.4081/ija.2020.1280

Omondi, J. O., Chiduwa, M. S., Kyei-Boahen, S., Masikati, P., & Nyagumbo, I. (2024). Yield gap decomposition: quantifying factors limiting soybean yield in southern africa. *npj Sustainable Agriculture*, *2*(1), 32.

Oteng-Darko, P., Yeboah, S., Addy, S., Amponsah, S., & Danquah, E. O. (2013). Crop modeling: A tool for agricultural research–a review. *Journal of Agricultural Research and Development*.

Phalan, B., Green, R. E., Dicks, L. V., Buchanan, G. M., Bradbury, C., Crowe, C. A., ... Donald, P. F. (2014). Closing yield gaps: perils and possibilities for biodiversity conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1639), 20120285. Retrieved from https://royalsocietypublishing.org/doi/10.1098/rstb.2012.0285 (Accessed: 2025-02-17) doi: 10.1098/rstb.2012.0285

Plantinga, A. J., & Irwin, E. G. (2017). Overview of empirical methods. In *Economics of rural land-use change* (pp. 129–150). Routledge.

Rattalino Edreira, J. I., Andrade, J. F., Cassman, K. G., van Ittersum, M. K., van Loon, M. P., & Grassini, P. (2021). Spatial frameworks for robust estimation of yield gaps. *Nature Food*, *2*(10), 773–779.

Schils, R., Olesen, J. E., Kersebaum, K.-C., Rijk, B., Oberforster, M., Kalyada, V., ... others (2018). Cereal yield gaps across europe. *European Journal of Agronomy*, *101*, 109–120.

Schnug, E., Heym, J. M., & Murphy, D. P. L. (1995). Boundary line determination technique (bolides). In P. C. Robert, R. H. Rust, & W. E. Larson (Eds.), *site specific management for agricultural systems* (p. 899-908). Wiley Online Library. doi: https://doi.org/10.2134/1995.site-specificmanagement.c66

Steduto, P., Hsiao, T., Fereres, E., & Raes, D. (2009). Aquacrop, a crop-water productivity model. *Agricultural Water Management*, *96*(4), 659–678.

Tittonell, P., & Giller, K. E. (2013). When yield gaps are poverty traps: The paradigm of ecological intensification in african smallholder agriculture. *Field Crops Research*, *143*, 76–90. doi: https://doi.org/10.1016/j.fcr.2012.10.007

Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). Springer.

United Nations. (2023). *World population surpasses 8 billion: What are the implications for planetary health and sustainability?* Retrieved from https://www.un.org/en/un-chronicle/world-population-surpasses-8-billion-what-are-implications-planetary-health-and (Accessed: 2025-02-17)

Van Asten, P., Wopereis, M., Haefele, S., Isselmou, M. O., & Kropff, M. (2003). Explaining yield gaps on farmer-identified degraded and non-degraded soils in a sahelian irrigated rice scheme. *NJAS-Wageningen Journal of Life Sciences*, *50*(3-4), 277–296.

van Bussel, L. G., Grassini, P., Van Wart, J., Wolf, J., Claessens, L., Yang, H., ... others (2015). From field to atlas: upscaling of location-specific yield gap estimates. *Field Crops Research*, *177*, 98–108.

Van Diepen, C. v., Wolf, J. v., Van Keulen, H., & Rappoldt, C. (1989). Wofost: a simulation model of crop production. *Soil Use and Management*, *5*(1), 16–24.

van Ittersum, M. K., Alimagham, S., Silva, J. V., Adjei-Nsiah, S., Baijukya, F. P., Bala, A., ... others (2025). Prospects for cereal self-sufficiency in sub-saharan africa. *Proceedings of the National Academy of Sciences*, *122*(24), e2423669122.

van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P., & Hochman, Z. (2013). Yield gap analysis with local to global relevance—a review. *Field Crops Research*, *143*, 4–17.

van Ittersum, M. K., & Rabbinge, R. (1997). Field crops research concepts in production ecology for analysis and quantification of agricultural input-output combinations. *Field Crops Research*, *52*, 197-208. doi: https://doi.org/10.1016/S0378-4290(97)00037-3

van Ittersum, M. K., Van Bussel, L. G., Wolf, J., Grassini, P., Van Wart, J., Guilpart, N., ... others (2016). Can sub-saharan africa feed itself? *Proceedings of the National Academy of Sciences*, *113*(52), 14964–14969.

Walters, J., Light, K., & Robinson, N. (2021). Using agricultural metadata: a novel investigation of trends in sowing date in on-farm research trials using the online farm trials database. *F1000Research*, *9*, 1305.

Webb, R. A. (1972). Use of the boundary line in the analysis of biological data. *Journal of Horticultural Science*, *47*(3), 309-319. doi: https://doi.org/10.1080/00221589.1972.11514472

Yoong, S. L., Turon, H., Grady, A., Hodder, R., & Wolfenden, L. (2022). The benefits of data sharing and ensuring open sources of systematic review data. *Journal of Public Health*, *44*(4), e582–e587.

Appendix A

BLA for yield gap analysis Workshop: Questionnaire Analysis

A.1 Introduction

This appendix presents the analysis from the questionnaire administered to participants of the Boundary Line Analysis workshops held in Zimbabwe and Kenya, aimed at eliciting their views on the various boundary line fitting methods and yield gap analysis approaches which were introduced in Chapter 4 of this thesis. This exercise was done to get information from stakeholders on which methods they preferred, how we could improve the boundary line methodology (implementation and interpretation) and preview the then prototype of the BLA R package for yield gap analysis.

A.2 Methodology

To achieve the outlined objectives, two workshops were organized at CGIAR centres in Kenya and Zimbabwe where participants with various specialities and levels of experience in statistical analysis were drawn from various institutions including International Institute of Tropical Agriculture (IITA), International Centre of Insect Physiology and Ecology (ICIPE), Alliance of Bioversity International and CIAT, Kenya Agricultural and Livestock Research (KALRO), University of Nairobi, French Agricultural Research Centre for International Development (CIRAD) and International Maize and Wheat Improvement Centre (CIMMYT). These were recruited by CGIAR research leads in Harare (CIMMYT) and Nairobi (IITA) who contacted potential users of the boundary line methodology for yield gap analysis. The criteria used to identify potential participants was that they work in research on agricultural crop production.

A comprehensive yield gap analysis was demonstrated to the participants using the

four boundary line fitting methods, binning, bolides, quantile regression and cbvn. **R** scripts for each of the boundary line fitting methods were provided and analysis was demonstrated. Participants were then given a separate dataset to carry out the yield gap analysis independently using the different boundary line fitting methods. After this exercise, an anonymous structured questionnaire was given to participants. Each questionnaire had a consent form attached for participants to indicate informed consent to participate in the study. The content of the consent form and Questionnaire can be found in Appendix B.

Items 1–3 of section 1 and items 1–2 of section 2 of the questionnaire solicited information on participants research areas, research focus, whether they are directly engaged with farmers, experience with statistical methods, previous experience on yield gap analysis and which methods were applied respectively. For these items, we reported absolute and relative frequencies as well as graphical representations using bar plots.

Items 3–9 of section 2 involved ranking the preferred boundary line fitting methods and getting more information on reason for ranking, agronomic interpretation of outputs, adequacy of outputs and whether the boundary line in general was sufficient for yield gap analysis. For item 3, the key analysis was to determine if there was uniformity of ranking for the different boundary line fitting methods (participants preferred method). To test the uniformity of ranking, a hypothesis test was conducted on the mean rank. Uniform ranking means that all possible rankings have the same probability of being observed. Under the null hypothesis of uniform ranking for k items, the expected value of the mean rank is given as:

$$\mu_{\text{rank}} = (k+1)/2, \tag{A.1}$$

and the evidence against this null hypothesis is measured using the test statistic:

$$\frac{12n}{k(k+1)} \sum_{i=1}^{k} \{m_i - \mu_{\text{rank}}\}^2, \tag{A.2}$$

where m_i is the mean rank of the i^{th} item, and n is the number of rankings. Under the null hypothesis, this statistic is distributed as χ^2 with $n-1$ degrees of freedom (Lee & Yu, 2013). Only those responses which had complete ranking provided by the respondent were considered for this test.

For items 6 and 9, the adequacy of output for interpretation for the different boundary line methods and the adequacy of boundary line methodology as a whole for yield gap analysis was analysed using absolute frequency. Information from items 1–3 of section 1 and items 1–2 of section 2 including the location, research area, research focus, past use of boundary line methods and statistical experience provided a framework to further

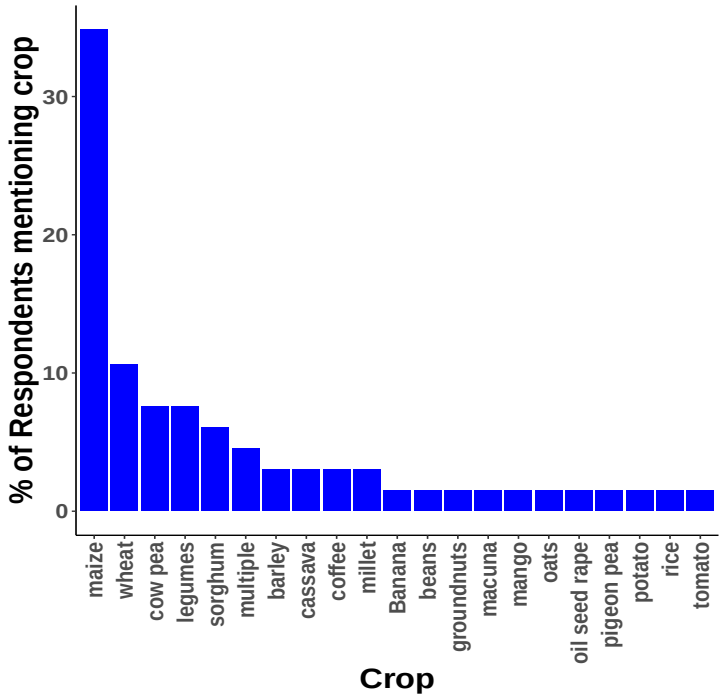
check if they influenced the ranking of boundary line fitting method if indeed there was non-uniformity in ranking.

A.3 Results and Discussion

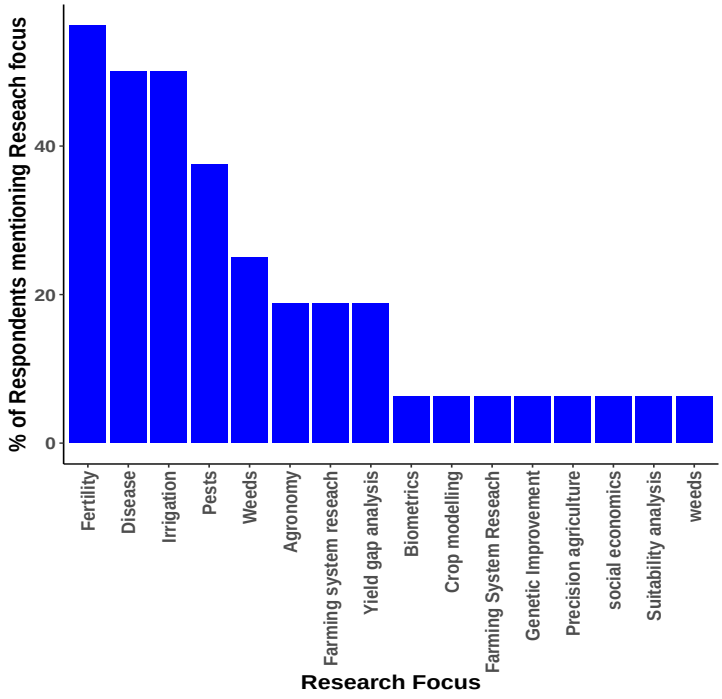
A total of 32 participants took part in the questionnaire exercise. All the 32 invited participants responded to the questionnaire corresponding to a 100 % response rate. Of these, 20 participants completed all questions corresponding to a completion rate of 62.5 %. The response rate for individual questions (items) was 100 %, except for items 3, 4, 5, 7 and 9 relating to ranking of boundary line methods (90.62 %), reasons for ranking (96.87 %), ease of interpreting method outputs (93.75 %), usefulness of uncertainty information (71.87 %) and sufficiency of the BLA for yield-gap analysis (93.75 %).

Figure A.1 and A.2 shows the characteristics of the participants of the survey from item 1–3 and 1 of sections 1 and 2 of the questionnaire respectively. Note that percent of respondents in Figure A.1(a) and A.1(b) represent the proportion of the total respondents that reported working with that particular crop and research focus respectively. This is because participants could reports multiple crops or research focus. A large portion of the participant were involved in maize research (about 30%). The majority of participants were involved in cereal research covering crops such as maize, wheat, sorghum, cassava, millet, and oats accounting for approximately 64% of respondents, while those working with legumes represented about 24%. Most participants focused their research on soil fertility (50%), Disease (48%), irrigation (48%) and pests (38%). Of these, 27 participants (87%) were directly engaged with farmers of which a total of 22 participants reported intermediate statistics experience while 6 and 4 reported advanced and beginner level of experience respectively (Figure A.2). These results indicate a good diversity among respondents in terms of the crops they worked with, their research focus, statistical experience, and experience with yield gap analysis, allowing for balanced opinions. A majority of the participants had prior experience in yield gap analysis, though not necessarily using the boundary line methodology (Figure A.2(c)). This, however, provided a valuable opportunity to obtain first impressions of the various boundary line models, including the one proposed in this study.

Item 3 of section 2 of the questionnaire relating to ranking of boundary line methods, the mean ranks for binning, bolides, cbvn and quantile regression were 3.17, 2.45, 2.31 and 2.07 respectively (Figure A.3). The test statistic ($\chi^2 = 2.71$, df= 3) showed that there was no evidence ($p = 0.56$) against the null hypothesis of uniform ranking. As the ranking of the boundary line fitting methods was uniform, we did not further split the data by location, research area, research focus, past use of boundary line methods or statistical experience for further analysis.

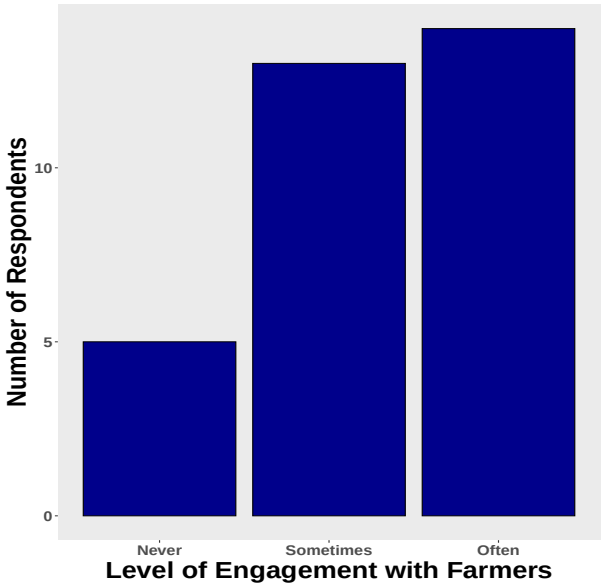


(a)

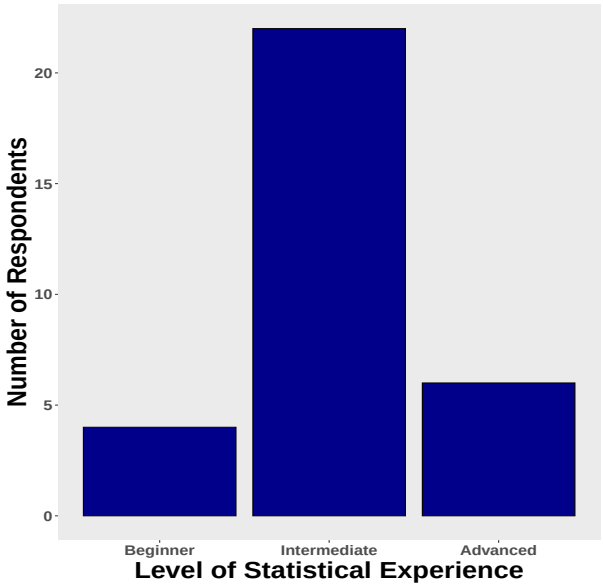


(b)

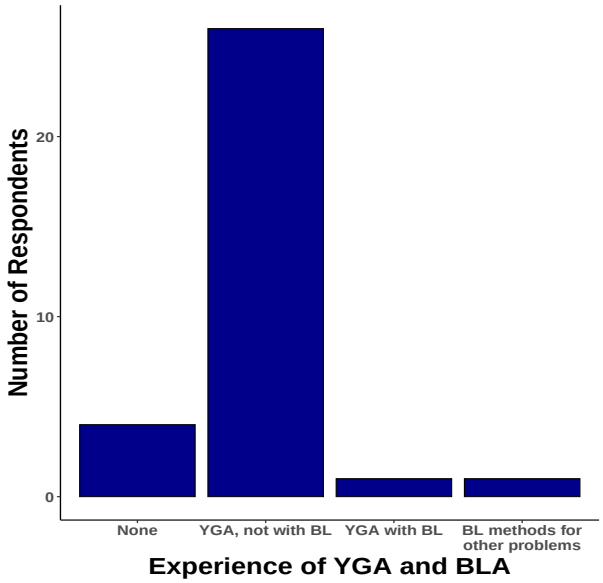
Figure A.1: Characteristics of participants including (a) area of research and (b) research focus



(a)



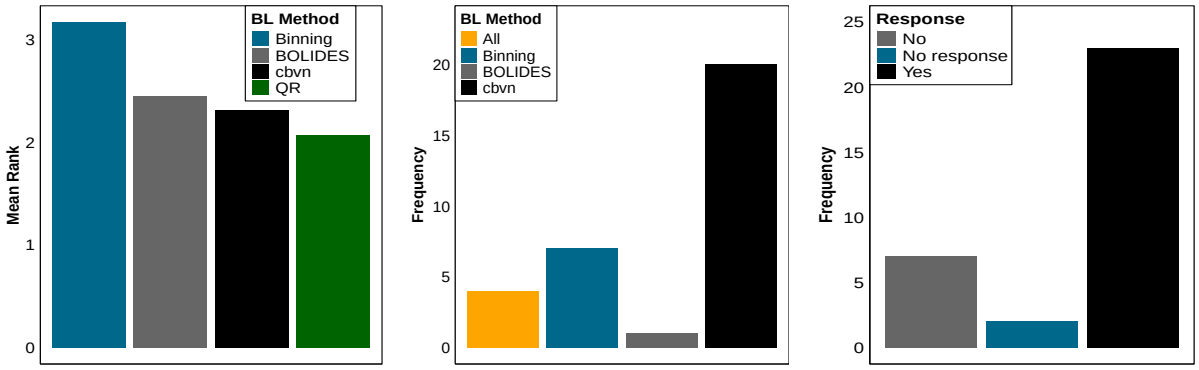
(b)



(c)

Figure A.2: Characteristics of participants including (a) engagement with farmers, (b) level of statistical experience (c) previous experience of yield gap analysis (YGA) and boundary line (BL) methods.

For item 6 of section 2 of the questionnaire relating to adequacy of output from the different boundary line fitting methods for yield gap analysis (Fig. A.3), 4 participants stated all methods, 7 binning, 1 bolides and 20 cbvn. Twenty three participants stated that information on uncertainty was an important output for interpretation while 9 did not give an opinion. For the adequacy of the boundary line methodology in general for yield gap analysis (Fig. A.3), 23 participants stated that it is adequate, 7 stated it is not adequate and 2 did not give an opinion.



(a) (b) (c)

Figure A.3: (a) Ranking of the boundary line fitting methods according to user-friendliness for all participants, (b) Adequacy of output for interpretation (c) Adequacy of boundary line methodology as a whole for yield gap analysis

Item 4 of section 2 of the questionnaire provided some useful information for the reasons for ranking by the participants. Some participants stated that the input parameters for the cbvn may be unavailable and difficult to estimate (especially measurement error) as reason for ranking it lower. However, the information on uncertainty of parameters was recognized as an important output for boundary lines using the cbvn by a majority of the participants (23) as indicated by item 7 of section 2. While more participants stated that the boundary line methodology was sufficient for yield gap analysis as a whole, some participants stated that it is inadequate as it does not incorporate the interaction effects on biological response, and in the cases of the cbvn method, cannot be applied to categorical variables including social economic factors.

A.4 Conclusion

According to the questionnaire responses, there was no clear preference for any particular boundary line fitting method, as indicated by the results of the uniformity test. This suggests that all methods are equally likely to be adopted for yield gap analysis by researchers across different disciplines, research focuses, locations, and levels of statistical

experience. Nonetheless, the feedback provided valuable insights into the interpretation and perceived adequacy of the boundary line methodology for yield gap analysis in general. These insights were instrumental in improving the functionality and outputs of the BLA R package, which was under active development at the time of the survey.

Appendix B

Consent form and Questionnaire for BLA workshop

Below is a concent form and questionnaire that was given to participants of the boundary line analysis workshop held in Zimbabwe and Kenya to elicit their views on the various boundary line fitting methods and yield gap analysis presented in Chapter 4.

University of Nottingham
UK | CHINA | MALAYSIA

Participant Consent Form

Name of Study: Boundary line methodology for yield gap analysis of farm systems

Name of Researcher(s): **Chawezi Miti, Murray Lark, Alice Milne and Ken Giller**

You are being invited to be involved in a research study. Taking part in this study involves participating in training in an R package for boundary line analysis, and then completing a survey questionnaire as a written form. In order to go forward with your participation, it is necessary for you to give consent.

Name of Participant:

By signing this form I confirm that (please initial the appropriate boxes)**:**	**Initials**
I have read and understood the Participant Information Sheet. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason, or withdrawing from the training if I want to continue with it	
I understand that no personal information about me will be collected, as the form is anonymous and is separated from this sheet when I hand it in.	
The word codes on this form (bottom right corner) and the questionnaire have been explained to me. I understand that (i) only the Chair of the University of Nottingham School of Biosciences holds the key to these codes and that (ii) the Chair will use them only if it is necessary to withdraw a questionnaire from the study if a consent form is missing, or if I write later to withdraw my consent.	
My words can be quoted in publications, reports, web pages and other research outputs.	
I give permission for the de-identified (anonymised) data that I provide to be used for future research and learning.	

I agree to take part in the study

_____	_____	_____
Name of Participant	Signature	Date
_____	_____	_____
Researcher's name	Signature	Date

2 copies: 1 for the participant, 1 for the project file

WORD CODE

Questionnaire

Section 1. Background		
1. Please describe your research area with a few words in each box below		
a. Crop(s) (e.g. maize, wheat, barley, tef, cassava, coffee)		
b. Research focus (e.g. nutrient deficiency, water relations and irrigation, diseases, pests, weeds, cultivations, rotations, economics etc).		
2. In the course of your work are you directly engaged with farmers? (PLEASE TICK A BOX BELOW)		
i. Never	ii. Sometimes	iii. Often
3. How would you rate your experience of statistical methods ? (PLEASE TICK A BOX BELOW)		
i. Beginner	ii. Intermediate	iii. Experienced

WORD CODE

Section 2. Use of Boundary Line Methods

1. What is your previous experience of yield gap analysis (YGA) and boundary line (BL) methods? (Please tick a box below)

i. None	ii. YGA, *not* with BL	iii. YGA with BL methods	iv. BL methods for other problems

2. If you have used BL methods, please indicate which method was used to fit the boundary (Please tick a box below)

i. Visual	ii. Binning	iii. BOLIDES	iv. Quantile regression	v. Stochastic frontier analysis	vi. Other (please describe below)

3. Rank the methods from the BLA package according to ease of fitting the model (Give 1 to the method you found most easy to use, 2 to the next-easiest, 4 to the least).

METHOD	RANKING (1,2, 3 or 4)
Binning	
BOLIDES	
Quantile Regression	
Censored bivariate normal model	

WORD CODE

4. Add comments on the reasons for your ranking below

5. Which method produced outputs that were easiest to interpret agronomically, and why?

6. Which methods (possibly more than one) gave adequate outputs to support a comprehensive yield gap analysis? Please add any comments.

WORD CODE

7. Do you think that statements about the uncertainty of boundary line parameters (e.g. the inflection point of a broken stick model) important? Please give brief reasons for your answer

8. Do you have suitable data and supporting information to use your preferred method (Q3) in your research environment?

9. Do you think that boundary line analysis in general suffices to support a comprehensive yield gap analysis? If not, then please indicate what additional information and methods you think are needed

WORD CODE

Appendix C

BLA R package downloads

Below is a plot showing the number of times the `BLA R` package has been downloaded per continent from the Comprehensive `R` Archive Network in a period of 10 months since it was published. There were some downloads which could not be identified from which country they cam from. These are labeled as Unknown.

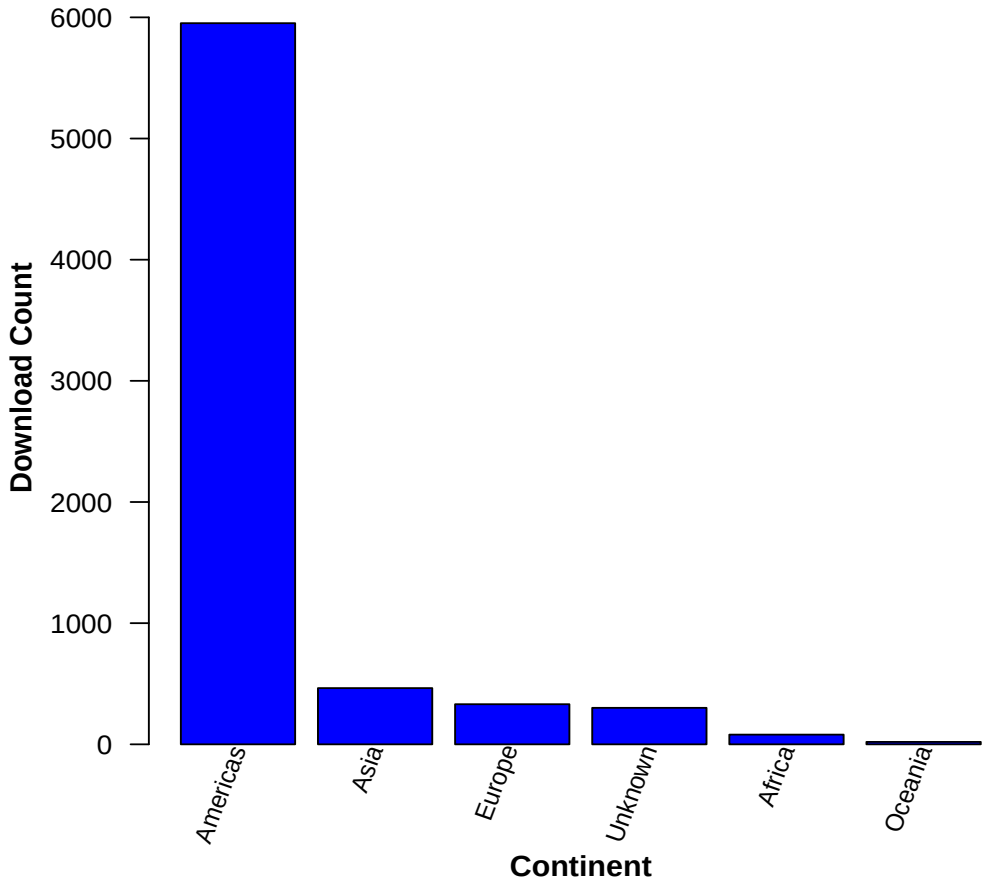


Figure C.1: Downloads of the `BLA R` package by continent, as recorded on CRAN since May 28, 2024.

Appendix D

Literature Review Methodology

This appendix provides the supplementary material for the article *"Concepts and Quantification of Yield Gaps Using Boundary Lines: A Review"*, presented in Chapter 2. It shows the methodology that was applied in the literature search.

Supplementary Material

Table S. 1: A Prisma protocol for the literature search for articles used in this paper

Checklist item	**Description**
Title	The concepts and quantification of yield gap using boundary lines: A review
Rationale	There are various methods used to fit boundary lines to data for use in yield gap analysis studies. These include the heusteric and statistical methods. However, there is no agreed standard procedure for setting out the boundary line. Currently, there is no information available in literature on how the use of different methods compare and its effects on yield gap analysis.
Objectives	The objectives of the review are to 1. gives an overview of the different agronomic interpretations of the boundary line that have been applied in yield gap analysis. 2. identify different approaches to fitting boundary line that are available for yield gap analysis and which ones are commonly used. 3. analyse to what extent the different approaches provide an objective way of fitting the boundary line, their weaknesses and strengths, factors influencing the choice of which approach to use and it's the impact on decomposition of yield gaps

Table S. 1: (Continued)

Checklist item	Description
Eligibility criteria	We included all studies that: 1. Evaluated yield gaps of agricultural crops using the boundary line methodology. We excluded all studies that: 1. Evaluated yield gaps of agricultural crops using other methodologies (not boundary line). 2. Review papers
Information sources	We conducted the search in the Web of science (`https://www.webofscience.com/wos/woscc/basic-search`) on 1^{st} June 2023. Google scholar was searched for articles that appeared in the reference list of the papers selected but did not appear in the initial search on web of science.
Search strategy	The following terms were searched in the web of science database (`https://www.webofscience.com/wos/woscc/basic-search`) as a Boolean "Yield gap*" AND "Boundary*"
Selection process	Studies that analysed yield gaps using any boundary line methodology were included in this study. Studies on yield gap analysis which used methods other than boundary line and review papers were excluded. One reviewer (CM) carried out the initial selection of articles and these were confirmed by three other reviewers (RML, AM and KG)
Data collection process	We designed a data extraction form based on the objectives of the study which one reviewer, (CM), used to extract data from eligible studies. These data were entered into an excel database. These data were verified by three other reviewers (RML, AM and KG)

Table S. 1: (Continued)

Checklist item	Description
Data items	We collected data on 1. The year of publishing 2. Domain of study 3. Crop studied. 4. Criteria for selection of factors used for yield gap analysis. 5. The agronomic interpretation of the boundary line 6. Criteria used to identify outliers. 7. The method used to estimate the boundary line and its classification i.e. whether it is based on a statistical model or heuristic method ('non-statistical') 8. The assumptions made for the boundary line method used (e.g., the percentile value assumed to be the boundary for binning methods).
Study selection	We found 70 articles in our data base search. We screened 70 articles from which 53 met the criteria and were reviewed. Later, we searched for articles that used boundary lines for yield gap analysis which were cited in the reviewed studies. Eleven articles were included in these searches. In total, 64 articles were used in this review.

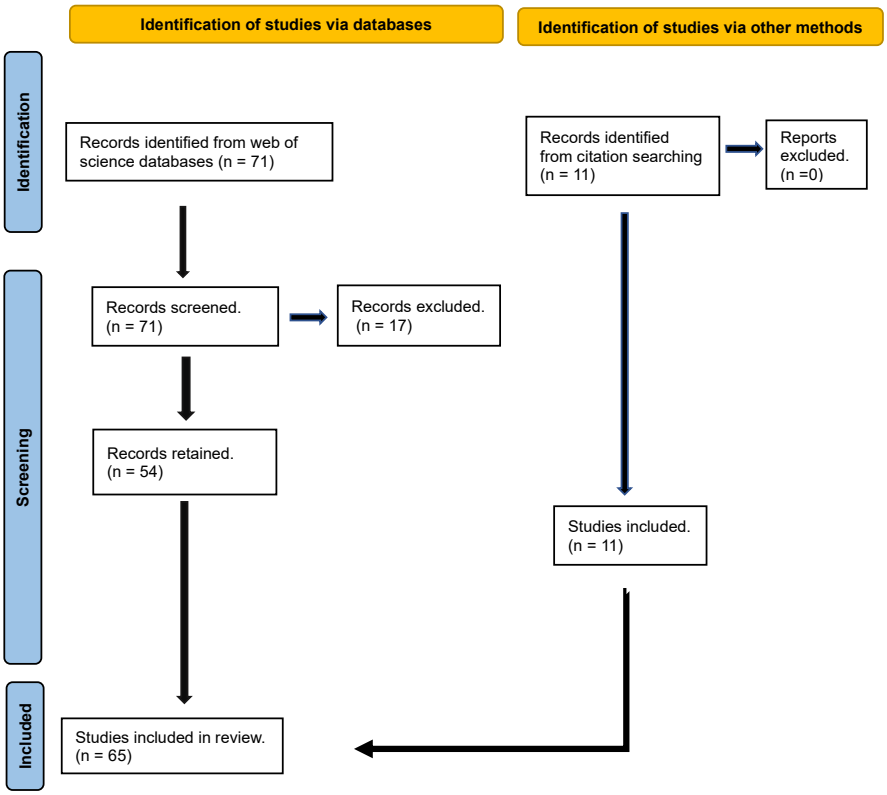


Fig S. 1: Flow diagram for the literature search using the prisma protocol

Table S. 2: Summary of studies that have evaluated yield gaps using boundary line (BL) methodology. BOLIDES is the boundary line determination tecnique, SFA is stochastic frontier analysis, QR is quantile regression and CM is the censored bivariate normal model methods

Article	**Country**	**Crop**	**BL Method**	Factors considered
Abravan et al. (2016)	Iran	Rapeseed	Visual	N, P, K, S, Plant density, sowing date, weed density, percent infected plants
Andrade et al. (2023)	Brazil	Grapes	QR	P, K, Ca, Mg, S, Cu, Zn, B, Mn
Affholder et al. (2013)	Senegal	Wheat,Maize, Millet, Rice	Binning	rainfall, check others
Gorjizad et al. (2019)	Iran	Rice	Visual	time of tillage operations, sowing date, transplanting date, seeding rate, seedling age, plant density, N, P and K, irrigation frequency, time of weed, disease and pest controls and harvesting date. Time of operations
Hajjarpoor et al. (2018)	Iran	Wheat	Visual	irrigation frequency, sowing date,seed rate, N rate, K and P
Baudron et al. (2019)	Ethiopia	Wheat	SFA /QR	Season, elevation, soil type, N application, P application, variety, seed rate, number of weeding operations, field size, planting date,fertilizer applied to previous crop
Bucagu et al. (2014)	Rwanda	Maize	BOLIDES	Organic carbon
Cao et al. (2019)	China	Wheat	BOLIDES	Plant density, N rate, P rate, K rate, sowing date
Casanova et al. (1999)	Spain	Rice	Binning	Soil texture, pH , EC, CEC, P, N, Na, K, Ca, Mg, Organic matter, chlorides, Sulphates, bicarbonates and nitrates
Dehkordi et al. (2020)	Iran	Wheat	Visual	NDVI
Tagliapietra et al. (2018)	Brazil	Soybean	Visual	Leaf area index, Maximum leaf area index, sowing dates
Evanylo et al. (1987)	USA	Soybeans	Binning	P, K, Mg, Ca, pH

Table S. 2: (Continued)

Article	**Country**	**Crop**	**BL Method**	Factors considered
Fermont et al. (2009)	East Africa	Cassava	BOLIDES	Soil pH, soil texture, weed management score, rainfall, available P, exchangeable K, SOC
French & Schultz (1984b)	Australia	Wheat	Visual	evapotranspiration
French & Schultz (1984a)	Australia	Wheat	Visual	evapotranspiration
Grassini et al. (2009)	USA	Maize	QR	Seasonal water supply, seasonal evapotranspiration
Chen et al. (2018)	China	Maize	BOLIDES	N rate, P rate, K rate, Plant density, sowing date, timing of irrigation, herbicide rate, insecticide rate, bactericide rate
Chen et al. (2019)	China	Maize	BOLIDES	Plant density, N rate, P rate, K rate, sowing date
Guo et al. (2021)	China	Maize	BOLIDES	Plant density, K, N, AWC, pH, EC
Haefele et al. (2003)	West Africa	Rice	Visual	Grain N, Grain P, Grain K
Fu et al. (2021)	China	Pear	BOLIDES	Organic matter, N,P,K,Ca,Mg, Plant density, pest control time, tree age and leaf N, P,K, Ca, Mg
Huang et al. (2008)	USA	Maize, Wheat, Soybean	Binning	wetness index
Silva et al. (2019)	Ethiopia	Wheat	SFA/BOLIDES	N fertilizer, P fertilizer, labour for ploughing, sowing and herbicide application
Kindred et al. (2015)	UK	Wheat	CM	Crop N demand, Soil N supply, Fertilizer N recovery
Kintché et al. (2017)	DRC	Cassava	Binning	SOC, pH, Total N, available P, exc K, CN ratio, exc Al, (Ca+Mg)/k
Lark et al. (2020)	UK	Wheat	CM	P, K, Mg, Ca, pH
Yousefian et al. (2021)	Iran	Rice	Visual	Transplant date, seedling age, Plant density, N, P, Harvest date, Pest, disease problem and weed problem
Wang et al. (2022)	China	Maize	QR	ET

Table S. 2: (Continued)

Article	**Country**	**Crop**	**BL Method**	Factors considered
Nehbandani et al. (2020)	Iran	Soybean	Visual	number of irrigations,sowing date,N rate and P consumed
Nezamzade et al. (2020)	Iran	Rapeseed	Visual	Field area, N,P,K, Pesticide, Herbicide, Pest, disease, Weed problem, lodging problem
Luo et al. (2020)	China	Maize	Binning	Plant density
Patrignani et al. (2014)	USA	Wheat	Binning	Growing season rainfall
Rhebergen et al. (2018)	Ghana	Oil palm	BOLIDES	year after planting
Sadras & Angus (2006)	Australia	Wheat	QR	evapotranspiration
Schnug et al. (1996)	Denmark	Rapeseed	BOLIDES	P, K, Soil pH, Clay
Bhattarai et al. (2017)	Costa Rica	Coffee	BOLIDES	Plant density, N application, Altitude, sunrise time,slope, pH, soil texture, (Ca+Mg)/K, Soil N, Soil P, Soil Fe
Shatar & McBratney (2004)	Australia	Wheat	Binning	Soil organic carbon, K, Ca, Mg, pH, Fe, P, water holding capacity
Silva et al. (2021)	Ethiopia	Wheat	SFA	seed rate, N rate, Herbicide rate, Weeding rate(persday/ha), labour use, field size, altitude, growing degree days, temperature, farming system, agro-ecological zone,variety, aridity index, soil water, soil depth
Silva et al. (2017)	Philipines	Rice	SFA	N ferilizer, P fertilizer, K fertilizer, seed rate, insecticide rate, herbicide rate, method of transplanting, variety, season, clay content, Sand content, SOC, available P, exch K
Ndabamenye et al. (2013)	East Africa	Banana	BOLIDES	pH,Mg,SOC,total N, P,Ca,Mg, K, Leaf area index, rainfall,soil water content, Plant density

Table S. 2: (Continued)

Article	**Country**	**Crop**	**BL Method**	Factors considered
Li et al. (2017)	China	Grape	Binning	soil water storage change
Asten et al. (2003)	Mauritania	Rice	Visual	EC, pH, Ca,P, ESP,Na, sowing density, N application
van Vugt & Franke (2018)	Malawi	Sweet potato	Binning	Altitude,planting date, total rainfall, rainfall exposure days, plant establishment
Wairegi et al. (2010)	East Africa	Banana	BOLIDES	Corn damage(%), N-total, K/(Ca + Mg), plant Population, Root necrosis (%), pH, SOM, soil clay content,Weeds pressure index, Fertilizer rate, Mulch and Rainfall
Walworth et al. (1986)	Worldwide	Maize	Binning	Leaf N, Leaf P, Leaf K
Wang et al. (2015)	Uganda	Coffee	BOLIDES	Pest and disease incidence, soil pH,SOC, soil P, Soil N, Soil K, Ca, Mg, elevation,tree density, tree age, mulch depth, shade tree density, hand weeding frequency
D. Zhang et al. (2019)	China	Mango	BOLIDES	Fertilise N, Fertilizer P, Fertilizer K, Plant density, Tree age, irrigation frequency, Plant height, planting scale, farmer age, experience
D. Zhang et al. (2019b)	China	Apple	BOLIDES	elevation, tree age, plant density,mulch grade, number, diameter and length of twigs, fruit weight, pest and disease incidence,number of sprays, Fertilizer N,P and K
Z. Zhang et al. (2020)	China	Rapeseed	BOLIDES	Fertilizer N, P and K
Niang et al. (2017)	West Africa	Rice	Spline	Total rainfall, total radiation
Hoogmoed et al. (2018)	Australia	Wheat	Binning	Nitrogen nutrition index
Lollato et al. (2017)	USA	Wheat	QR	Seasonal water supply, Seasonal evapotranspiration

Table S. 2: (Continued)

Article	**Country**	**Crop**	**BL Method**	Factors considered
Dossou-Yovo et al. (2020)	SSA	Rice	SFA	soil total N, soil clay content, soil pH, N fertilizer rate, K fertilizer application, P fertilizer application, agro-ecological zone
Wairegi et al. (2018)	Kenya	Coffee	Not stated	root necrosis, corn damage, Soil pH, SOC, Soil N, (Ca+Mg)/K, clay, mulch depth, weed pressure index, plant density, rainfall
Edreira et al. (2017)	USA	Soybean	QR	sowing date, seasonal water supply
Berrueta et al. (2020)	Uruguay	Tomato	BOLIDES	Cumulative PAR intercepted
Rizzo et al. (2021)	Uruguay	Soybean	QR	Seasonal rainfall
Fink et al. (2022)	USA	Alfalfa	QR	seasonal evapotranspiration
Baral, Bhandari, et al. (2022)	USA	Alfalfa	Visual	growing season rainfall
Baral, Lollato, et al. (2022)	USA	Alfalfa	Visual	growing season rainfall
Duan et al. (2022)	China	Apple	Not stated	Soil OC, pH, N, P and K, Fertilizer N, P and K, irrigation frequency, pruning times, pesticide spray times, pest and disease incidence, Tree age, plant density, number of bags per tree
Neuhaus & Sadras (2018)	Australia	Wheat	QR	Nitrogen nutrition index
Mohammadi-Kashka et al. (2023)	Iran	Soybean	Visual	Number of tillages, seedrate, irrigation frequency, farm size and Fertilizer N,P,K & S
Scarlato et al. (2017)	Uruguay	Strawberry	Binning	ground cover (%), number of leaves, number of crowns, plant density, date of planting

References

Abravan, P., Soltani, A., Majidian, M., & Mohsenabadi, G. (2016). Factors limiting canola yield and determining their optimum range by boundary line analysis. *Iioab Journal*, *7*(8), 161–167.

Affholder, F., Poeydebat, C., Corbeels, M., Scopel, E., & Tittonell, P. (2013). The yield gap of major food crops in family agriculture in the tropics: Assessment and analysis through field surveys and modelling. *Field Crops Research*, *143*, 106-118. doi: https://doi.org/10.1016/j.fcr.2012.10.021

Andrade, C. B., Comin, J. J., Moura-Bueno, J. M., & Brunetto, G. (2023). Obtaining reference values for nutrients in vineyard soils through boundary line approach using Bayesian segmented quantile regression on commercial farm data. *European Journal of Agronomy*, *150*, 126928.

Asten, P. J. V., Wopereis, M. C., Haefele, S., Isselmou, M. O., & Kropff, M. J. (2003). Explaining yield gaps on farmer-identified degraded and non-degraded soils in a Sahelian irrigated rice scheme. *Netherlands Journal of Agricultural Science*, *50*, 277-296. doi: https://doi.org/10.1016/s1573-5214(03)80013-1

Baral, R., Bhandari, K., Kumar, R., & Min, D. (2022). Yield gap analysis of alfalfa grown under rainfed condition in Kansas. *Agronomy*, *12*(9), 2190.

Baral, R., Lollato, R. P., Bhandari, K., & Min, D. (2022). Yield gap analysis of rainfed alfalfa in the United States. *Frontiers in Plant Science*, 2492.

Baudron, F., Ndoli, A., Habarurema, I., & Silva, J. V. (2019). How to increase the productivity and profitability of smallholder rainfed wheat in the Eastern African highlands? Northern Rwanda as a case study. *Field Crops Research*, *236*, 121-131. doi: https://doi.org/10.1016/j.fcr.2019.03.023

Berrueta, C., Heuvelink, E., Giménez, G., & Dogliotti, S. (2020). Estimation of tomato yield gaps for greenhouse in Uruguay. *Scientia Horticulturae*, *265*, 109250.

Bhattarai, S., Alvarez, S., Gary, C., Rossing, W., Tittonell, P., & Rapidel, B. (2017). Combining farm typology and yield gap analysis to identify major variables limiting yields in the highland coffee systems of Llano Bonito, Costa Rica. *Agriculture, Ecosystems and Environment*, *243*, 132-142. doi: https://doi.org/10.1016/j.agee.2017.04.016

Bucagu, C., Vanlauwe, B., Van Wijk, M. T., & Giller, K. E. (2014). Resource use and food self-sufficiency at farm scale within two agro-ecological zones of Rwanda. *Food security*, *6*(5), 609–628. doi: https://doi.org/10.1007/s12571-014-0382-0

Cao, H., Li, Y., Chen, G., Chen, D., Qu, H., & Ma, W. (2019). Identifying the limiting factors driving the winter wheat yield gap on smallholder farms by agronomic diagnosis in North

China Plain. *Journal of Integrative Agriculture*, *18*, 1701-1713. doi: https://doi.org/10.1016/S2095-3119(19)62574-8

Casanova, D., Goudriaan, J., Bouma, J., & Epema, G. (1999). Yield gap analysis in relation to soil properties in direct-seeded flooded rice. *Geoderma*, *91*(3-4), 191–216. doi: https://doi.org/10.1016/S0016-7061(99)00005-1

Chen, G., Cao, H., Chen, D., Zhang, L., Zhao, W., Zhang, Y., ... Zhang, F. (2019). Developing sustainable summer maize production for smallholder farmers in the North China Plain: An agronomic diagnosis method. *Journal of Integrative Agriculture*, *18*, 1667-1679. doi: https://doi.org/10.1016/S2095-3119(18)62151-3

Chen, G., Cao, H., Liang, J., Ma, W., Guo, L., Zhang, S., ... Zhang, F. (2018). Factors affecting nitrogen use efficiency and grain yield of summer maize on smallholder farms in the North China Plain. *Sustainability*, *10*(2), 363.

Dehkordi, P. A., Nehbandani, A., Hassanpour-bourkheili, S., & Kamkar, B. (2020). Yield gap analysis using remote sensing and modelling approaches: Wheat in the Northwest of Iran. *International Journal of Plant Production*, *14*, 443-452. doi: https://doi.org/10.1007/s42106-020-00095-4

Dossou-Yovo, E. R., Vandamme, E., Dieng, I., Johnson, J.-M., & Saito, K. (2020). Decomposing rice yield gaps into efficiency, resource and technology yield gaps in sub-Saharan Africa. *Field Crops Research*, *258*, 107963.

Duan, Z., Zheng, C., Zhao, S., Feyissa, T., Merga, T., Jiang, Y., & Zhang, W. (2022). Cold climate during bud break and flowering and excessive nutrient inputs limit apple yields in Hebei Province, China. *Horticulturae*, *8*(12), 1131.

Edreira, J. I. R., Mourtzinis, S., Conley, S. P., Roth, A. C., Ciampitti, I. A., Licht, M. A., ... others (2017). Assessing causes of yield gaps in agricultural areas with diversity in climate and soils. *Agricultural and Forest Meteorology*, *247*, 170–180.

Evanylo, G. K., Sumner, . M. E., & Francis, T. . (1987). Utilization of the boundary line approach in the development of soil nutrient norms for soybean production. *Commun. in Soil Sci. Plant Anal.*, *18*, 1397-1401. doi: https://doi.org/10.1080/00103628709367906

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 Research*, *112*, 24-36. doi: https://doi.org/10.1016/j.fcr.2009.01.009

Fink, K. P., Grassini, P., Rocateli, A., Bastos, L. M., Kastens, J., Ryan, L. P., ... Lollato, R. P. (2022). Alfalfa water productivity and yield gaps in the US Central Great Plains. *Field Crops Research*, *289*, 108728.

French, R. J., & Schultz, J. E. (1984a). Water Use Efficiency of Wheat in a Mediterranean-type Environment 1. The Relation between Yield, Water Use and Climate. *J. Agric. Res*, *35*, 743-64.

French, R. J., & Schultz, J. E. (1984b). Water Use Efficiency of Wheat in a Mediterranean-type Environment 2. Some Limitations to Efficiency. *Aust. J. Agric. Res*, *35*, 765-75.

Fu, H., Ma, Q., Ma, Z., Hu, Y., Liu, F., Chen, K., ... Wu, L. (2021). Quantifying key internal and external yield-limiting factors for Chinese pear in smallholder dominant areas. *HortScience*, *56*, 1395-1401. doi: https://doi.org/10.21273/HORTSCI16115-21

Gorjizad, A., Dastan, S., Soltani, A., & Norouzi, H. A. (2019). Large scale assessment of the production process and rice yield gap analysis by comparative performance analysis and boundary-line analysis methods. *Italian Journal of Agronomy*, *14*, 123-131. doi: https://doi.org/10.4081/ija.2019.1174

Grassini, P., Yang, H., & Cassman, K. G. (2009). Limits to maize productivity in Western Corn-Belt: A simulation analysis for fully irrigated and rainfed conditions, volume = 149. *Agricultural and Forest Meteorology*, 1254-1265. doi: https://doi.org/10.1016/j.agrformet.2009.02.012

Guo, X., Shukla, M. K., Wu, D., Chen, S., Li, D., & Du, T. (2021). Plant density, irrigation and nitrogen management: three major practices in closing yield gaps for agricultural sustainability in North-West China. *Frontiers of Agricultural Science and Engineering*, *8*(4), 525–544.

Haefele, S., Wopereis, M., Ndiaye, M., Barro, S., & Isselmou, M. O. (2003). Internal nutrient efficiencies, fertilizer recovery rates and indigenous nutrient supply of irrigated lowland rice in Sahelian West Africa.

Hajjarpoor, A., Soltani, A., Zeinali, E., Kashiri, H., Aynehband, A., & Vadez, V. (2018). Using boundary line analysis to assess the on-farm crop yield gap of wheat. *Field Crops Research*, *225*, 64-73. doi: https://doi.org/10.1016/j.fcr.2018.06.003

Hoogmoed, M., Neuhaus, A., Noack, S., & Sadras, V. O. (2018). Benchmarking wheat yield against crop nitrogen status. *Field Crops Research*, *222*, 153–163.

Huang, X., Wang, L., Yang, L., & Kravchenko, A. N. (2008). Management effects on relationships of crop yields with topography represented by wetness index and precipitation. *Agronomy Journal*, *100*(5), 1463–1471. doi: https://doi.org/10.2134/agronj2007.0325

Kindred, D. R., Milne, A. E., Webster, R., Marchant, B. P., & Sylvester-Bradley, R. (2015). Exploring the spatial variation in the fertilizer-nitrogen requirement of wheat within fields. *Journal of Agricultural Science*, *153*, 25-41. doi: https://doi.org/10.1017/S0021859613000919

Kintché, K., Hauser, S., Mahungu, N. M., Ndonda, A., Lukombo, S., Nhamo, N., ... 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. *European Journal of Agronomy*, *89*, 107-123. doi: https://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. *European Journal of Soil Science*, *71*, 334-351. doi: https://doi.org/10.1111/ejss.12891

Li, T., Hao, X., & Kang, S. (2017). Spatial variability of grape yield and its association with soil water depletion within a vineyard of arid North West China. *Agricultural Water Management*, *179*, 158-166. doi: https://doi.org/10.1016/j.agwat.2016.05.006

Lollato, R. P., Edwards, J. T., & Ochsner, T. E. (2017). Meteorological limits to winter wheat productivity in the US Southern Great Plains. *Field Crops Research*, *203*, 212–226.

Luo, N., Wang, X., Hou, J., Wang, Y., Wang, P., & Meng, Q. (2020). Agronomic optimal plant density for yield improvement in the major maize regions of China. *Crop Science*, *60*(3), 1580–1590. doi: https://doi.org/10.1002/csc2.20000

Mohammadi-Kashka, F., Pirdashti, H., Tahmasebi-Sarvestani, Z., Motevali, A., Nadi, M., & Aghaeipour, N. (2023). Integrating life cycle assessment (LCA) with boundary line analysis (bla) to reduce agro-environmental risk of crop production: a case study of soybean production in Northern Iran. *Clean Technologies and Environmental Policy*, 1–20.

Ndabamenye, T., Asten, P. J. V., Blomme, G., Vanlauwe, B., Uzayisenga, B., Annandale, J. G., & Barnard, R. O. (2013). Nutrient imbalance and yield limiting factors of low input East African highland banana (Musa spp. AAA-EA) cropping systems. *Field Crops Research*, *147*, 68-78. doi: https://doi.org/10.1016/j.fcr.2013.04.001

Nehbandani, A., Soltani, A., Hajjarpoor, A., Dadrasi, A., & Nourbakhsh, F. (2020). Comprehensive yield gap analysis and optimizing agronomy practices of soybean in Iran. *Journal of Agricultural Science*, *158*, 739-747. doi: https://doi.org/10.1017/S0021859621000241

Neuhaus, A., & Sadras, V. O. (2018). Relationship between rainfall-adjusted nitrogen nutrition index and yield of wheat in Western Australia. *Journal of Plant Nutrition*, *41*(20), 2637–2643.

Nezamzade, E., Soltani, A., Dastan, S., & Ajamnoroozi, H. (2020). View of Factors causing yield gap in rape seed production in the East of Mazandaran Province, Iran. *Italian Journal of Agronomy*, *15*, 10-19. doi: https://doi.org/10.4081/ija.2020.1280

Niang, A., Becker, M., Ewert, F., Dieng, I., Gaiser, T., Tanaka, A., ... others (2017). Variability and determinants of yields in rice production systems of West Africa. *Field Crops Research*, *207*, 1–12.

Patrignani, A., Lollato, R. P., Ochsner, T. E., Godsey, C. B., & Edwards, J. T. (2014). Yield gap and production gap of rainfed winter wheat in the Southern Great Plains. *Agronomy Journal*, *106*(4), 1329–1339. doi: https://doi.org/10.2134/agronj14.0011

Rhebergen, T., Fairhurst, T., Whitbread, A., Giller, K. E., & Zingore, S. (2018). Yield gap analysis and entry points for improving productivity on large oil palm plantations and smallholder farms in Ghana. *Agricultural Systems*, *165*, 14-25. doi: https://doi.org/10.1016/j.agsy.2018.05.012

Rizzo, G., Monzon, J. P., & Ernst, O. (2021). Cropping system-imposed yield gap: Proof of concept on soybean cropping systems in Uruguay. *Field Crops Research*, *260*, 107944.

Sadras, V. O., & Angus, J. F. (2006). Benchmarking water-use efficiency of rainfed wheat in dry environments. *Australian Journal of Agricultural Research*, *57*, 847-856. doi: https://doi.org/10.1071/AR05359

Scarlato, M., Giménez, G., Lenzi, A., Borges, A., Bentancur, Ó., Dogliotti, S., et al. (2017). Analysis and hierarchization of factors explaining strawberry cultivation yield gap in Southern Uruguay. *Agrociencia (Montevideo)*, *21*(1), 43–57.

Schnug, E., Heym, J. M., & Murphy, D. P. L. (1996). *Boundary line determination technique*. Retrieved from `https://www.researchgate.net/publication/282843591`

Shatar, T. M., & McBratney, A. B. (2004). Boundary-line analysis of field-scale yield response to soil properties. *Journal of Agricultural Science*, *142*, 553-560. doi: https://doi.org/10.1017/S0021859604004642

Silva, J. V., Baudron, F., Reidsma, P., & Giller, K. E. (2019). Is labour a major determinant of yield gaps in sub-Saharan Africa? A study of cereal-based production systems in Southern Ethiopia. *Agricultural Systems*, *174*, 39-51. doi: https://doi.org/10.1016/j.agsy.2019.04.009

Silva, J. V., Reidsma, P., Baudron, F., Jaleta, M., Tesfaye, K., & van Ittersum, M. K. (2021). Wheat yield gaps across smallholder farming systems in Ethiopia. *Agronomy for Sustainable Development*, *41*. doi: https://doi.org/10.1007/s13593-020-00654-z

Silva, J. V., Reidsma, P., Laborte, A. G., & van Ittersum, M. K. (2017). Explaining rice yields and yield gaps in Central Luzon, Philippines: An application of stochastic frontier analysis and crop modelling. *European Journal of Agronomy*, *82*, 223-241. doi: https://doi.org/10.1016/j.eja.2016.06.017

Tagliapietra, E. L., Streck, N. A., da Rocha, T. S. M., Richter, G. L., da Silva, M. R., Cera, J. C., ... Zanon, A. J. (2018). Optimum leaf area index to reach soybean yield potential in subtropical environment. *Agronomy Journal*, *110*(3), 932–938. doi: https://doi.org/10.2134/agronj2017.09.0523

van Vugt, D., & Franke, A. C. (2018). Exploring the yield gap of orange-fleshed sweet potato varieties on smallholder farmers' fields in Malawi. *Field Crops Research*, *221*, 245-256. doi: https://doi.org/10.1016/j.fcr.2017.11.028

Wairegi, L. W., Bennett, M., Nziguheba, G., Mawanda, A., de los Rios, C., Ampaire, E., ... van Asten, P. J. (2018). Sustainably improving Kenya's coffee production needs more participation of younger farmers with diversified income. *Journal of Rural Studies*, *63*, 190–199.

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 Research*, *117*, 146-153. doi: https://doi.org/10.1016/j.fcr.2010.02.010

Walworth, J., Letzsch, W., & Sumner, M. (1986). Use of boundary lines in establishing diagnostic norms. *Soil Science Society of America Journal*, *50*(1), 123–128.

Wang, N., Chen, H., Ding, D., Zhang, T., Li, C., Luo, X., ... Siddique, K. H. (2022). Plastic film mulching affects field water balance components, grain yield, and water productivity of rainfed maize in the Loess Plateau, China: A synthetic analysis of multi-site observations. *Agricultural Water Management*, *266*, 107570. doi: https://doi.org/10.1016/j.agwat.2022.107570

Wang, N., Jassogne, L., van Asten, P. J., Mukasa, D., Wanyama, I., Kagezi, G., & Giller, K. E. (2015). Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. *European Journal of Agronomy*, *63*, 1-11. doi: https://doi.org/10.1016/j.eja.2014.11.003

Yousefian, M., Soltani, A., Dastan, S., & Ajamnoroozie, H. (2021). Yield gap assessment in rice-grown fields using CPA and BLA approaches in Northern Iran. *International Journal of Plant Production*, *15*, 203-217. doi: https://doi.org/10.1007/s42106-020-00128-y

Zhang, D., Wang, C., & Li, X. (2019). Yield gap and production constraints of mango (Mangifera indica) cropping systems in Tianyang County, China. *Journal of Integrative Agriculture*, *18*, 1726-1736. doi: https://doi.org/10.1016/S2095-3119(18)62099-4

Zhang, D., Wang, C., Li, X., Yang, X., Zhao, L., & Xia, S. (2019b). Correlation of production constraints with the yield gap of apple cropping systems in Luochuan County, China. *Journal of Integrative Agriculture*, *18*(8), 1714–1725.

Zhang, Z., Cong, R., Ren, T., Li, H., Zhu, Y., & Lu, J. (2020). Optimizing agronomic practices for closing rapeseed yield gaps under intensive cropping systems in China. *Journal of Integrative Agriculture*, *19*, 1241-1249. doi: https://doi.org/10.1016/S2095-3119(19)62748-6