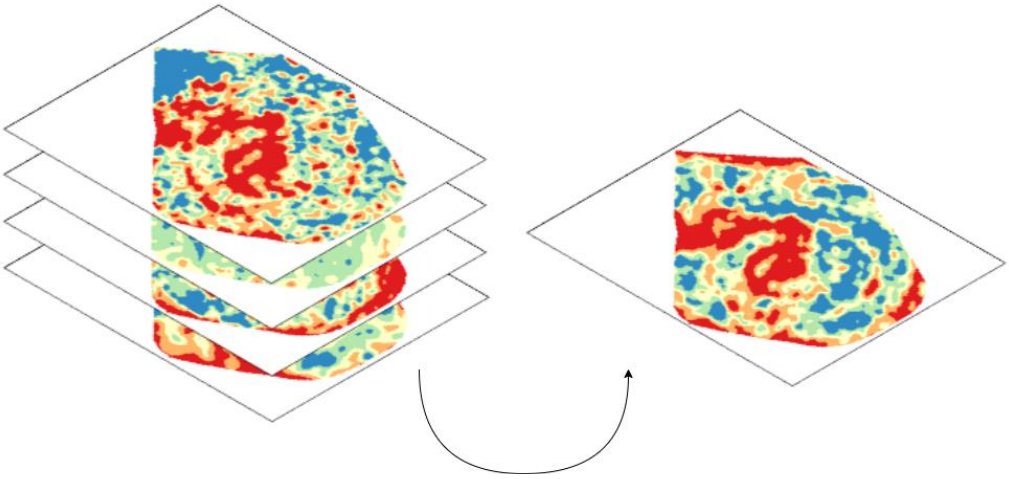


Filling the maize yield gap based on precision agriculture – A Maxent approach

Marcos Alexandre Mota Norberto
Mestrado em Engenharia Agronómica
Departamento de Geociências, Ambiente e Ordenamento do Território
2021

Orientador
Doutor Mário Manuel de Miranda Furtado Campos Cunha, Professor Associado, Faculdade de Ciências da Universidade do Porto

Co-Orientador
Doutor Neftalí Sillero Pablos, Investigador Principal, Centro de Investigação em Ciências Geo-Espaciais (CICGE), Faculdade de Ciências da Universidade do Porto

Supervisor
Engenheiro João Coimbra, Administrador da Quinta da Cholda

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri,
e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/_____

“Out of the night that covers me,
Black as the pit from pole to pole,
I thank whatever gods may be
For my unconquerable soul.

In the fell clutch of circumstance
I have not winced nor cried aloud.
Under the bludgeonings of chance
My head is bloody, but unbowed.

Beyond this place of wrath and tears
Looms but the Horror of the shade,
And yet the menace of the years
Finds and shall find me unafraid.

It matters not how strait the gate,
How charged with punishments the scroll,
I am the master of my fate,
I am the captain of my soul.”

Invictus by *William Henley*

Agradecimentos

Desde ao momento em que decidi ingressar no mestrado em Engenharia Agronómica e até ao seu término, houve um conjunto de pessoas que foram essenciais nas diferentes etapas, e espero desta forma honrá-las pelo contributo que tiveram.

Ao meu Orientador, o Professor Mário Cunha, pelo "insight" que teve durante o desenvolvimento de toda a tese. Sem o seu entusiasmo, disponibilidade, visão e orientação nos momentos certos, dificilmente esta tese veria a luz do dia.

Ao Professor Neftalí Sillero, por ter disponibilizado o espaço e equipamentos do CICGE, por me ter transmitido uma área de conhecimento desconhecida por mim, e pelo auxilio que forneceu ao longo da dissertação.

Ao Engenheiro João Coimbra e á AgroAnalítica, pelo desafio proposto, por me ter disponibilizado um conjunto de dados que em condições normais eu não teria tido acesso, e por toda sua disponibilidade para me retirar dúvidas.

To Daria Lowicz, for having the patient in talking to me at the right moments in a particular hard time in my life, and for being a good friend.

To the Moore family and Sonia in particular, for the warm embrace that I received during my stay in the UK. Truly, the time spent with you all was one of the most amazing learning experiences in my life.

Ao P.E.S.T Lab (Joana Neto, Leonor "Chicharro", Diogo Saraiva, António "Toy" Neto-Parra, Pedro Sousa e Leandro Rodrigues), por toda o companheirismo e camaradagem que se criou, o meu muito obrigado. Partilhar as nossas dores com outras pessoas que fazem o mesmo percurso torna esta viagem suportável.

Á Patrícia Machado, pela boa companhia nas longas tardes de trabalho, riso, e boas conversas com umas cervejinhas á mistura.

Aos meus amigos de longa data, Nuno Fernandes, Duarte Teixeira , Vasco Gouveia , Susana Teixeira e Jéssica Correia, obrigado pela amizade ao longo de todos estes anos.

E finalmente á minha família. Sem eles nada disto seria possível. Ao meu irmão, Jorge Norberto, e cunhada, Raquel Monteiro, em se mostrarem disponíveis em me receber sempre que precisei para espairecer, pensar e trabalhar. Aos meus sobrinhos, João e Sara Norberto, por me fazerem companhia e um esforço por perceber a minha tese. E finalmente á Dona Manuela, a minha mãe. Por perceber quando eu estava particularmente de mau humor, em me motivar sempre que precisei, em me mandar "ir correr" (literalmente) para espairecer.

A todos vós, o meu profundo obrigado.

Abstract

Precision Agriculture (PA) is seen as one of the European Union's strategies to achieve the goal of producing more with less, in a world with fewer available resources. The application of PA techniques in the collection of information from maize allows detecting the existence of spatiotemporal variations in the fields, but traditional approaches cannot deal with the multidimensionality of the information to determine the main effects of these variations on productivity. A yield gap approach is a promising strategy to assess the available biophysical potential of a specific location by comparing a potential yield with the field average yield. Following this approach, this work hypothesized that the edaphic characteristics that cause consistent high and low productivity patterns in time and space can be interpreted as an ecological niche. In light of this interpretation, this work sought to apply an ecological niche model, the maximum entropy method by the Maxent algorithm, to analyze the low and high yields patterns and determine the key factors responsible for the yield gaps in three different plots. This work was carried out at Quinta da Cholda, which has an extensive database of various agronomic indicators with spatiotemporal dimensionality, such as productivity maps for 2015-2020, soil electrical conductivity, digital elevation maps and fertility maps. From the productivity maps, low and high productivity locations were identified and georeferenced and the key factors were determined using the Maxent algorithm. In the Cerca plot, productivity is controlled by the apparent electrical conductivity for the regions of high productivity (17-25 mS/m) and low productivity (7-12 mS/m). In the Lourenço plot, total phosphorus (95-135 mg/kg) and pH (7.75 -7.85) are the main factors that characterize the high yields. Low yields are characterized by high elevation regions and pH (>7.9). In the Vinha plot, regions with low productivity are located in high elevation regions (>70m) associated with low water availability. The high yields are located in regions with a low topographical position index (-0.7 - 0), magnesium (170-220 mg/kg) and organic matter (>2%). A factor common to the three plots was the influence of topography in both high and low productivity regions. The methodology developed here allowed the identification of the main factors responsible for the yield gaps, although certain patterns its agronomic interpretation may not be relevant from a quantitative point of view. This work provides agriculture with an innovative modeling approach to efficiently manage high-dimensional spatiotemporal data to support advanced AP solutions.

Keywords: Maize; Precision farming; Yield maps; Ecological niche models; Maximum Entropy; Machine Learning; Maxent;

Resumo

A Agricultura de Precisão (AP) é vista como uma das estratégias da União Europeia para atingir o objetivo de produzir mais com menos, em um mundo com menor quantidade de recursos disponíveis. A aplicação de técnicas de AP na recolha de informação do milho permite detetar a existência de variações espaço-temporais nos campos mas as abordagens tradicionais não conseguem lidar com a multidimensionalidade da informação para determinar os principais efeitos destas variações na produtividade. Uma abordagem de *yield gap* é uma estratégia promissora para avaliar o potencial biofísico disponível de um local especifico através da comparação de uma produtividade potencial com a produtividade média existente. Seguindo esta abordagem, este trabalho colocou a hipótese de que as características edáficas que provocam padrões de elevada e baixa produtividade consistentes no espaço e tempo podem ser interpretados como um nicho ecológico. Á luz desta interpretação este trabalho procurou aplicar um modelo de nicho ecológico, o método da entropia máxima pelo algoritmo Maxent, para analisar os padrões de baixa e alta produtividade e determinar os fatores chave responsáveis pelos *yield gaps* em três parcelas diferentes. Este trabalho foi desenvolvido na Quinta da Cholda, que possui uma extensa base de dados de vários indicadores agronômicos com dimensionalidade espácio-temporal, tais como mapas de produtividade de 2015-2020, condutividade elétrica do solo, mapas de elevação digital e mapas de fertilidade. Apartir dos mapas de produtividade, foi identificado e georreferenciado os locais de baixa e alta produtividade e os fatores chave foram determinados através algoritmo Maxent. Na parcela Cerca, a produtividade é controlada pela condutividade elétrica aparente para as regiões de alta produtividade (17-25 mS/m) e de baixa produtividade (7-12 mS/m). Na parcela Lourenço, o fosforo total (95-135 mg/kg) e o pH (7,75 -7,85) são os principais fatores que caracterizam as altas produtividades. As baixas produtividades são caracterizadas por regiões elevadas e pelo pH (>7.9). Na parcela Vinha as regiões de baixa produtividade situam-se em regiões de maior elevação (>70m) associado a uma baixa disponibilidade hídrica. As altas produtividades situam-se em regiões com um baixo índice de posição topográfico (-0.7 - 0), magnésio (170-220 mg/kg) e matéria orgânica (>2%). Um fator comum ás três parcelas foi a influência da topografia tanto nas regiões de alta e baixa produtividade. A metodologia desenvolvida permitiu a identificação dos principais fatores responsáveis pelos *yield gaps*, embora certos padrões a sua interpretação agronómica poderá não ser relevante do ponto de vista quantitativo. Este trabalho disponibiliza à agricultura uma abordagem de modelação inovadora para gerir com eficiência dados espaço-temporais de alta dimensão para suporte a soluções de AP avançados.

Palavras-chave: Milho; Agricultura de Precisão; Mapas de Produtividade; Modelos de Nicho Ecológico, Entropia Máxima, Machine Learning, Maxent.

Index

List of Tables

List of figures

1. Introduction

Since the food price crisis of 2008-2009, it has become evident that global agricultural commodity markets remain highly volatile. There is a structural situation of low stocks and stagnating productivity, yet demand steadily increases due to changing diets in emerging countries. Extreme events of climate change are also influencing agricultural production. Medium-term projections by the Food and Agriculture Organization of the United Nations estimate that agricultural production would need almost double to meet the need for an anticipated global population of 9 billion people in the year 2050, and is the first of the Millennium Development Goals of eradicating hunger (Sachs et al., 2019). Agriculture is thus recognized as a public good of top priority, although being among the major drivers of negative environmental externalities. It accounts for more than 10% of the total greenhouse gas emissions in the EU- 28, is among the major contributors to water, soild and biodiversity loss (Recanati et al., 2019). Still, it faces the dramatic challenge of producing more with less, with more sustainable use of natural resources, considering both water and land scarcity and the need to mitigate, as well as to adapt to climate change (Pachauri et al., 2014).

Precision agriculture (PA) is recurrently pointed out as one of the strategies of the European Union to achieve this objective (European Commission, 2019). Specifically, the European Green-Deal, which is an integral part of the European Commission strategy to implement the 2030 Agenda and achieve the sustainable development goals of the United Nations, refers to the PA as one of the mechanisms to operationalize its strategy "*from farm to fork*"[1]. The Common Agricultural Policy, since its foundation, as also has been steadily evolving in time to respond to EU society changes and needs. Since 2013, it provides direct support to producers to respond to long term objectives reflecting the three dimensions of sustainability: Viable food production, balanced rural development and sustainable natural resources management.

The yield gap approach is a promising way to achieve necessarily sustainable intensification. This concept indicates the biophysical potential available to improve agricultural production in a specific location (Van Ittersum et al., 2013) and can be estimated as the difference between a benchmark and the actual yield (Beza et al., 2017). Several studies have examined yield gaps at the scale of the region or agro-climatic zones (Mueller et al., 2012; Neumann et al., 2010). However, there is a lack of

[1] Cunha, M. 2020. Agriculture challenges: context and directions – Agronomic project. Course resources for Agronomic Project (2 year, 2. MSc in Agronomy, Faculdade de Ciências, Universidade do Porto. https://www.fc.up.pt/pessoas/mccunha/Projeto_agronomico/bases/ICC.htm

field/farm studies to understand the yield gaps further. At this level, the availability of spatio-temporal agronomic data and reliable modelling approaches are the major drawbacks in a yield gap analysis (Beza et al., 2017).

Since yield is a complex relationship between genetics, environment and management (GxMxE), additional complexity arises when spatial and temporal within-field variations are added up. But if a permanent characteristic exists that impacts the yield, those spatial variations will be similar each year (Blackmore et al., 2003; Ping & Dobermann, 2005). Still, such an approach is not efficient to support better agricultural decisions and does not provide the scientific understanding of the biophysical process of the yield gap so there is a necessity to understand the main drivers of the within-field variability and quantify them in order to correct the existing yield gaps.

Maxent is based on a machine learning approach designed to make predictions from incomplete information (Phillips et al., 2006). The Maxent algorithm is mainly used in species distribution modelling (SDM) (Sillero & Barbosa, 2021). This approach allows to identify and quantify the main factors that influence the distribution and the habitat selection of living organisms through the use of two components: (i) a georeferenced dataset of where the species was detected and (ii) the environmental layers that characterize the study area (Merow et al., 2013).

The basis of ENMs is the ecological niche concept (Hutchinson, 1957) which states that the habitat containing suitable environmental conditions enables a species to survive and reproduce (Grinnell, 1917). It is possible to express this relation through the BAM diagram (biotic, abiotic and movement relationships) (Soberón & Peterson, 2005) and outside of the region defined by these factors, the habitat is unsuitable for a species to exist. Extending this logic to crop yield, high and low productivity areas affected by permanent characteristics that induce a spatial pattern every year can be interpreted as a BAM interaction, representing a region where the species (yield class) possess a more suitable environmental area to exist.

So, in light of this interpretation, this thesis proposes using a Maxent approach combined with yield maps to identify and quantify the main yield gap driving factors using data from three different fields. To our knowledge, this sort of approach has never been done before at the farm level. Three types of multiyear yield maps were built to identify the existing high and low yield patterns to reach this objective. These areas were then analysed using Maxent with an agronomic dataset consisting of electrical conductivity, primary and secondary topographic attributes, and fertility maps.

Literature Review

1.1. Precision Agriculture

Precision Agriculture (PA) is a management strategy that looks to understand, interpret, and manage special and temporal variability in agricultural fields to increase environmental and economic performance (Braga & Pinto, 2011). The parcels are treated heterogeneously, divided by management zones where the soil and the crop requirements are matched to their needs. Each operation is chosen at the correct place, on the exact intensity, in the right timing (Bongiovanni & Lowenberg-DeBoer, 2004; Gebbers & Adamchuk, 2010).

This definition encompasses the idea that a PA is a management strategy that evolves through time. The focus is given to the decision making process regarding resource management with the inference that well-supported agronomic decisions will have positive repercussions (economic, social, environmental) that may be quantifiable or not (Whelan & Taylor, 2013). From a production perspective, PA can be considered as the application of information at the site-specific level, with the objective of: a) optimizing production efficiency b) optimizing quality c) minimizing environmental impacts and d) minimizing risks (agronomic, economic and environmental).

Variability can be decomposed into two main elements: spatial and temporal variability. Spatial variability refers to the soil, crop and environmental characteristics that change in distance and depth, while temporal variability measures the same variations in relation to time (Shannon et al., 2020). The existence of spatial-temporal variabilities in the fields has always been a constant in agriculture. Before the introduction of mechanization, small fields allowed the early agricultures to manage these variabilities manually to assure that the crops were well adjusted to the existing conditions on the farm (Oliver, 2010). Meaning, the factual base of PA – the existence of soil, crop and environment variabilities – has always been taken into account since the beginnings of agriculture.

However, with the enlargement of fields, intensive production and mechanization in the latter half of the last century, managing within-field spatial variability was not possible unless a significant improvement in the technology (Stafford, 2000). Hence, the typical approach adopted by farmers was the whole field approach. When handling large areas, the farmers would simply ignore the existence of variabilities, and the inputs would be applied uniformly throughout the entire field. This type of management strategy is appealing because it enables processing large amounts of areas in a short time frame.

However, the inefficient application of crop production inputs leads to overapplying and underapplying in other sites(Shannon et al., 2020). Sites that cannot lock the inputs lead to environmental impacts due to leaching, runoff, and greenhouse gas emissions (Balafoutis et al., 2017; Guignard et al., 2017).

In 1993, the USA department of Defence finished the Global Positioning System (GPS), which comprises 24 satellites. This system enabled the development of PA as it is today (Oliver, 2010; Stafford, 2000; Zhang et al., 2002; Zhang, 2016). The GPS makes it possible to determine a position (latitude, longitude, altitude) anywhere on the planet, with an accuracy of a few centimetres using real-time kinematics GPS (González-García et al., 2020) . This information is essential for PA. It enables mapping and georeferencing the existing variabilities in topography, fertility or productivity, allowing the application of crop production inputs using variable rate technology (VRT) in the exact location needed (Mulla & Khosla, 2016; Stafford, 2000).

Nowadays, technological development reached a stage that enables the farmer to analyse, measure and make decisions to handle the existing variabilities (Zarco-Tejada et al., 2014). It is considered that the agricultural sector is going through a fourth technological revolution (Agriculture 4.0) supported mainly by the advances in information technologies (Zhai et al., 2020). Advances in several areas, such as remote detection, GPS, big data analysis and artificial intelligence, promises to optimize agricultural operations and inputs to improve yield and reduce losses (Porter & Heppelmann, 2014; Wolfert et al., 2017). However, the explosive information available (crop, economic and environmental data) makes it hard to transform into practical knowledge (Taechatanasat & Armstrong, 2014). Decision support systems (DSS) are necessary to help in the decision-making process based on evidence (Zhai et al., 2020).

1.2. Precision Agriculture Cycle

PA is a cyclic system of data acquisition used to manage the information extraction and decision-making process, with the cycle continuing in the following years (Braga & Pinto, 2011; Gebbers & Adamchuk, 2010). The system is organized in four stages (figure 1). The information acquired in the process is stored in a database to support future decisions (Cambouris et al., 2014; Pedersen & Lind, 2017). The first stage identifies where, how and how much variability is present in a given field. To this effect, several data acquisition technologies can be used (remote sensing, yield mapping, fertility maps, topography) to map these variabilities and georeferenced them (Arslan & Colvin, 2002; Bishop & McBratney, 2002; Cahn et al., 1994; Sishodia et al., 2020).

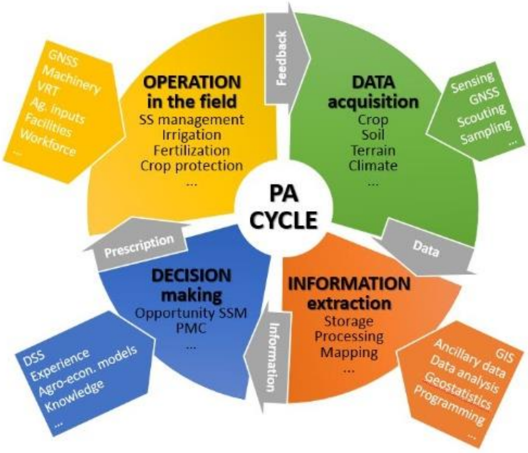


Figure 1: PA cycle. Adapted from Arnó Satorra and Martínez Casasnovas (2016)

The main task of the second stage is to extract and process the acquired data. The use of geographical information systems (GIS) is fundamental in this step. During the third stage, it is necessary to make decisions regarding complex data. Several methodologies exist in supporting the decision-making process, the most noteworthy approaches being the construction of management zones, the use of decision support systems (DSS) and crop models based on machine learning because of the capability to handle highly complex and non-linear agricultural problems (Nawar et al., 2017; Wolfert et al., 2017; Zhai et al., 2020). Finally, the decision is operationalized in the field through VRT technologies. Two main approaches exist: I) Reactive II) Predictive. In the reactive approach (real-time), the crop/soil condition estimation is made and the VRT application rate changes in response to local conditions assessed by a sensor at the time of the application. In contrast, the predictive approach (map-based) sets the condition of the soil/crop off-site, using several different sensors to generate soil property maps (yield, topography, fertility). Through the combined use of this information, VRT applications recommendations are made (Adamchuk et al., 2011; Braga & Pinto, 2011).

1.3. Within Field Variability

The with-in field variability in agricultural soils can be decomposed into two main types of components - Temporal and Spatial (Srinivasan, 2006). Spatial variability measures the changes in the physicochemical properties of the soil and productive capability through space. These different variations can be inherent to the pedological factors of the field, but some variability can be introduced through management decisions (Iqbal et al., 2005). Spatial variability measures the same changes but through time. The main factors that influence the variability are listed in table 1.

Table 1: Factors influencing yield variations and survey methods. Adapted from Godwin and Miller (2003)

Group	Factor	Method
Soil-water	Soil texture, structure, available water and waterlogging	Soil mapping, profile description, electro-magnetic induction
Topography	Topography and micro-climate	Topography surveys, 3D-DGPS
Soil nutrition	Major nutrients, pH and trace elements	Targeted sampling, canopy density, yield maps, ADP
Crop weeds and pests	Weeds, pests and diseases	Field walking, ADP, reflectance imaging

3D-DGPS: three-dimensional differential global positioning system; ADP: Aerial digital photography.

1.3.1. Soil-Water

The impact of water availability on yield is well documented (Condon et al., 2002), with soil texture being an important environmental factor that influences crop productivity because of its direct effect on soil water and complex interactions with other environmental factors (He et al., 2014). The apparent electrical conductivity (EC_a) is used in PA to identify several physicochemical properties (Corwin & Lesch, 2005). In regions where salinity is not a significant factor, measurements of the EC_a is primarily a function of soil moisture, organic matter and texture, as illustrated in table 2 (Corwin & Lesch, 2005; Kuang et al., 2012).

Measurement of EC_a is done using electrical resistivity, electromagnetic induction and time domain frequency which is mainly a function of the number of ions present. Electrical resistivity and electromagnetic induction are well suited for field-scale applications because their volumes of measurement are large which reduces the influence of local scale variability (Corwin & Lesch, 2003; Grisso et al., 2005). Electrical resistivity introduces an electrical current into the soil through the electrodes and measures the difference in the current flow potential. But electrical resistivity is an invasive method that requires good contact between the four electrodes and the soil. Electromagnetic induction however, just requires a transmitter coil above the surface. The coil induces a circular eddy-current loop in the soil, with the magnitude of these loops directly proportional to the electrical conductivity in the vicinity of that loop. However, the measurement of EC_a with electromagnetic induction depends of a depth weighted response function, unlike electrical resistivity, where the measurement is linear over depth. Time domain reflectometry measures the time that a voltage pulse travels down through a soil probe and back, which is a function of the dielectric constant of the porous media being measured. This method is non-invasive and has a similar performance to the accepted methods of EC_a measurement (Corwin & Lesch, 2005). However, time domain reflectometry is a stationary instrument and its use has been limited due to its high cost and need for complex wave form analysis (Hardie, 2020).

Table 2: Soil texture and their electrical conductivity. Adapted from:(Heege, 2013)

Soil texture class or influence of salt	Electrical conductivity (mS/m)
Sand	*0.1 – 1*
Loamy sand	*1.0 – 5.0*
Loam	*5.0 – 12.5*
Silt	*12.5 – 25.0*
Clay	*25.0 – 100*
Saline Soil	*> 100*

The properties measured by EC_a can be divided into two categories: a) Static b) Dynamic. Dynamic properties change through time, like the soil water content. Water affects the EC_a measurements because many ions are in the soil solution, and its temporal variation provokes different impacts on the EC_a (Brevik et al., 2006). Static properties, on the other hand, remain constant through time. Among the static properties, the texture is considered the property with the greatest influence (table 2) and among the textural classes, clay exerts the greatest influence in the EC_a because it possesses the highest ion exchange capacity (Heege, 2013).

1.3.2. Topography

The elevation is a critical layer in PA because it provides critical information on topography's impact on yield variability. To develop a high-resolution digital elevation map (DEM), the data from a differential GPS or Real time Kinematic GPS is required. This allows resolutions from up to 1m on the ground, which is enough to reflect the continuous nature of the topography (Bishop & McBratney, 2002). From the DEM, it is possible to generate new topographical parameters to characterize it. Wilson and Gallant (2000a) divided the topographical parameters into a) primary and b) secondary. Primary parameters are calculated directly from a DEM and include slope, aspect, flow accumulation, curvature (planar, profile, tangential) and shaded relief. Secondary parameters result from the combination of two or more primary parameters and are used as indexes that characterize the spatial variability of specific processes in the fields (Moore et al., 1991). Several of these indexes describe and quantify the influence of the topography in water distribution, such as the topographical wetness index (TWI) and distance to flow lines (DFL).

The relationship between several primary and secondary topographical characteristics with yield has been investigated, and several relations with yield has been found. Kravchenko et al. (2000) studied the effect of slope in corn and soybean yield. In growing seasons with dry weather, larger yields were observed at low slope locations, and moderate and high slopes had a wide range of values. During wet seasons, lower yields prevailed at locations with low slopes. Kaspar et al. (2003) analysed the effect of elevation, slope and curvature on yields. In years with less than normal precipitation, corn yield was negatively correlated with elevation, slope and curvature. In wetter years, yield was positively correlated with relative elevation and slope. Bakhsh, Colvin, et al. (2000) showed that lower yield areas of corn were consistent in regions of higher elevation, and higher yielding areas were variable. Hansen et al. (2013) demonstrated that maize in the summit or shoulder positions had less water, leaf area, biomass, N and P uptake when compared with maize grown in backslope areas. Mishra et al. (2008) also showed that summit/shoulder areas were associated with low yield values, soil organic carbon and pH values. And the opposite relation was found in footslope regions, with high yield values, soil organic carbon and pH. Kravchenko and Bullock (2002) found that in seasons with enough precipitation, higher elevation and steep slopes and areas with convex curvature would produce soybeans with high protein and oil content. But in dry seasons, the opposite relation was found. Kravchenko et al. (2005) showed that the yield variation was high as 45% in years of low rainfall and low as 14% with above average precipitation. Maestrini and Basso (2018) confirmed this relation by showing that the performance of areas of a field with a high TWI (depressions) depends on the rain patterns of the season. In wet years these regions may be waterlogged, yielding less that the rest of the field. In dry years, these regions are wetter than the rest of the field, yielding more than the rest of the field. Da Silva and Silva (2008a, 2008b) evaluated the relationship between several topographic attributes in irrigated maize and found that the highest correlating attribute was the distance do flow accumulation lines (DFL). That is, the yield increases when the distance is reduced to water flow accumulation areas. Kumhálová et al. (2011) analysed the correlation between flow accumulation and yield and found the correlation weak for wet years and strong for drier years.

The impact of terrain topography in water displacement is due to the gravitational potential energy gradients from the elevation differences (Murphy et al., 2009), which controls the vertical and horizontal water distribution (Verity & Anderson, 1990). Topography's influence on water distribution is one of the main factors for yield variabilities (Mishra et al., 2008; Sadler et al., 2000). Godwin and Miller (2003) supported these findings. They also stated that topography was one of the most obvious causes of yield variations in field crops for its direct effect on micro-climate related factors.

1.3.3. Soil Nutrition

To address soil fertility's spatial variability, extensive soil sampling is recognized as a basis for site-specific fertilizer applications (Cambardella & Karlen, 1999). Spatial patterns of soil properties and nutrient concentrations need to be characterized to establish site-specific farming practices (Cahn et al., 1994). Their quality affects the efficacy of soil fertility management (Mueller et al., 2001). There are two primary approaches to develop VRT fertilization maps: a) grid soil sampling and b) management zone delineation. These two approaches depend on the farmers' managing options, fields history, and the available spatial information resources (Ferguson et al., 2017).

Grid soil sampling is more commonly used if (i) there is no prior knowledge on the field; (ii) previous decisions significantly altered the fertility; (iii) several small fields were merged into a single field; (iv) if an accurate base map of soil organic matter is desired (Ferguson & Hergert, 2009). A grid sampling strategy is enough to reveal fertility patterns but only if a high sampling density is used (Flowers et al., 2005; Franzen & Peck, 1995; Nanni et al., 2011). It consists in sampling pre-determined spots following a regular or irregular grid pattern (figure 2a). It is retrieved between 5-8 samples surrounding the pre-determined spots in a radius of 2-3m, which are aggregated in a single composite sample. The regular grid pattern was a common sampling method before the GPS but is susceptible to systematic errors (Franzen et al., 2018). The irregular grid pattern minimizes the effects of these errors in two directions, and it is the sampling method most adapted to kriging interpolation (Gotway et al., 1996).

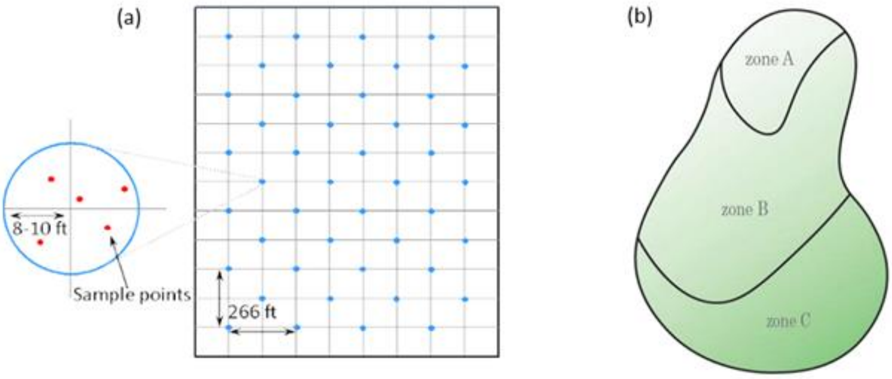


Figure 2: The two main approaches for developing soil specific maps: (a) Grid sampling (b) Zone sampling. Adapted from Ferguson et al. (2017)

A management zone approach may be preferable if several sources of spatial information exist. A management zone is a sub-region in a field that exhibits a homogenous combination of limiting factors for which a single rate of a crop input is necessary for optimal efficiency (figure 2b) (Doerge, 1999; Fraisse et al., 2001; Haghverdi et al., 2015). These regions are generated from the combination of several layers such as yield maps, topography, aerials photos, electrical conductivity, remote sensing, or the farmer's experience. The resulting management zone needs to be simple, precise, stable over the years and identify regions that need to be managed differently (Khosla et al., 2010). Soil sampling on a management zone approach can be randomly collected within each zone and integrated into a single sample for laboratory analysis (Chang et al., 2003). This process significantly reduces the cost of the analysis, in opposition to grid sampling.

There is still a third approach, by using real-time sensors. Traditional sampling methods are slow and costly (Rossel & Bouma, 2016), so it is necessary to develop new strategies for mapping different soil properties with a high sampling density at a low cost (Adamchuk et al., 2004; Viscarra Rossel et al., 2010). For this effect, the use of proximal soil sensors can understand and quantify the existing spatial-temporal variabilities in a field (Kuang et al., 2012). By being connected to a GPS to register the sampled position, it is possible to quickly produce maps of several properties at a high spatial resolution (Adamchuk et al., 2011). The type of sensors can be electrical, electromagnetically, optical, radiometric, mechanic, acoustic, pneumatic and electrochemical (Viscarra Rossel et al., 2010).

Ideally, a sensor corresponds to a single soil traits variability and is highly correlated to a conventional analyses method. But in reality the sensors respond to several properties, and the separation of their effects is difficult (Adamchuk et al., 2004; Viscarra Rossel et al., 2010), which allows only indirect information about their properties.

1.4. Grain yield monitors

To acquire yield data, yield monitors need to be installed near the grain elevator. The most used commercial systems are classified in two categories: i) mass flow meters and ii) volume flow meters(Arslan & Colvin, 2002). The mass flow meters calculate the productivity directly from the force that hits the impact plate, which is connected to a potentiometer that generates an electrical signal that is proportional to the mass of the grain (figure 3), while the volume flow potentiometers needs to incorporate the density of the grain to determine the rate of flow mass (Arslan & Colvin, 2002; Whelan & Taylor, 2013). For cereal crops, mass flow meters are the most suitable because the beans are partially dry at the time of the harvest and can deal with mechanical handling without damage (Fulton et al., 2018).

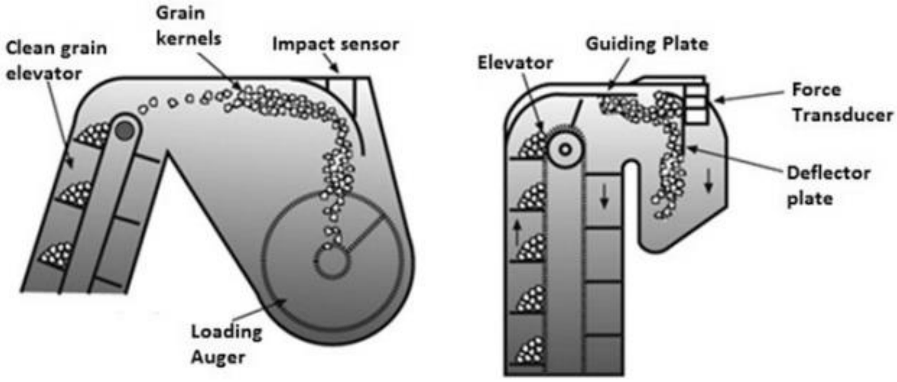


Figure 3: Mass flow meter. Adapted from Fulton et al. (2018)

To produce a yield map, the number of grains per a specific area (kg/ha) is recorded. The yield monitor incorporates several different types of information relayed by the sensor system (figure 4). All this information is necessary to determine two important parameters for yield estimation: weight of the grain harvested and harvested area. The weight of the grain is expressed by the grain flow (kg/s) and the harvested area is determined by the speed and operating width of the harvester combine. With this information, the yield calculation is done through equation 1 (Whelan & Taylor, 2013):

$$Y_{t/ha} = ((\frac{f_{kg/s} \times i_s}{1000_{\,kg/s}}) \times (\frac{10\,000_{m2/ha}}{d_m \times w_m}) \times (\frac{100 - m_{\%}}{100 - sm_{\%}})) \qquad \text{(Eq.1)}$$

Where:

$Y_{t/ha}$ = yield in t/ha
$f_{kg/s}$ =grain flow in kg/s
i = sampling interval (s)
d_m = distance travelled
w_m = width of the harvester (m)
$m_{\%}$ = grain moisture (%)
$sm_{\%}$ = reference moisture content for the market (%)

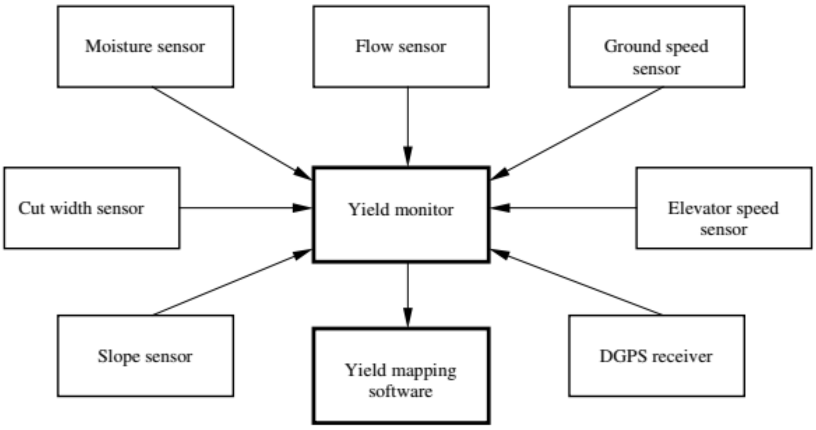


Figure 4: Sensors system associated with the yield monitor. Adapted from: Arslan and Colvin (2002)

1.4.1. Yield Maps

Yield maps have been recognized as a valuable source of information for their effectiveness in mapping the within-field yield variability (Diker et al., 2004; Pringle et al., 2003). Storing yield georeferenced data allows to identify the specific productive potential of some sites of the field, gives feedback on how the crop reacted in the function of decisions made in previous years and makes it possible to determine the number of nutrients that were exported by the crop (Arslan & Colvin, 2002; Reyns et al., 2002). However, interpreting multiple years of data may be challenging because yield variability is caused by many factors (Wibawa et al., 1993). For example, Machado et al. (2002) found that crop stress, pests and diseases could explain up to 50% of yield variability across years and sites. As a result, yield maps vary from year to year, making it difficult to use them as the basis for site-specific management. Despite strong temporal variability of crop yield, it is often possible to detect consistent yield patterns across years (Dobermann et al., 2003; Kitchen et al., 2005; Taylor et al., 2007). The assumption is that if permanent soil characteristics affected the crop yields, in the same way, each year, the spatial yield pattern that affects that particular area would be similar each year (Blackmore et al., 2003; Ping & Dobermann, 2005). Some yield patterns were found consistent even when under different crops and varying climate conditions and can deliver relevant information regarding the soil characteristics within the field or depict the influence of other external factors, such as management practices and weather conditions (Diker et al., 2004). Taylor et al. (2007) showed that in specific portions of their field study, crop rotation management originated variations in yield spatial patterns in previous years. The influence of dry and wet years was also studied, and it has been found that high-yielding areas in dry years could be at the same time low yielding areas in wet years (Maestrini & Basso, 2018).

Although yield data is a valuable source of information, a few issues remain. The spatial yields patterns that originate from the interaction between management, climate and environment conditions within cropping seasons should not be used to generate VRT applications maps directly for the year *n* by solely relying on yield data in year *n*-1 (Ping & Dobermann, 2005). Although a single year yield map is useful, it should only be used for posterior interpretation of possible yield variations (Ping & Dobermann, 2005). It is acknowledged that the yield temporal variability is often stronger than yield spatial variability, which can hinder any analysis (Blackmore et al., 2003; Bramley & Hamilton, 2004). Temporal variability is essentially due to non-stable factors, such as climate patterns or the type of crop grown each year (Basso et al., 2012). Also, the number of years of yield data available to conduct yield temporal analysis is critical (Bakhsh, Jaynes, et al., 2000; Kitchen et al., 2005). Ping and Dobermann (2005) suggested that a minimum of 5 years of yield data for irrigated crops is necessary for yield classification. On top of that, yield data often comes with a large number of erroneous observations, such as flow delay, filling and emptying times, the abrupt speed changes of the combine, inaccurate sensor measurements of yield and moisture, the accuracy of the positioning system and errors dealing with the harvester operator(Arslan & Colvin, 2002; Blackmore, 1999; Lyle et al., 2014; Simbahan et al., 2004; Sudduth & Drummond, 2007).

1.5. Crop Yield Gap

The crop yield gap is a concept that describes the difference between the average on-farm yield and the yield under optimum management (figure 5) (Van Ittersum et al., 2013).This information provides the foundation for identifying the most important crop, soil and management factors limiting current farm yields (Lobell et al., 2009; Van Ittersum et al., 2013). This concept is based on ecological principles, and it is highly relevant because it indicates the biophysical potential available to improve agricultural output according to its specific location (Van Ittersum et al., 2013).

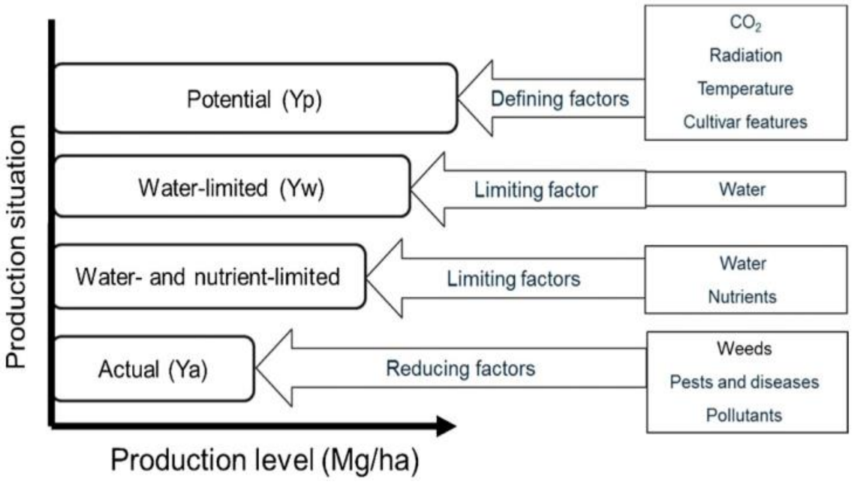


Figure 5: Different production levels as determined by growth defining, limiting and reducing factors. Adapted from (Van Ittersum et al., 2013)

Yield under optimum management is labelled as potential yield (Y_p), and is defined as “*the yield of a cultivar when grown in environments to which it is adapted, with nutrients and water non-limiting and with pests, diseases, weeds, lodging, and other stresses effectively controlled*” ” (Evans & Fischer, 1999). Under irrigated conditions, Y_p is determined by solar radiation, temperature, atmospheric CO2 and genetic traits that govern the growing period and the canopy architecture. In rainfed conditions, Y_p is most affected by water availability and is referred to as water-limited yield (Y_w) (Fischer, 2015; Lobell et al., 2009). Y_p can be estimated using field experiments and crop growth simulations models, evaluating the potential experimental yield (Y_e) and potential climatic yield (Y_p/Y_w). Y_e are defined as the maximum yield possible to obtain with the best management practices (Liang et al., 2011; Lobell et al., 2009). Table 3 summarizes the critical benchmarks used in yield gap analysis.

Table 3: Common methods to estimate yield benchmarks

Yield benchmark	Estimation method	Limitations
Potential yield		
Climatic potential yield (Y_p)	Crop growth simulation models for irrigated systems	Requires site-specific data that is not widely available at a local scale and rigorous validation at field conditions
Water-limited potential yield (Y_w)	Crop growth simulation models for rainfed systems	As above
Experiment potential yield (Y_e)	Field experiments with no yield constraints	Difficult to eliminate yield limiting and yield reducing factors. May not represent agro-ecological conditions of yields
Economically attainable yield		
Exploitable yield (Y_{ex})	70-80% of potential yield	Depends on price ratios and environmental conditions
Attainable farm yield (Y_{at})	Mean of the upper 10th percentile	Related to the best technologies that are available and affordable

Adapted from: (Cassman et al., 2003); (Lobell et al., 2009); (Van Ittersum et al., 2013); (Fischer, 2015)

Average field yield (Y_a) is the crop yield actually achieved in the farmer's field. It is defined as the average yield reached using the most widely used management practices, such as the sowing date, cultivar maturity, plant density, the type of nutrient management and crop protection. The yield gap (Y_g) is the difference between Y_p, Y_w or Y_e and Y_a's actual yields.

It is necessary to take into consideration the environmental, economic and social consequences of reaching target yields because in most cases is not desirable or practical to completely close the yield gaps (Cunningham et al., 2013; Mueller et al., 2012). It is also impossible to achieve perfect crop management needed to reach Y_p on a large scale (Koning & van Ittersum, 2009; Lobell et al., 2009; Van Ittersum et al., 2013). The response to applied inputs follows a diminishing return curve when yields approach celling yields making it increasingly difficult, time consuming and costly to eliminate the small imperfections to make small gains in yield (Cassman et al., 2003).

Due to these constraints, alternative definitions that are functionally relevant for real farming conditions can be used such as economic yields or best farmers yield (Y_{at}) (table 3). Comparing these benchmarks with the Y_a may be an alternative to assess yield gaps. Economically attainable yield is defined as the possible exploited yield attained when economically optimal practices are in place. Because of diminishing returns on investment, the exploitable yields (Y_{ex}) usually lie around 20-30% below the Y_p if favourable weather conditions exist (Fischer, 2015; Lobell et al., 2009; Van Ittersum et al., 2013).

Alternatively, economic yield may be determined by identifying the best yields in the farmers' fields through skilful use of the best available technology (Tittonell & Giller, 2013), having been measured as the mean of the upper 10th percentile in rice production systems (Laborte et al., 2012; Van Ittersum et al., 2013). This benchmark is defined as attainable farm yield (Y_{at}). Using the definitions above Y_p is only constrained by genotype and environment, where Y_{at} is limited by both these factors as well by optimal agronomic practices, socio-economic and institutional issues that have an impact on decision making and access to inputs and technology. The difference between the Y_{ex} and Y_{at} is considered the "exploitable yield gap" whereas the difference between Y_p and Y_{ex} is the "unexploitable yield gap".

Several studies have examined yield gaps at the scale of the region, agro-climatic zone and global level (Hochman et al., 2016; Licker et al., 2010; Mueller et al., 2012; Neumann et al., 2010; Schils et al., 2018) but in order to compare different regions in relative terms, it is necessary to use harmonized data (Van Ittersum et al., 2013). The Global Yield Gap Atlas (www.yieldgap.org) is an international project with the goal of establishing improved methods for estimating the yield gap of existing cropland worldwide (Grassini et al., 2017). It is based on local data from each of the worlds's major crop production countries (Grassini et al., 2015), and the estimates of these analyses are used as inputs to economic models to assess food security and identify areas where research and development interventions should be prioritised (Foley et al., 2011; Mueller et al., 2012; Van Ittersum et al., 2013).

However, more local studies are required to bring the role of the farmer and the biophysical conditions into the picture (Beza et al., 2017). Analysing yield gaps at the farm and farming levels can better understand whether yield gaps can be closed (Giller et al., 2006). However, a major drawback of this type of analysis at the farm level is the high data standards required which typically refers to (a) large sample size (b) fine resolution (c) great level of detail (Beza et al., 2017).

1.6. Ecological Niche Models

Understanding the relationship between a species or community with the spatiotemporal variation in abiotic and biotic conditions is crucial in ecology and conservation (Elith & Leathwick, 2009). The use of species distribution models (SDMs) and niche ecological models (ENMs) has advanced the understanding of several ecological mechanisms responsible for their distribution and is a key tool in predicting species response to environmental change (Elith et al., 2006; Elith & Leathwick, 2009; Zimmermann et al., 2010). But there are distinct differences among these two terms. SDMs refer to the potential distribution of suitable habitats that are being predicted by the model (Peterson & Soberón, 2012), while ENMs refers directly to ecological niche theory, and forecasts the species realized niche, according to the type of species records available (Sillero, 2011).

These methods are most often used in one of four ways: (1) to estimate the relative habitat suitability or occurrence probability known to be occupied by a species, (2) to estimate the relative habitat suitability in geographic areas not known to be occupied by the species, (3) to estimate changes in the suitability in a specific scenario of environmental change, and (4) as estimates of a species niche (Hampe, 2004; Sillero et al., 2021; Soberón & Peterson, 2005). The increased popularity of these techniques is due to two main reasons: the increased availability of presence of species data in large quantities (Edwards, 2004) and the availability, ease of use and resolution of digital environmental layers (Soberón & Peterson, 2004).

The theoretical framework of ENM/SDMs is based on the ecological niche concept, which is understood as the subdivision of the habitat containing the environmental conditions that enable individuals of a species to survive and reproduce (Grinnell, 1917). It is possible to express this relation through the interaction of three factors that limit the distribution of a species (Soberón & Peterson, 2005): A is the environmental conditions (abiotic) under which a species can establish a population, survive and reproduce; B is the biotic environment which is determined by the interaction between species such as competition or predation in which a species can persist; M is the area that is accessible to the species via its movement or dispersal capabilities. The biotic (B), Abiotic (A) and Movement (M) is the BAM diagram (figure 6), which is an abstract representation of the geographic space that represents the region where the species occur (Soberón & Peterson, 2005), and has become a central concern in model design (Barve et al., 2011; Saupe et al., 2012). Outside of the region defined by these factors, the habitat is unsuitable for a species to exist. Region A is commonly known as the ‘fundamental niche’, which is a N-dimensional Space where a species can maintain a viable population and persist through time (Hutchinson, 1957). The common area of A and B represents the ‘realized niche’ which is the space that a species occupies that does not use the entire fundamental niche due to constraints by species interactions (Hutchinson, 1957). M constrains the ‘realized niche’ to the “*region that has the right set of biotic and abiotic factors and that is accessible to the species, as is equivalent to the geographic distribution of the species*” (Soberón & Peterson, 2005). The constrained space is known as the occupied niche (Pearson, 2007).

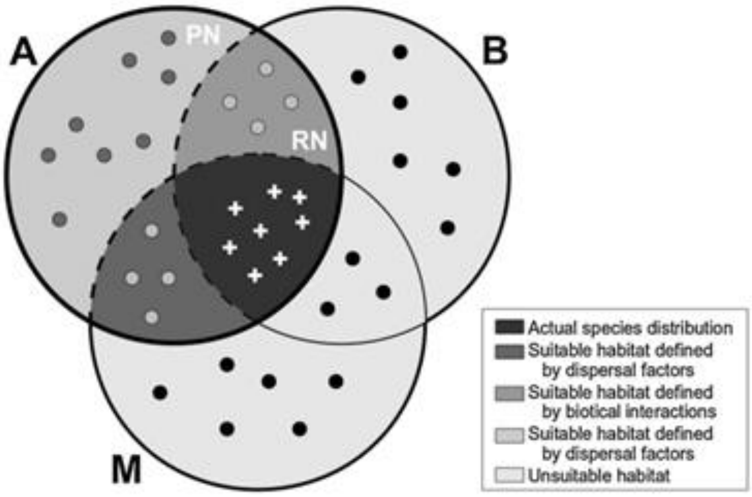

A
PN
B
RN
M
Actual species distribution
Suitable habitat defined by dispersal factors
Suitable habitat defined by biotical interactions
Suitable habitat defined by dispersal factors
Unsuitable habitat

Figure 6: BAM diagram represents the Biotic (B), abiotic (A) and dispersal factors (M). Outside of the space defined by three factors, the habitat becomes unsuitable for a species. Adapted from: Sillero (2011)

Three main categories of models are typically used in ENM/SDMs studies such as: (i) correlative models estimates a species ecological requirements by relating georeferenced locality records to a set of environmental variables (Araujo & Guisan, 2006; Franklin, 2010) (ii) mechanistic models use detailed bio-physiological information about a species and link them to the spatial habitat data, translating the interaction of the organism with its environment into key fitness components (Kearney & Porter, 2009); (iii) and process-oriented models, which is a combination of certain aspects of the other two modelling approaches. A hypothesis about niche, dispersal and biotic interactions can be integrated in these hybrid models, estimating species distribution (Peterson et al., 2015).

Correlative approaches are far more frequent than the mechanistic or process based approaches due to the great availability of detailed data about biodiversity and environment, advances in statistical techniques, development of GIS tools and specialized modelling software (Elith & Leathwick, 2009; Peterson et al., 2015). Among the correlative approaches, the most used methods are machine-learning methods, followed by statistical methods and then similarity based and expert rule approaches (Melo-Merino et al., 2020).

1.6.1.Maxent

Maxent is the Java software that implements the maximum entropy method, a machine learning algorithm capable of modelling complex, non-linear relationships between presence-only data and predictors (Elith et al., 2006; Phillips et al., 2006). The Maxent method is one of the most common and popular algorithms for the spatial distribution of fossils and living organisms (Kaky et al., 2020; Merow et al., 2013; Phillips et al., 2006; Phillips & Dudík, 2008). Its popularity is due to its robustness to low sample sizes and relative high predictive accuracy (Pearson et al., 2007) coupled with flexibility in model construction, and it is easy to implement automatically (Elith et al., 2006; Merow et al., 2013; Townsend Peterson et al., 2007). Maxent allows for many parameters to be manually determined by the user but also offers robust default values for accurate species distribution models (Phillips & Dudík, 2008).

To predict an unknown probability distribution, "*the best approach is to ensure that the approximation satisfies any constraints on the unknown distribution that we are aware of, and that subject to those constraints, the distribution should have maximum entropy"* (Jaynes, 1957). This means that the probability distribution that best represents the current state of knowledge of a system is the distribution that is most spread out while considering the constraints of the environmental variables of known locations. The constraints are expressed as functions of the environmental variables, called features. Since the set of constraints underspecifies the model, among the probability distributions that satisfy the constraints, the most unconstrained one is maximum entropy (Jaynes, 1957).

For this effect, Maxent calculates the probability density for the presence points and for the background (figure 7). The background represents the available environment and is defined by many random sample points that characterise the study area. The probability density for the background characterizes the available knowledge, whereas the probability density of the presence points characterizes the environment where the species has been found. Then, Maxent chooses the distribution that maximizes the similarity between the environmental characteristics of the total environment and those of the locations where the species is known to exist.

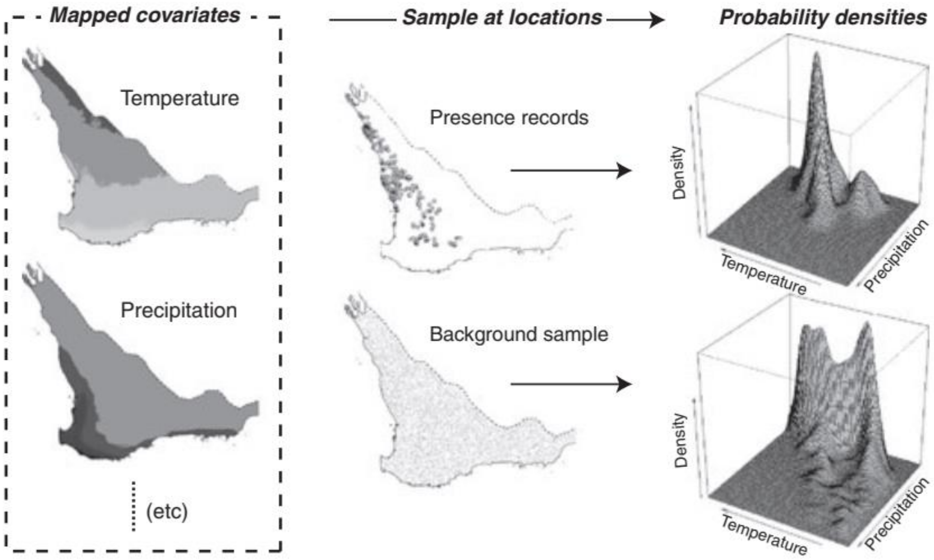


Figure 7: A diagrammatic representation of the probability densities for the presence and background samples. Adapted from: Elith et al. (2011).

This is known as the maximum entropy principle, where the true distribution of a species is represented as a probability distribution π over the set X of sites in the study area. The π assigns a non-negative value to each point x, and the values π(x) sum equals to 1. Following Phillips et al. (2006), the approximation of π is a probability distribution, defined as $\hat{\pi}$ in equation 2 and the entropy is:

$$H(\hat{\pi}) = -\sum_{x \in X} \hat{\pi}\,(x) ln\, \hat{\pi}(x) \tag{Eq.2}$$

According to (Dudik et al., 2004), the Maxent distribution belongs to the Gibbs distribution. Gibbs distributions (equation 3) are exponential distributions parameterized by a vector of feature weights λ=(λ_1, …,λ_n) defined by:

$$q_\lambda(x) = \frac{exp\left(\sum_{j=1}^{n} \lambda_j f_j(x)\right)}{Z_\lambda} \tag{Eq.3}$$

Where Z_λ is a normalization constant ensuring that the probabilities of a distribution $q_\lambda(x)$ sum to 1 over the study area. Which means that the value of a Maxent model q_λ at a site x depends only on the environmental variables at x. Equation 4 is the Maxent q_λ distribution that maximizes a penalized log likelihood of the presence locations, namely:

$$\frac{1}{m}\sum_{i=1}^{m} ln\big(q_\lambda(x_i)\big) - \sum_{i=1}^{n} \beta_j\lambda_j \vee \qquad \text{(Eq.4)}$$

Where the regularization parameter $\beta_j$is the width of the error bound for the features $f_j$and x_1….,x_m are the presence locations. In the first term, the loglikelihood (or gain) gets larger to better fit the data. The second term is the regularization, known as LASSO penalty (Tibshirani, 1996).This penalty gets larger as the weights λ_j get larger, and typically means that the model is more complex and is likely to overfit. Maximizing the difference between the two terms is a Gibbs distribution that fits the data well and is not too complex. The regularization parameters control the trade-off.

The best performance of Maxent is when the regularization parameters β_j are as small as possible (equation 5). This ensures that the true feature means (under π) is within the error bounds and serves as an incentive to keep the error bounds as tight as possible. The regularization parameters are defined as:

$$\beta_j = \beta\sqrt{\frac{s^2[f_j]}{m}} \qquad \text{(Eq.5)}$$

Where β is a regularization parameter that depends on the feature class and $s^2[f_j]$ is the empirical variance of feature f_j so $\sqrt{s^2[f_j]}$ / m is an estimate of the standard deviation of the empirical average.

Regularization reduces overfitting in two ways. It ensures that the imprecision in the empirically measured constraints approximately satisfies the constraints required by the prediction. Secondly, regularization penalizes the model proportionally to the coefficients magnitude, shrinking several coefficients to zero. The regularization coefficient can be tuned to amplify or dampen its effect to produce more or less complex models (Merow et al., 2013; Warren & Seifert, 2011).

Maxent derives several ‘features’ from the predictors, each of which is a simple mathematical transformation of the predictor. Six types of features exist: Linear, quadratic, product, threshold, hinge and category indicator features. Linear features are equal to continuous environmental variables, quadratic features to their squares and product features equal products of pairs of continuous environmental variables. Threshold features make a continuous predictor binary, 0 below the threshold and 1 above. Hinge features are similar to the threshold, except that a linear function is used instead of a step function (Merow et al., 2013; Phillips & Dudík, 2008).

2. Materials and methods

The materials and methods section is structured in six different subsections. First, the objective of this thesis is presented, followed by the characterization of the study area. From the yield data, three type of yield maps were calculated to determine if permanent yield impacting factors exist. The soil dataset is described and from the DEM, several primary and secondary topographic attributes were derived. Using a yield gap approach, the georeferenced locations of the permanent yield impacting factors were extracted for the Maxent analysis. The variable selection process screened all variables for correlation and collinearity, where the highest contributing variables for the models were selected for model building. Finally, several models were built for the high and low yielding regions using the Kuenm R package, and the best model for each yield class was kept for the analysis of the response curves generated by Maxent. Figure 8 summarizes the work that was developed in this thesis.

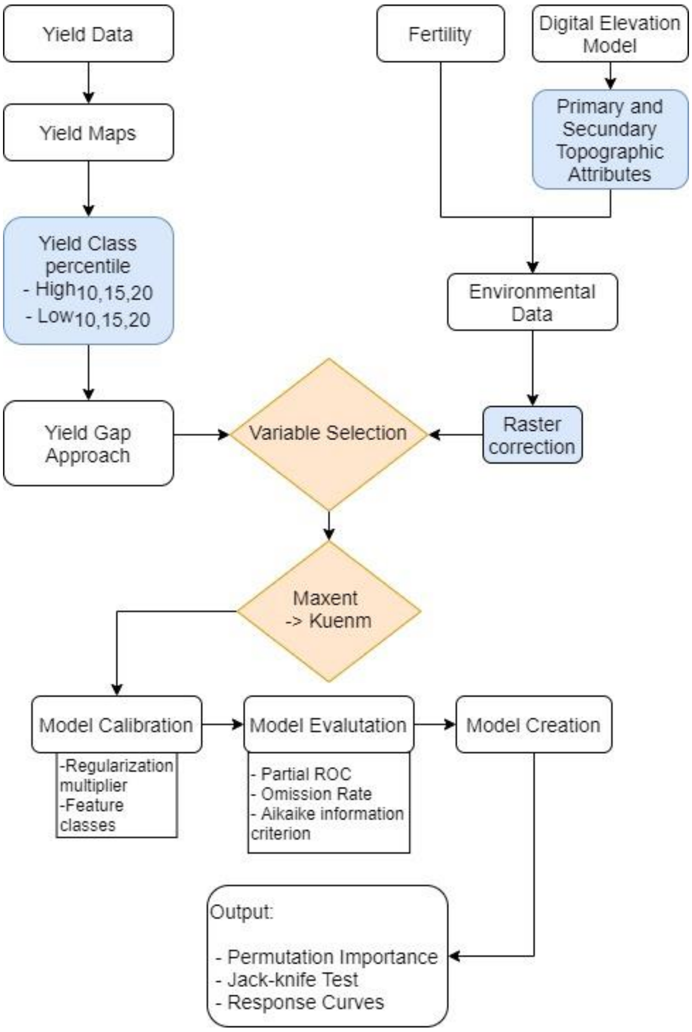


Figure 8: Flowchart summarizing the methodology used in this study. Blue represents data preparation for the Maxent model. Orange indicates Maxent processing.]

2.1. Objectives

The purpose of this work is to determine the main yield impacting factors through a yield gap approach using an established method in the field of ecology, the maximum entropy method (Maxent).

The niche ecological concept states that suitable environment conditions enable a species to survive, reproduce and maintain a species in that particular ecological niche. Since specific edaphic conditions characterize a yield pattern, it can also be interpreted as an ecological niche. So, we hypothesise that high and low yield patterns can be interpreted as a BAM interaction. Using this established framework developed in ecology, we propose to analyse low and high yielding locations generated by a multiyear yield map through a yield gap approach using a niche ecological model, Maxent.

It is the expectation that since Maxent can quantify the main biophysical factors that influence the distribution and the habitat selection of living organisms, the existing yield patterns will be properly characterized by this approach, with the end results being realistic, interpretable and viable for future agronomical recommendations.

2.2. Study Area

Quinta da Cholda is a large maize farm located in the county of Golegã, district of Santarém (39°21'48.0"N; 8°32'25.9"W), which is managed by João Coimbra. The farm produces grain maize on 560 hectares, which all fields are irrigated by a centre pivot, is one of the largest maize producers in Portugal. The farm is specialized in maize and has been sowing it continuously over the last 30 years, obtaining average productivity of 16.7 ton/ha in 2020. It is regarded as one of the most efficient farms in Europe.

The study area (figure 9) is inserted in the aquifer system of Tejo, whose materials are of fluvial origin (IUSS Working Group, 2006): modern alluviums (Holocene) and terraces (Pleistocene) belonging to the most important hydrogeological unit in the country, the Tejo-Sado basin. On the right side of the Tejo, the quaternary terraces have a large extension near Entroncamento, Golegã, Azinhaga and Pombalino (Almeida et al., 2000).

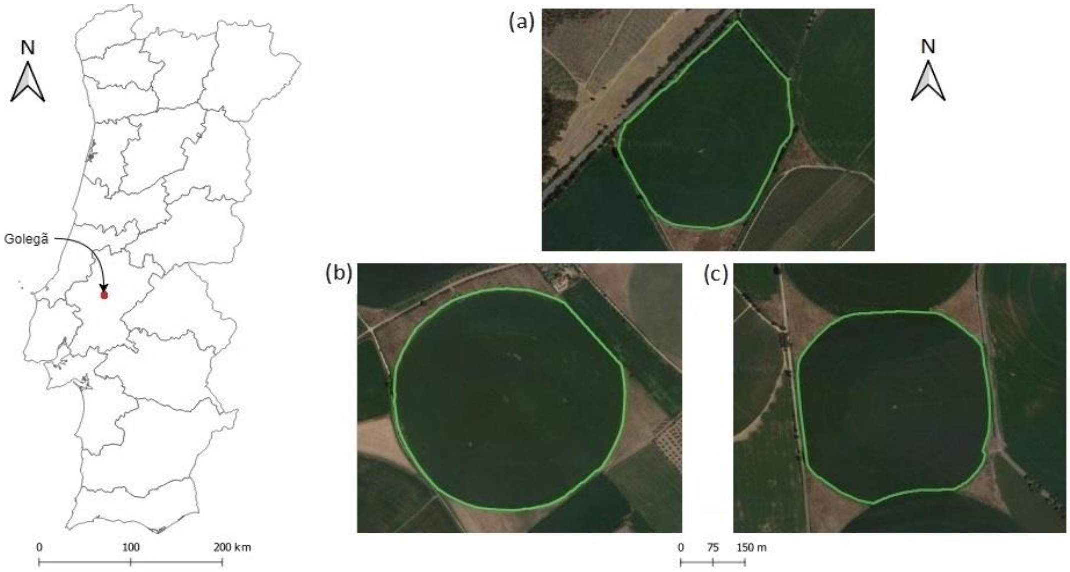


Figure 9: Location of the study area, Golegã in the district of Santarém. The studied fields were (a) Vinha, (b) Cerca, and (c) Lourenço.

According to Cardoso (1965), the Golegã soils belong to order of soils incipient, suborder of alluvisols and group of modern alluviosols. Following Figueira (1997), these soils are modern alluviosols of light or medium calcareous texture. They are non-evolved soils, without differentiated genetic horizons, practically reduced to the original material consisting of mineral and organic detrital materials, transported by river water, from gravel and coarse sand to the finest clay particles.

The farm has a weather station, but it only has data since 2015. As the Santarém weather station provides data from 1970 – 2018, it was decided to use this data to characterize the weather patterns from 1970 to 2000 and the farm station to characterize the weather for 2015-2020 (figure 10).

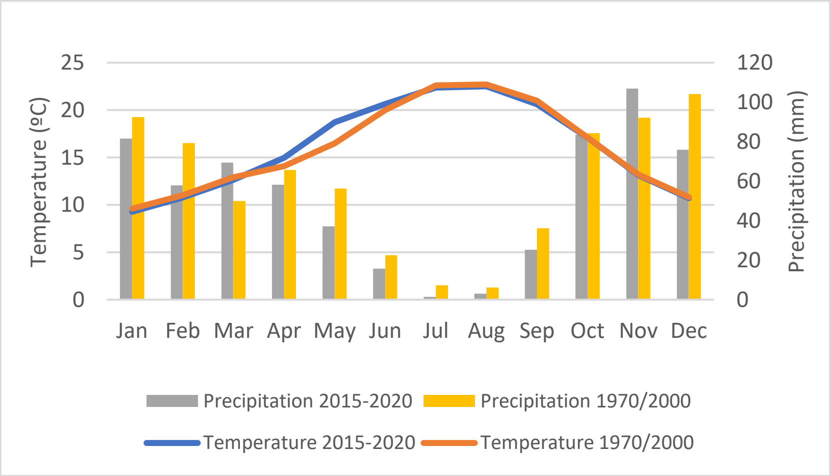


Figure 10: Ombrothermic diagram. The meteorologic weather data of Santarém for the period of 1970/2000 and 2015-2020.

According to the rational climatic classification of Thornthwaite (1948), the study region is classified as a dry-humid climate (mesothermal B4). Following figure 10, the Koppen (1936) scale classifies this region as temperate since it is rainy and moderately warm, with intense winter rains (Type Cs). Since the average of the warmest month is above 22ºC, its subtype is classified as Csa: humid mesothermal, with warm and dry summer and a cool and rainy winter.

2.3. Yield Data

The maize yield data was collected in the three studied parcels between 2015-2020 . A few years of data is missing due to wild boars and heavy rust attacks. The maize yield data was collected with a Fendt harvester model 5275c and processed by an *Ag Leader* yield monitor. The yield monitor is a fundamental part of producing high quality yield data because it receives information from the various combined sensors (figure 4) to compute the final yield maps. The maximum, minimum, mean and standard deviation of all yield maps are described in table 4.

Table 4: Descriptive properties of the yield maps

Field	Area (ha)	years	Descriptive statistics			
			Max	Mean	Min	SD
Cerca	23.3	2016	21.62	15.86	4.92	2.78
		2018	21.28	17.18	9.72	1.42
		2020	21.06	18.08	12.76	1.09
Lourenço	17.9	2015	20.85	17.57	8.40	1.19
		2016	20.83	18.04	9.36	1.35
		2018	20.88	17.13	10.56	1.13
		2019	21.66	17.69	12.26	1.22
		2020	19.64	16.53	13.47	0.92
Vinha	13.25	2016	19.21	15.77	5.25	1.74
		2018	16.13	12.87	6.45	1.21
		2019	19.12	14.53	7.51	1.64
		2020	17.24	14.85	8.20	1.24

To map the yield variabilities and quantify them, we initially built three types of different yield maps and did a visual assessment to figure out if any kinds of patterns emerged. Following the assumption that permanent soil characteristics directly impact the spatial yield, repeating yield patterns emerge with multiple years of data (Ping & Dobermann, 2005). If it does not, the yield pattern fades off after a few years of accumulated data (Blackmore et al., 2003). Three yield maps were developed to assess this hypothesis: a standardized multiyear yield map, a frequency yield map and a spatiotemporal yield map. Different types of yield maps map out different patterns because they are built differently. Upon visual confirmation, if the same pattern is present in all three, we follow the Maxent analysis.

A buffer was applied around the yield maps to remove the low yield areas typically affected by soil compaction. This compaction is particularly relevant at the edges of parcels due to machine manoeuvres (tractors, trailers and harvesters). The buffer zones were 20m for Cerca, Lourenço and Vinha. QGIS was used to create the multiyear yield maps and the buffer zones.

2.3.1. Standardized multiyear yield map

The yield data was first standardized according to Blackmore (2000) to compare yield patterns across years. Following equation 8, the relative yield (S_i) was determined by dividing the yield data of each location (y_i) by the average yield of that particular field for that year ($\overline{y}$):

$$S_i = \frac{y_i}{\overline{y}} * 100 \qquad \text{(Eq.8)}$$

After the yield maps for all years were standardized, the yield values were aligned then averaged according to the number of years of data available for each field. The rationale for standardization had a twofold objective. The first was to remove the climate influence from the data since several years of data are being aggregated. Different years have different climate influences, and the climatic influence is not the focus of this work. The second was to put the data from different years on a relative, comparable scale**.** The resulting maps will look like figure 11.

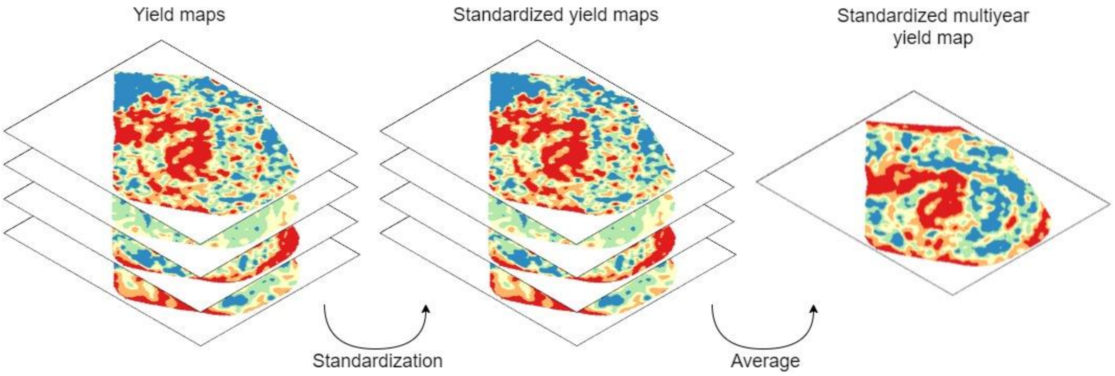


Figure 11: Standardization process for the construction of a multiyear yield map. The maps of several years are standardized, then a simple average of the resulting maps is done

2.3.2. Yield frequency map

A yield frequency map was computed using an adapted method of Franzen et al. (2008). For each year, the mean and standard deviation is determined. Then, using QGIS, the pixels are given a value of +1, 0 or -1, depending on whether the average of the pixel was greater than the field average within one standard deviation (+1); whether if the pixel was between one standard deviation of the average (0); or if the pixel is less than the field average within one standard deviation (-1). After constructing the yield frequency maps for the years available, the data was summed up, resulting in a frequency yield map (figure 12).

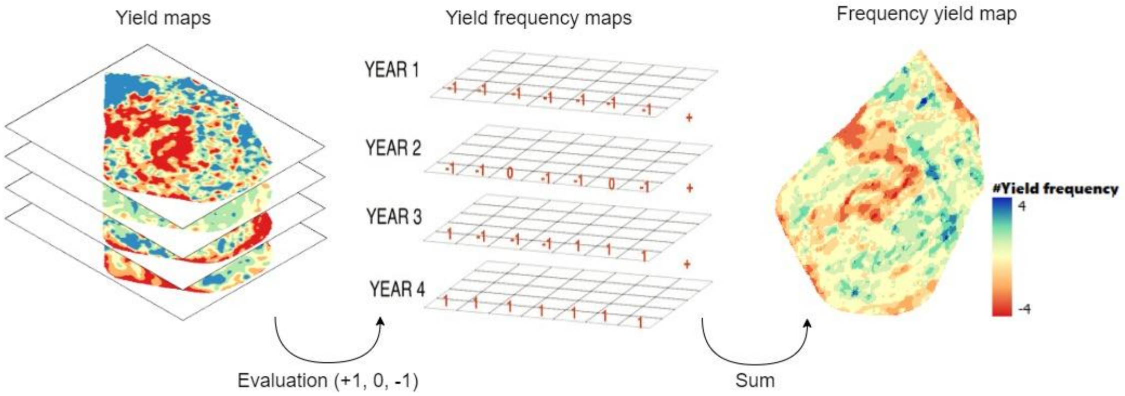


Figure 12: Frequency yield map construction process.

2.3.3. Spatio-temporal yield map

The spatial-temporal variability was assessed according to Blackmore et al. (2003) and Clay et al. (2017), where a yield variability map is combined with a temporal stability map. For the spatial variability, we used the map calculated in 2.3.1. Then, following Whelan and McBratney (2000), temporal stability was evaluated according to equation 9:

$$\sigma_i^2 = \frac{\sum_{i=1}^{n}\left(Y_{i,n} - \overline{Y}\right)^2}{n-1} \qquad \text{(Eq.9)}$$

Where σ_i^2 is the temporal variance at point *i*, Y_i is the maize yield at point *i* at year *n, the* $\overline{Y}$ is the mean yield for all selected harvest years, and *n* is the number of harvest years.

The grand mean of the spatial yield map dataset and the grand standard mean of the temporal yield map dataset are determined. Following figure 13, the regions of the spatial yield map below or above the grand mean are attributed a unique identifier, 1 and

2 respectively and regions on the temporal yield map below or above the grand standard mean are attributed 1 and 3.

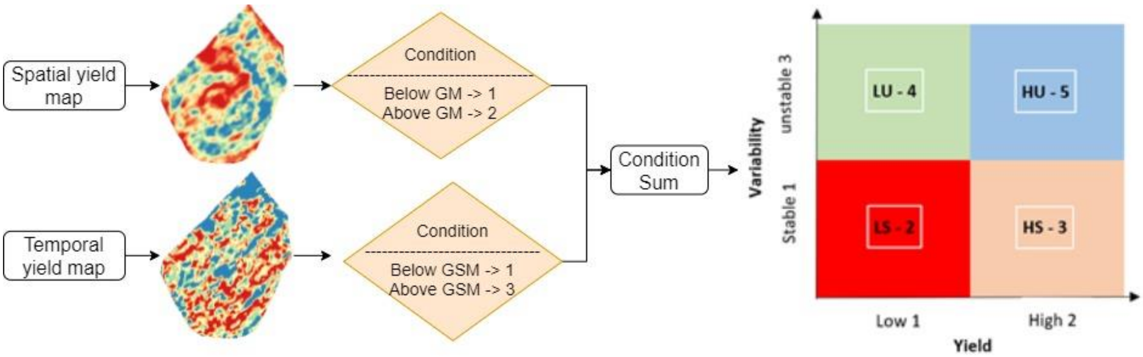


Figure 13: Workflow for the spatio-temporal yield maps, with the four possible conditions for each map.

As yield and stability are not exclusive, summing up the spatial and the temporal identifiers, the resulting spatial-temporal yield map will have four possible combinations to describe the existing variability, each with a unique identifier (figure 13): high yield and stable temporal variability (HS - 3); high yield and unstable variability (HU - 5); low yield and stable variability (LS - 2) and low yield and unstable variability (LU - 4). The resulting spatio-temporal yield maps will look like figure 14.

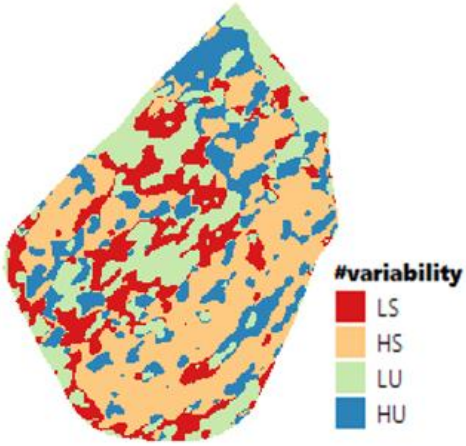


Figure 14: Spatio-temporal yield map for Vinha with the four possible combinations for spatial and temporal variability.

2.4. Soil Data

The sampling and processing of the soil, digital elevation models and electrical conductivity data was made available by a precision farming company, *Agroanalitica.* All rasters were in 2x2m resolution and scripts were developed in R *4.1.1* (R Core Team) using the *raster* package (Hijmans et al., 2015) to convert the rasters into ASCII format, snap the rasters to the same size, extend, orientation, projection and to extract the number of occurrences according to percentile and yield classes. Each of the environmental layers must be in the form of an ASCII grid with matching raster cell size and grid placement to be compatible with the Maxent software (Phillips, 2005).

The soil dataset consists of soil fertility information (pH, total phosphorus, potassium, magnesium, organic matter) a digital elevation model to characterize the topography, and the soils apparent electrical conductivity layer (table 5). Limestone was

Table 5: Descriptive statistics of the soil properties in the studied field (Cerca, Lourenço, and Vinha).

Field	Variable	unit	Descriptive statistics			
			Max	Mean	Min	SD
Cerca	Limestone	g/kg	0.531	0.29	0	0.18
	Electrical conductivity	mS/m	30.5	14.39	7.19	4.54
	Digital elevation model	m	70.05	69.2	68.41	0.36
	Total Phosphorus	mg/kg	453.1	243.9	116.8	76.5
	Magnesium	mg/kg	130.9	98.4	79.6	14.3
	Organic matter	%	1.56	1.09	0.65	0.19
	pH	---	7.96	7.73	7.36	0.08
	Potassium	mg/kg	265.7	203.2	132.7	25.4
Lourenço	Electrical conductivity	mS/m	43.41	24.147	10.6	5.66
	Digital elevation model	m	69.91	68.75	67.64	0.61
	Total Phosphorus	mg/kg	135.81	85.72	43.23	13.93
	Magnesium	mg/kg	240.57	178.11	121.71	28.55
	Organic matter	%	2.20	2.06	1.87	0.06
	pH	---	7.92	7.78	7.39	0.10
	Potassium	mg/kg	174.1	138.4	99.2	17.76
Vinha	Electrical conductivity	mS/m	48.4	18.97	12.34	4.09
	Digital elevation model	m	70.19	68.41	67.05	0.92
	Total Phosphorus	mg/kg	157.01	76.4	32.17	26.9
	Magnesium	mg/kg	298.6	169.6	108.1	38.4
	Organic matter	mg/kg	2.85	1.72	0.97	0.54
	pH	---	7.88	7.57	6.83	0.19
	Potassium	mg/kg	262	173.1	106.9	35.2

Existing variables for each field

only available for Cerca. From the DEM, several primary and secondary topographical attributes were extracted. The primary attributes are known to impact yield, and the secondary parameters describe and quantify the spatial variability of specific processes in the fields.

2.4.1. Topographic attributes

Before deriving topographic attributes used for this analysis based on elevation data (table 6), pre-processing was required to remove depressions known as sinks in the data. Sinks are local elevation minima without lower neighbourhoods that exist in a grid DEM and most of these depressions are artefacts that have an undesirable effect of altering the simulated flow networks (Tarboton et al., 1991; Tribe, 1992). There are different strategies to deal with sinks, such as depression filling, breaching or a combination of both (Wang et al., 2019).

In this study, a few depression filling and breaching algorithms were tested to remove the sinks from the data, such as the ones proposed by Jenson and Domingue (1988), Lindsay (2016), Planchon and Darboux (2002) and Wang and Liu (2006). Still, all of them were warping the terrain too much. According to Lidberg et al. (2017), breaching creates the most accurate stream networks on all resolutions, whereas filling is the least accurate. Following his advice, I used the algorithm 'Breach depressions least cost' proposed by Lindsay (2020), achieving a non-warped DEM free of sinks. This algorithm was implemented in WhiteboxTools 2.0.0 (Lindsay, 2018) using the plugin *WhiteboxTools for Processing* for QGIS 3.18.3.

Table 6: Topographic attributes computed from the DEM data. Adapted from (Wilson & Gallant, 2000b)

Topographic Attributes	Unit	Definition
Primary attributes		
Altitude	meters	Terrain elevation
Plan Curvature	deg.m-1	Contour curvature
Profile Curvature	deg.m-1	Slope profile curvature
Aspect	degree	Direction that a slope faces
Slope	degree	Percent change in that elevation over a certain distance
Secondary attributes		
Topographic Wetness Index	dimensionless	Estimates areas of water concentration
Distance to Flow Lines	meters	Yield relation to flow accumulation lines
Topographic Position Index	dimensionless	Landscape characterization

2.4.1.1. Primary Attributes

After removing the sinks from the DEM, the DEM was introduced into the SAGA-GIS 2.3.2 (System for Automated Geoscientific Analysis) to extract several topographic attributes in raster format that are known to have an impact on yields, such as aspect, slope, profile and planar curvature (Olaya, 2004).

The slope represents the rate of elevation change for each cell in the DEM. It was calculated in degrees using the method of Haralick 10 parameters 3° order polynomial.

Profile curvature is a measure of the curvature in the vertical plane (figure 15, a). Positive values indicate that the surface is upwardly concave, and negative values indicate upwardly convex. Planar Curvature is the curvature in a horizontal plane and measures the topographical concavities (figure 15,b). Positive values indicate that the surface is laterally convex, and negative values indicate that a surface is laterally concave (Wilson & Gallant, 2000b). Plan and profile curvature was calculated using the method of Zevenbergen and Thorne (1987).

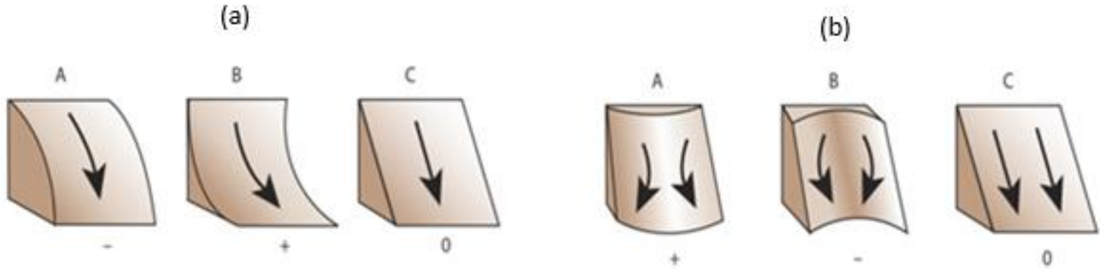


Figure 15: Measures of terrain curvature: (a) Profile curvature (b) Planar Curvature

Aspect describes the orientation of the slope in degrees starting clockwise from the north (Wilson & Gallant, 2000b). Since aspect is a circular variable, making aspect a suitable parameter for inclusion in the analysis is necessary. It was converted into two linear components: aspect Eastness and aspect Northness. The aspect raster is first converted into radians by multiplying it by π and dividing by 180. Then the Eastness is determined by applying the sine to the aspect layer, and the Northness by applying the cosine. Northness is an index from +1 to −1 of how north (+1) or south (−1) a site faces. Eastness is an index from +1 to −1 of how east (+1) or west (−1) a site faces. Eastness and Northness were calculated using QGIS 3.18.3.

2.4.1.2. Secondary Attributes

Topographic Wetness Index

The basic concept of the Topographic Wetness Index (TWI) is the expression of terrain mass-balance of the catchment water supply and local drainage. This index assumes steady-state conditions and quantifies soil water distribution's tendency, which is affected by topography and is determined according to equation 6:

$$TWI = ln\left[\left(\frac{TCA}{FW}\right)/tan\,(S)\right] \qquad \text{(Eq.6)}$$

The TWI has three key components: total catchment area (TCA), flow width (FW), and slope gradient (S) (figure 16). TCA determines the size of the upslope area (number of cells) draining into a given cell, FW is the length of a contour orthogonal to the flow from the cell, and S is the slope of the focal cell or the slope between the focal cell and a further cell downslope (Gruber & Peckham, 2009).

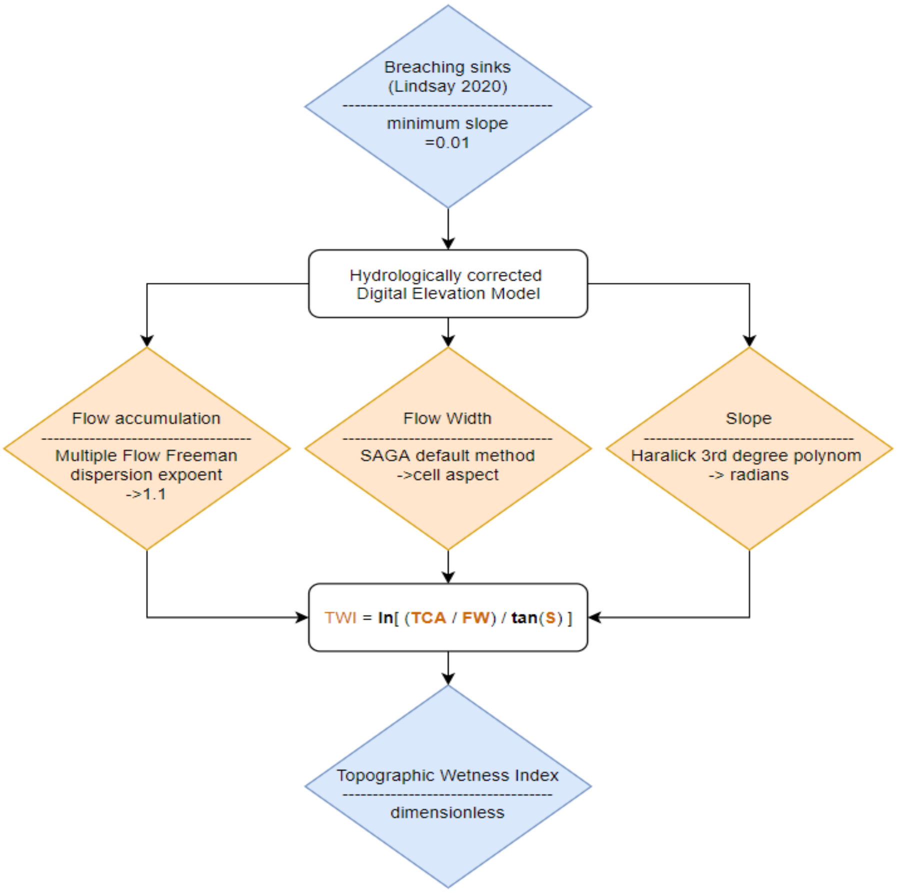


Figure 16: TWI quantifies the terrain mass-balance of the catchment water supply and local drainage. Blue font indicates the name of the tools used, and the orange font indicates the three TWI components used. Adapted from: Kopecký et al. (2021)

A flow accumulation algorithm is used to calculate de TCA, and it establishes the direction of the flow for every cell. Flow accumulation can be calculated using either single flow (SF) or multiple flow (MF) algorithms (figure 17). They can describe the downslope water movement, following the path of the steepest descent. The most common SF algorithm is the deterministic eight-direction, mainly used for approximating flow directions in a topographic surface. Its simplicity lies in the use of discrete flow angles, and each pixel has a single flow direction that does not capture the effect of divergent flow over hillslopes (figure 17,a). A SF algorithm assumes that subsurface flow occurs only in the steepest downslope direction from any given point, while MF diverts the flow to multiple downslope cells in proportion to the slope between them (figure 17,b) (Qin et al., 2007). A value of the exponent parameter in MF algorithms needs to be specified since it controls the degree of dispersion in the resulting flow-accumulation grid (Qin et al., 2007).

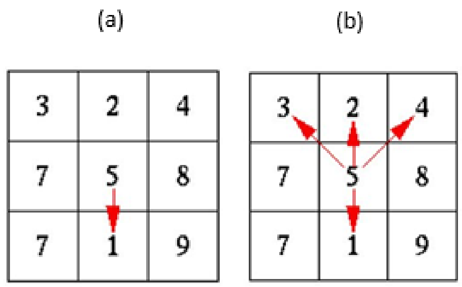


Figure 17: Direction of the flow (a) Single flow direction (b) Multiple flow direction. Adapted from Stojanovic and Stojanovic (2019)

To calculate the TWI, we followed the guidelines of Kopecký et al. (2021) using the open-source software SAGA-GIS (figure 16). The choice of flow accumulation algorithm is a critical parameter in the TWI calculation, followed by the slope gradient and the flow width. Kopecký et al. (2021) compared SF and MF's performance and found that the best performing MF algorithms explained twice as much variance in the measured soil moisture than the SF algorithms. The MF-Freeman's best performing algorithm was chosen with a flow accumulation unit of many cells. The chosen flow dispersion exponent value was 1.1 because it performed substantially better when compared to the higher values previously recommended Kopecký et al. (2021). Then, the TCA was converted into a specific catchment area using the flow width calculated with the SAGA-GIS default method based on cell aspect. Finally, the chosen slope method used the SAGA Haralick 10 parameters 3° order polynomial, with the slope gradient in radians. To calculate the final TWI maps, we used the raster calculator from QGIS 3.18.3, following equation 6.

Distance to Flow Accumulation Lines

Distance to flow accumulation lines (DFL) is a secondary attribute developed by (Da Silva & Silva, 2006). The calculation of this indicator involves determining the flow accumulation, deciding on the proper threshold to keep the main flow lines, and calculating the distance away from the flow accumulation areas (figure 18).

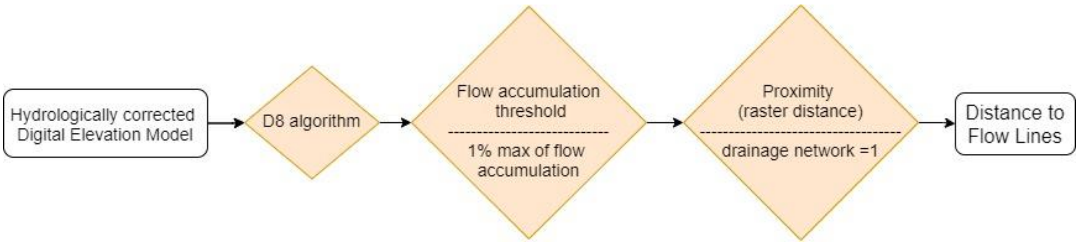

Hydrologically corrected Digital Elevation Model
D8 algorithm
Flow accumulation threshold
1% max of flow accumulation
Proximity (raster distance)
drainage network =1
Distance to Flow Lines

Figure 18: Workflow for calculating the Distance to flow accumulation lines.

The flow accumulation algorithm was the D8, due to its simplicity of use and its capacity to model the watershed draining structures (Turcotte et al., 2001). From the resulting flow accumulation lines (figure 19a), a flow accumulation threshold (FAT) was applied to extract the drainage network. The FAT is a user-defined parameter that directly affects the structure of the drainage networks extracted from DEMs (Ozulu & Gökgöz, 2018). There are several types of FATs and the chosen method was to apply 1% of the maximum flow accumulation value (Maidment & Morehouse, 2002), with the resulting drainage network outlined in figure 19b.

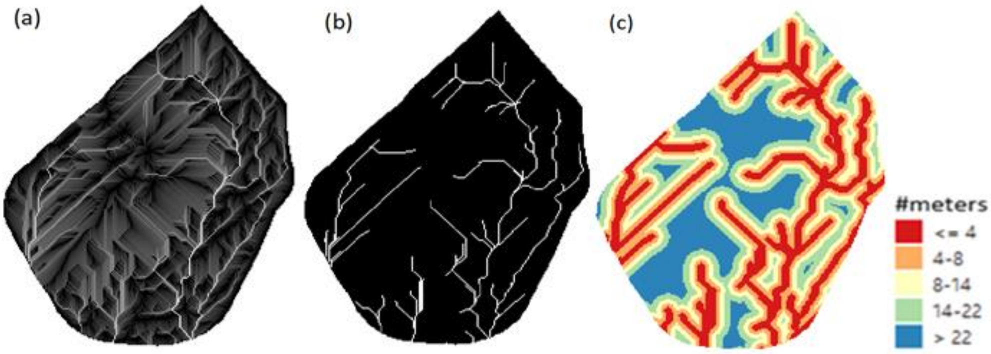


Figure 19: Resulting steps for the calculation of the DFL for Vinha: (a) log10 of Flow accumulation, whiter represents more water accumulation (b) 1% of Flow accumulation threshold (c) Distance to flow accumulation lines.

Following the FAT application, it is necessary to determine the distance to the drainage network by computing the function *Proximity* (raster distance). We assigned a pixel value of 1 to the drainage network, and the distance was computed from the georeferenced coordinates of said pixels, resulting in a continuous raster layer of distances to the water flow lines (figure 19c). The FAT and Proximity were computed using QGIS 3.18.3.

Topographic Position Index

The TPI compares the elevation of each cell in a DEM to the mean elevation of a specified neighbourhood around that cell (Weiss, 2001). According to Mieza et al. (2016), in regions where yield variability correlates with local minima and maxima topographic values, the TPI had a better explaining power than topography. So, taking this into account, we developed the TPI using equation 7:

$$TPI = \left(h - \left(x_h(r)\right)\right) \qquad \text{(Eq.7)}$$

Where h is the elevation of a grid cell in a meter above sea level, x_h is the mean elevation of grid cells in the neighbourhood with radius r. Positive values represent locations higher than the average of their surroundings, and negative values represent locations lower than their surroundings. Values near zero are flat areas or regions with constant slopes within the neighbourhood (figure 20). The radius controls the scale of the analysis, deciding what cells are to be considered “around” the cell. The chosen radius was 100 map units (200 m) obtained using the SAGA-GIS. We decided that this threshold was enough to reflect the elevation variability of the field.

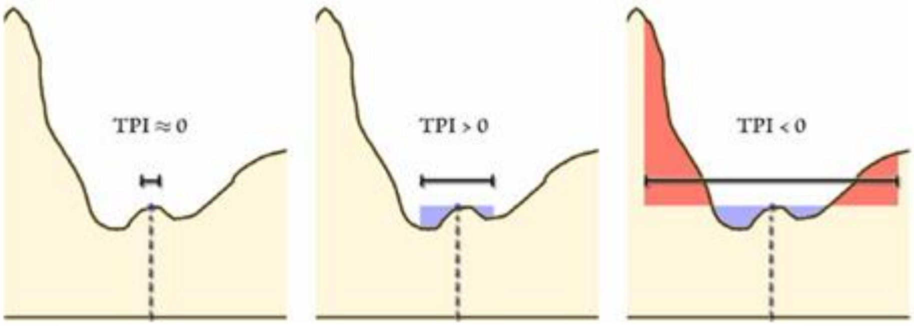


Figure 20: How the choice of the radius impacts the TPI evaluation of its immediate surroundings.

2.5. Yield gap approach

The high and low productivity areas were identified upon visual confirmation of similar variabilities in the three yield maps (Annex I). Following a yield gap approach, from the yield map calculated in 2.3.1 the existing maximum yield gap between the high and low productivity areas were quantified as 44% for Cerca, 24% for Lourenço and 36% for Vinha. The best yields were identified using the attainable farm yield (Y_{at}), with the upper 10th percentile as the benchmark (Van Ittersum et al., 2013). The lower 10th percentile productivity areas were used as the worst farm yield and identified using the histogram of the standardized yield maps (figure 21) calculated in 3.3.1. Then we did the same for the 15th percentile and 20th percentile, for the fields in this study.

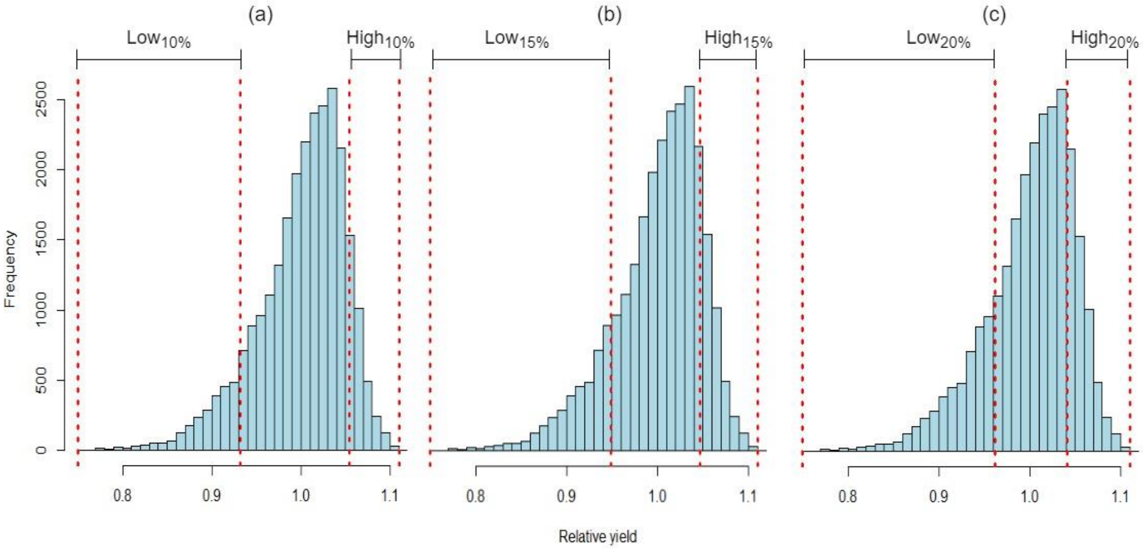


Figure 21: High and low yield classes for: (a) 10th percentile, (b) 15th percentile and (c) 20th percentile for Vinha.

Six different sets of georeferenced points ($High_{10,15,20\%}$, $Low_{10,15,20\%}$) were extracted for each field (table 7) to understand the main drivers of the within-field variability of the fields using the Maxent algorithm.

Table 7: Percentile and Number of presence points selected from each yield class

Percentile	Yield class	Number of presence points (#pixels)		
		Cerca	Lourenço	Vinha
10	Low	4973	3708	2638
	High	4974	3708	2638
15	Low	7460	5561	3956
	High	7460	5561	3956
20	Low	9946	7415	5275
	High	9947	7415	5275

2.5.1. Maxent Modelling

Modelling was done using the maximum entropy method with the Maxent software 3.4.4 (Phillips, Dudík, et al., 2017). Maxent is a general-purpose machine learning method that relies on presence-only and background data (Phillips, Anderson, et al., 2017; Phillips et al., 2006; Phillips & Dudík, 2008) to identify better variables correlate with the occurrence records. Occurrence data was previously obtained from the standardized yield maps, soil fertility parameters, topographical and soil electrical conductivity maps were the environmental layers used.

2.5.2. Variable Selection

A sequential approach was chosen to assess the models variables with the most information. By improving variable selection, collinearity can be minimized (Feng et al., 2019) because if two variables are highly correlated, it becomes difficult to separate the individual effects of each variable.

The variables were initially screened for correlations in combination with the jack-knife procedure of preliminary maxent models (Gąsiorek et al., 2021; Raghavan et al., 2019). The default parameters of Maxent were used in the preliminary models, and each model had its specific occurrence data set (table 7). The jack-knife and the correlation coefficients were used to select the variables to retain in the modelling approach. If two variables were correlated (r≥ 0.8), a highly contributing variable was chosen, and the other was removed. Next, we proceeded to fit another maxent model to remove variables that had a low contribution to model gain (less than 1% of percentage and permutation importance). Finally, multicollinearity was evaluated using the variance inflation factor (VIF) with a stepwise approach. Variables with a VIF < 10 were retained since higher values indicate multicollinearity and increase the risk of overfitting by the model (Dormann et al., 2013).

The correlation and VIF analysis were done in R *4.1.1* (R Core Team) using the *usdm* (Uncertainty Analysis for Species Distribution Models) package (Naimi et al., 2014) and the *raster* package (Hijmans et al., 2015).

2.5.3. Model calibration, evaluation and creation

Maxent has several modifiable parameters and while many studies still use the default settings (Morales et al., 2017), there is a growing amount of evidence that using the default parameters may not generate the best models (Morales et al., 2017; Radosavljevic & Anderson, 2014; Syfert et al., 2013).In addition, the area under the receiver operating characteristic curve, known as the AUC, is considered the standard method to assess the accuracy of predictive distribution models. It avoids the subjectivity in the threshold selection process by summarizing the overall model performance over all possible thresholds. However, has been heavily criticised for weighting omission and commission errors equally and giving no information about the spatial distribution of model errors (Lobo et al., 2008). Additionally, the AUC is only informative when true instances of absence are available (Jiménez-Valverde, 2012). Since this information usually does not exist, the use of background data in place of true absences can produce misleading results (Lobo et al., 2010). AUC values have also been shown to inflate systematically when the size of the study is increased relative to the extent of the geographical range of the organism in question (Lobo et al., 2008). Consequently, SDMs with high AUC values and excellent predictive performance can be obtained irrespective of whether the models identify plausible or causal relationships between environmental predictors and the distribution of the species (Veloz, 2009).

Preliminary models were calibrated using the Maxent algorithm with the Kuenm package in R (Cobos et al., 2019), avoiding the previously mentioned drawbacks. This package allows the fine-tuning of the two main modifiable parameters that control complexity, the regularization multiplier (RM) and the feature classes (FC), instead of the default parameters. The statistics of model performance implemented in Kuenm are partial ROC as a measure of statistical significance, omission rates (OR), and Akaike Information Criterion (AICc). Since AUC is not an appropriate measure in ENMs (Jiménez-Valverde, 2012; Lobo et al., 2008), the partial ROC is a more suitable indicator of statistical significance and is determined by a bootstrap resampling of 50% of testing data with 500 iterations, where probabilities are assessed by direct count of the proportion of bootstrap replicates for which a AUC ratio with values ≤1 reflects predictions indistinguishable from random predictions. But a ratio >1 indicates predictions that are better than random (Peterson et al., 2008).The OR is used as a measure of performance, indicating how well models created with training data predict test occurrences (Anderson et al., 2003). AICc is used to evaluate how well models fit to the data while penalizing complexity to favour simple models. Models selected by these three metrics will maximize model performance and simplicity (Cobos et al., 2019).

The purpose of the calibration is to evaluate the best potential combination of selectable parameters in Maxent to select the most appropriate model. The RM affects how focused or closely fit the output distribution is. It is a penalty that occurs in the form of a β regularization parameter specific to each feature class, limiting complexity and protecting against overfitting (Anderson & Gonzalez Jr, 2011; Phillips et al., 2006;

Warren & Seifert, 2011). The RM is a coefficient applied to the value of the respective β parameter of each feature class, altering the overall level of regularization rather than changing the β parameters individually. FC corresponds to the mathematical transformation of the different covariates used in the model to allow complex relationships to be modelled. These relationships can be linear (l), quadratic (q), product (p), threshold (t) and hinge (h) (Elith et al., 2010; Merow et al., 2013).

For the preliminary models for each yield class of table 7, a specific set of variables was selected according to section 3.7.1. 255 models were created for each yield class by combining 17 RM (0.1 to 1, 2 to 6, 8 and 10) and 15 possible combinations of four feature classes (l, q, p, h, lq, lp, lh, qp, qh, ph, lqp, lqh, lph, qph, lqph). The threshold feature was not included since this appears to improve model performance and results in smoother and simpler models, hence more likely to be realistic (Phillips, Anderson, et al., 2017). The models were calibrated using 70% of the records for training and 30% for testing.

The best models were evaluated according to: (1) Statistical significance using partial ROC (Peterson et al., 2008) (2) Predictive ability is evaluated through the use of OR at a threshold of 5% (Anderson et al., 2003); (3) AICc is used to evaluate model complexity and models with delta AICc values of $\leq$2 were selected (Warren & Seifert, 2011). When more than one best model was obtained, the best one based on the highest value of the AUC mean ratio was selected.

Then, using the best parameter settings selected during model evaluation, the final models were created using a 10 cross-validation replicates. All models were set with 500 iterations. The number of background points adapted to the study area extended automatically and model output is cloglog, an estimate of occurrence probability (Phillips, Anderson, et al., 2017). The Maxent output for the model is a map representing the environmental suitability of the available area through a continuous index (Sillero, 2011) of very suitable (value 1) to unsuitable (value 0) and the average of the 10 model outputs was used as the basis of interpretations. Model overfit was evaluated considering a threshold-independent measure, AUC_{DIFF} (Warren & Seifert, 2011), which is the $AUC_{training}$ minus AUC_{test}. Overfit models generally perform well on training data but poorly on test data, so low AUC_{DIFF} values indicates that the model is not overly specific to the training data, which reduces the risk that the model is over-parameterized (Warren & Seifert, 2011).

2.5.4. Final Model

The contribution of the variables to the final model was evaluated using permutation importance and the jack-knife. Permutation importance (PI) was used to determine the main impact factors on yield. Factors with more than 10% of PI were considered primary impacting factors. For each variable, the PI is determined by randomly permuting the values of that variable among the training points (presence and background), and the decrease in the training AUC is measured. A large decrease indicates that the model depends heavily on that variable (Phillips et al., 2006). Searcy and Shaffer (2016) have shown that the PI is biologically realistic, reflecting the role of the most critical variables in defining the species' environmental niche.

The jack-knife was used to assess variables contributing less than 10% of PI to the model. These variables were considered secondary impacting factors. The jack-knife analysis systematically excludes a variable in each run, and the model is created with the remaining variables. Then, a model is created using each variable in isolation and another one with all the variables. Each step increases the models gain by modifying the coefficient for a single feature, assigning the increase in gain to the environmental variable that the feature depends on (Phillips et al., 2006). Variables that showed a high model gain when used in isolation or a high drop in gain occurs when the variable is omitted were chosen to be analysed. But an issue remains: the jack-knife shows which variables have the most useful information for the model performance and does not consider the agronomical importance of the chosen variables.

Exploring the response curves for each predictor helps to understand which range of values are agronomically relevant. Maxent produces two sets of response curves that help to understand which range of values are agronomically relevant. When both sets are not similar, correlation among variables is present and should be considered when analysing the response curves. It is also necessary to define a cut-off value of the environmental suitability (ES) where values below this threshold are considered having low suitability for the yield class in analysis, so they are not analysed. It was decided on a threshold of 0.7 of ES for this effect.

In the average model, one variable varies with all the other variables set to their average value. The value shown on the y-axis is the predicted probability of suitable conditions, given in the cloglog output. But the model might depend on correlated variables in ways that are not evident in the response curves. Maxent produces a second set of response curves to make it easier to analyse correlated variables. The response curve is created by generating a model using only the corresponding variable and disregarding all the others (Phillips et al., 2006).

3. Results and discussion

The results obtained in this work will be structured in three sections, one for each field, where i present the main results of the high and low yield models, followed by the discussion. For all the images concerning yield maps, yield suitability, topography, electrical conductivity and fertility maps, we forward them to figure 9 if it is necessary to check the scale and the geographic north.

Six different yield class models (High$_{10,15,20}$, Low$_{10,15,20}$) were tested for each field, with 255 different models built for each yield class. The selected models for the highly productive areas were statistically significant with an OR lower than the 5% threshold, and among these, the least complex model was chosen. The same process was applied to the low yield models. As an example, Table 8 shows the results of Cerca obtained with the Kuenm package.

Table 8: Model selection process for the high and low yield classes of Cerca.

Percentile	TCM	SSM	MOr	MAIC	SSM + MOr	SSM + MAIC	SSM + Mor + Maic	Selected model
High yield models								
10	255	255	7	1	7	1	1	RM_0.2_F_lq
15	255	255	0	1	0	1	0	RM_0.1_F_lq
20	255	255	158	1	158	1	1	RM_0.2_F_ph
Low yield models								
10	255	255	126	1	126	1	1	RM_0.2_F_qp
15	255	255	78	1	78	1	1	RM_0.1_F_qp
20	255	255	61	1	61	1	1	RM_0.2_F_lph

TCM = total candidate models; SSM = statistically significant models; MOr = models that satisfy the criterion of omission rate, MAIC = models that satisfy the AICc; AICc = Akaike information criterion; RM = selected regulation multiplier; F = selected features

The final models produced in table 9 all showed good performance, except model high$_{15}$ which failed to reach an OR below 5% (table 8). The remainder of the models were statistically significant with an OR below 5.0%, had low standard deviations during model training and testing, while showing low overfit to the data. Finally, we selected the best models that characterize the high and low yields according to an OR below 5%. If 2 models have similar ORs, the least complex model is chosen according to the lowest AICc.

Table 9: Statistics of the best-selected models (goodness of fit) of Cerca.

Yield Class	Selected model	Mean AUC ratio	OR	AICc	AUC_{Train}	$Stdev_{train}$	AUC_{Test}	$Stdev_{test}$	AUC_{Diff}
High yield models									
10	RM_0.2_F_lq	1.41	0.049	99171.69	0.7573	0.000683	0.7568	0.00027	0.0005
15	RM_0.1_F_lq	1.369	0.051	150589.5	0.7153	0.000539	0.715	0.00907	0.0003
20	RM_0.2_F_ph	1.442	0.049	199849.6	0.6964	0.000382	0.6946	0.00505	0.0018
Low yield models									
10	RM_0.2_F_qp	1.459	0.045	98129.4	0.7705	0.00128	0.7701	0.00814	0.0004
15	RM_0.1_F_qp	1.364	0.047	149902.7	0.7208	0.00129	0.7206	0.00820	0.0002
20	RM_0.2_F_lph	1.438	0.05	199893.7	0.7035	0.00025	0.7011	0.00498	0.0024

The same selection process was made for the models that best characterize the high and low yielding areas of Lourenço and Vinha, also achieving good model performances. Model results are available in annexe II.

3.1. Cerca Field

3.1.1. High Yield

The model used to characterize the high yielding regions of Cerca was the $high_{10}$. According to the percentual contribution (PC) of model gain, the results show that LS (57.7%) and EC_a (18.1%) have the highest contributions (table 10). The permutation importance (PI) shows that EC_a is the primary impacting factor, with 43.5%, followed by DEM, Mg and OM with 18.2%, 18.1 and 14.6%, respectively. These four factors explain 94.4% of the high yielding areas. LS showed a high PC but a low PI (0.5%).

Table 10: Environmental variables used for yield class model $high_{10}$ in Cerca.

Variable	LS	EC_a	TPI	OM	K	Mg	pH	DEM	SL
Percent Contribution	57.7	18.1	6.5	5.6	4.3	3.1	1.8	1.5	1.3
Permutation Importance	0.5	43.6	2.6	14.6	1.7	18.1	0.1	18.2	0.6

LS = Limestone; EC_a = Apparent electrical conductivity; TPI = Topographic position Index OM = Organic matter; K = Potassium; Mg = Magnesium; pH = Soil acidity; DEM = Digital elevation model; SL = Slope;

The Jack-knife (figure 22) indicates which secondary impacting factors appear to be relevant. From these, only LS and K show high model gain when used in isolation.

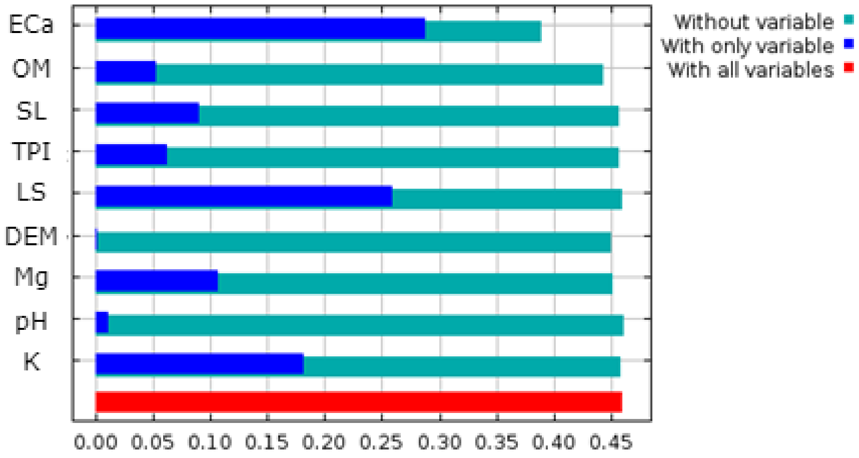


Figure 22: The Jack-knife test for evaluating the relative importance of environmental variables for yield class model $high_{10}$ in Cerca

The analysis identified several factors responsible for the high yielding areas, with different impacts in certain regions. Overall, the response curves show little correlation, with the high yielding locations characterized by figure 23, with LS of 0-0.5 mg / kg, EC_a of 17-25 mS / m, an OM of 1.1-1.5 %, a Mg of 100-130 mg / Kg, a K of 200-260 mg / kg. The DEM and LS presented correlation in their response curves. The single variable model for DEM indicates ES values below the 0.7 thresholds, decreasing ES values with increasing altitude. But the average model, when considering the other variables, suggests a high ES between the altitudes of 68.4-69.2m. Similarly, LS indicated high ES in the average model at all concentrations, while the single model showed high ES when LS is above 0.47 mg/kg.

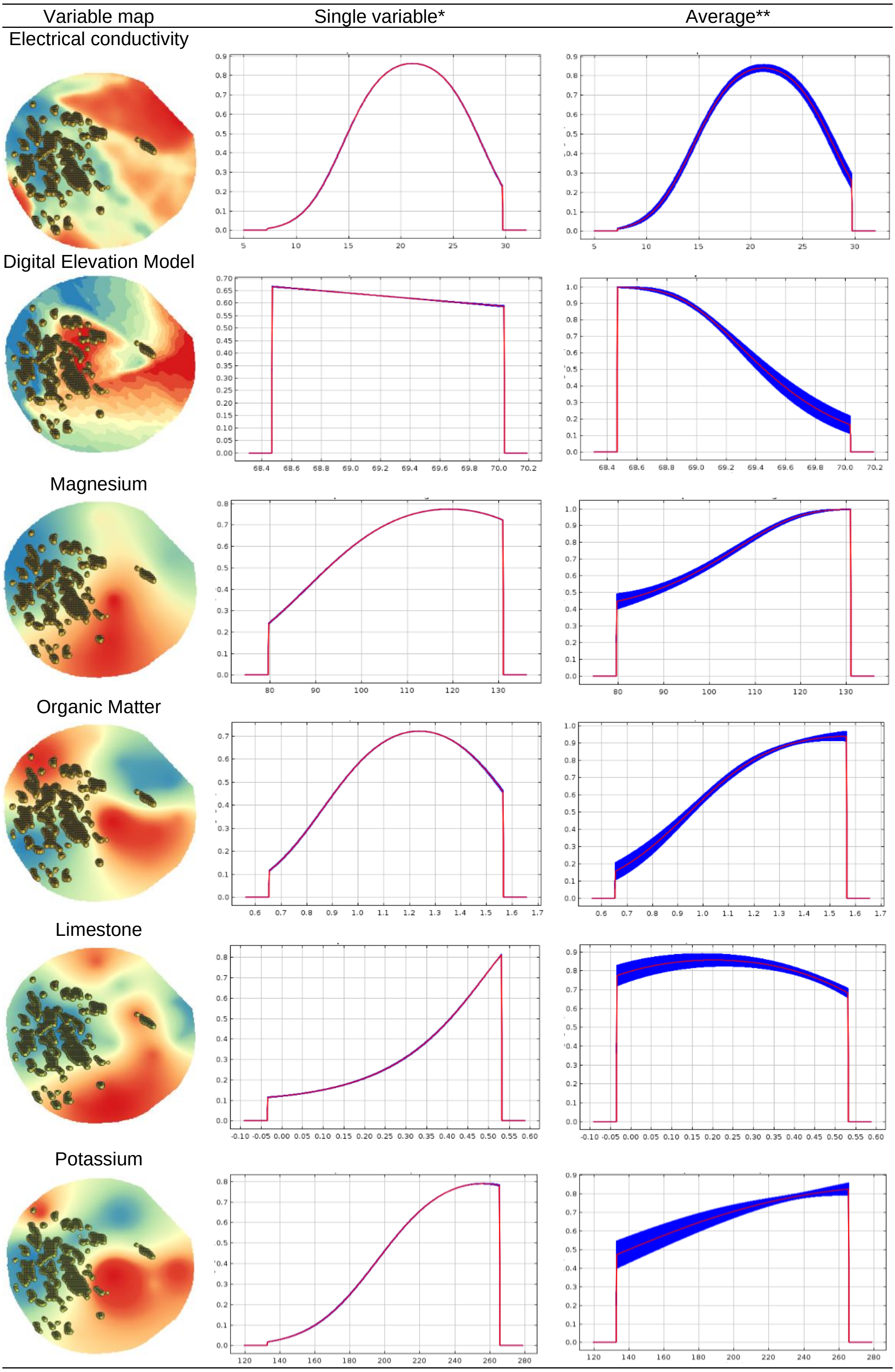


Figure 23: The environmental variables that control the high yield areas in Cerca. The dots in the variable map column represent the location of the presence points.

* Maxent model made with only the corresponding variable

** Maxent model where one variable is made to vary with all the other variables set to their average value

3.1.2. Low Yield

The model used to characterize the low yielding regions is the low_{10} model. PC shows that OM (36.1%) and EC_a (33.5%) contribute the most to model gain (table 11). PI shows that EC_a, P, K and TPI are the primary impacting factors for the low yield with 23.5%, 22.1%, 20.3% and 12.3%, respectively. These four factors explain 78.2% of the low yielding regions.

Table 11: Environmental variables used for yield class model low_{10} in Cerca.

Variable	OM	EC_a	TPI	K	pH	DEM	P	LS	Mg
Percent Contribution	36.1	33.5	8.2	5.6	4.5	4.4	P	2.1	2
Permutation Importance	4.5	23.5	12.3	20.3	2.9	1.8	22.1	6.1	6.5

OM = Organic matter; EC_a =Apparent electrical conductivity; TPI = Topographic position Index; K = Potassium; ; pH = Soil acidity; DEM = Digital elevation model; P = Phosphorus; LS = Limestone; Mg = Magnesium.

The Jack-knife (figure 24) indicates which secondary impacting factors appear to be relevant. From these, only OM show high model gain when used in isolation.

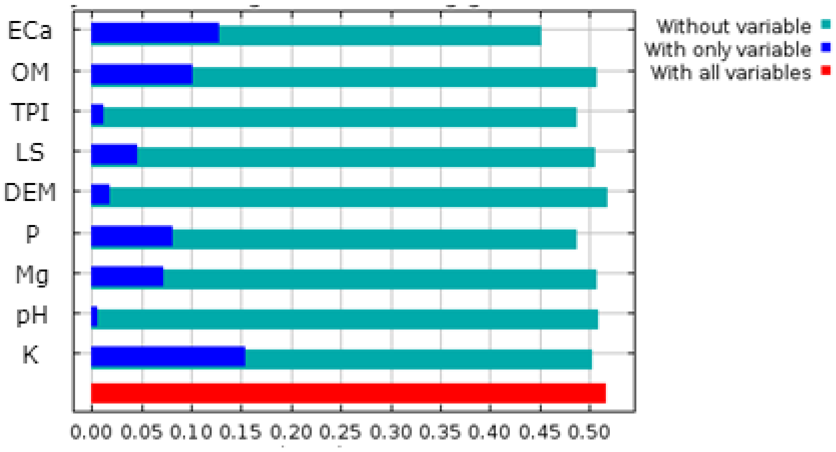


Figure 24: The Jack-knife test for evaluating the relative importance of environmental variables for yield class model low_{10} in Cerca

The response curves show little correlation between them, with the low yielding locations characterized by figure 25 with a EC_a of 7-12 mS/m, P of 120-150 mg/Kg, a K of 130-175, OM of 0.65-0.8% and a TPI of 0.1-0.35.

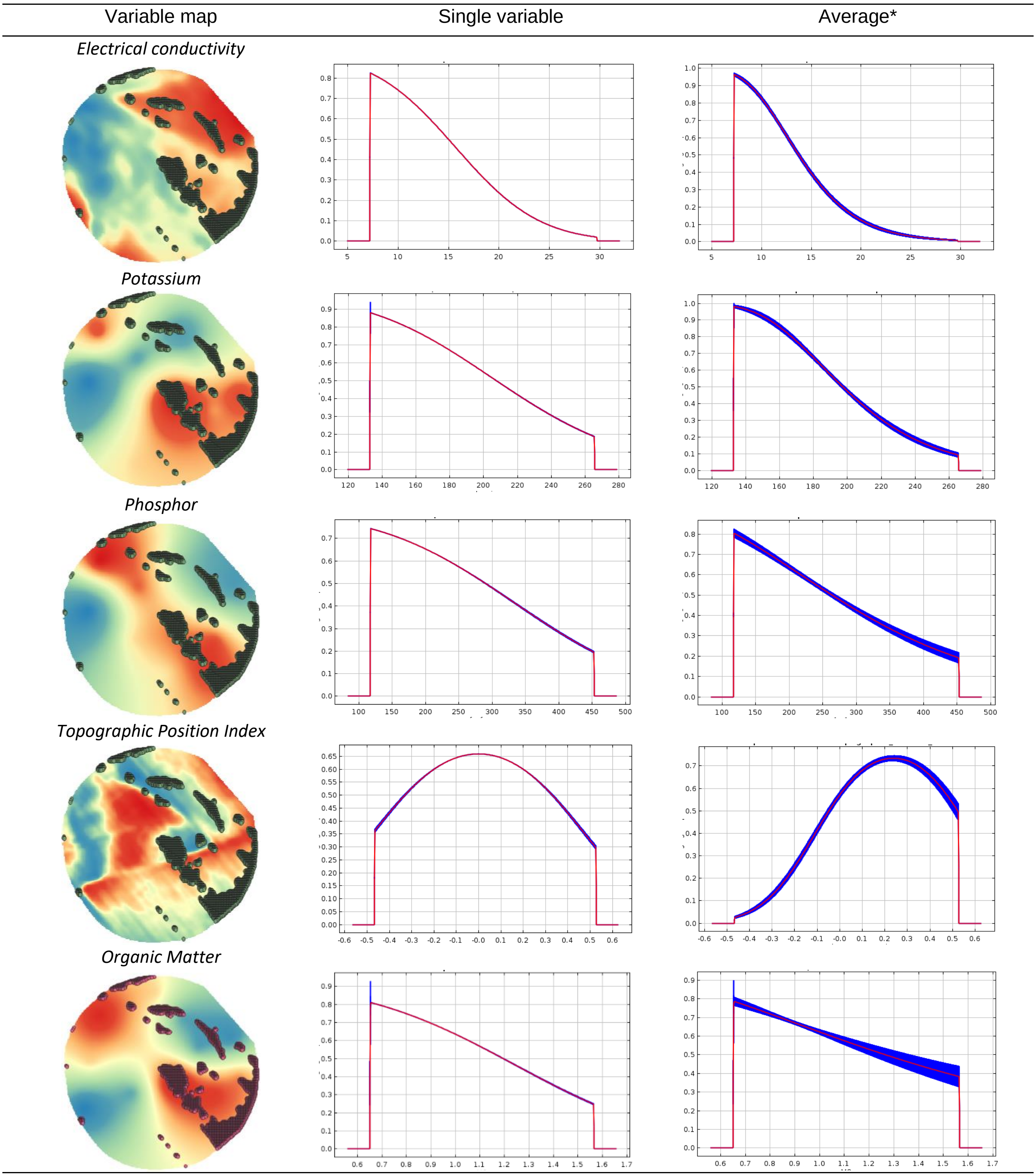


Figure 25: The environmental variables that control the low yield areas in Cerca. The dots in the variable map column represent the location of the presence points.

* Maxent model made with only the corresponding variable

** Maxent model where one variable is made to vary with all the other variables set to their average value

3.1.3. Discussion - Cerca

Apparent electrical conductivity (EC_a) is influenced by several soil properties such as salinity, texture, water content and bulk content (Corwin & Lesch, 2005; Sudduth et al., 2005). Given the importance that the models give to EC_a for high and low productivity areas, it appears that the stable soil properties directly impact the yield (Farahani & Buchleiter, 2004; Michael Mertens et al., 2008). The $high_{10}$ model indicates that high yielding areas have high environmental suitability (ES) in the range of 17-25 mS/m, while the low yielding areas were in the range of 7-12 mS/m (figure 26). EC_a in these ranges indicates that the texture of high yielding areas are characterized for having a higher clay content, and the low yielding areas possess fewer clay. Soils in the 17-25 mS/m range are characterized for being silty soils (Heege, 2013), and these tend to have higher water retention than sandy soils (Jalota et al., 2010; Kaspar et al., 2004), which is a favourable factor towards explaining the higher yields.

Regarding topography's impact on yield in Cerca, figure 23 indicates a high ES in the elevations of 68.4 – 69.2m for high yields, which appears to be a depression (figure 26). Regions with lower elevation than the surrounding areas tend to receive more water and be waterlogged in wet years or more productive in dry years (Kaspar et al., 2004; Maestrini & Basso, 2018). This indicates that the maize in this region is properly drained, receiving a higher level of water and nutrients from the surrounding areas, bolstering the yield. In opposition, the high TPI (0.1-0.35) (figure 25) indicates regions with higher relative altitude concerning their surroundings (figure 26). Higher elevations or summit regions are typically characterized for having lower yields (Bakhsh, Jaynes, et al., 2000; Hansen et al., 2013; Mishra et al., 2008) and the results indicate the same.

Although it is difficult to disentangle the OM effects on yield, several works have shown that building soil OM improves yields over time (Kaur et al., 2008; Majumder et al., 2008; Oldfield et al., 2018), although the increase in yield levels of at 2% of Soil OM (Lal, 2020; Oldfield et al., 2019).This is due to the improvement of several soil properties, including retaining water and nutrients, improving the water holding capacity and aeration, and minimising topsoil erosion (Doran & Zeiss, 2000; Lal, 2016; Philip Robertson et al., 2014). Our results for Cerca seem to be in line with the existing bibliography. Higher productive regions are situated in areas with OM around 1.1-1.5% (figure 23), while the low yielding areas have OM of 0.65-0.8% (figure 25). This indicates that it might be possible to increase the yields in this region if a correction is made.

Magnesium is responsible for several functions in maize (Cakmak & Yazici, 2010), such as supporting nitrogen uptake and simultaneously controlling processes responsible for photosynthesis and assimilate production and partition (Cakmak &

Kirkby, 2008; Gerendás & Führs, 2013; Senbayram et al., 2015). It is a component of the chlorophyl molecule, constituting 15-20% of the total leaf content (Farhat et al., 2016). Deficiencies in Mg can lead to impairments of growth and yield and having correct levels of Mg in the soil can improve crop yield (Wang et al., 2020). The high_{10} model identified Mg as a primary factor for the high yields with the ideal concentrations of 100-130 mg/kg (figure 23). The low_{10} model did not identify Mg as a responsible factor. This could be because potassium (K) and phosphor (P) are two main macronutrients necessary for plant growth and are shown to impact the low yielding regions.

Potassium (K) is the most abundant cation in plant tissues, and plays a key role in several regulatory systems, such as resistance to pests and diseases, photosynthesis, osmoregulation, enzyme activation, protein synthesis, phloem loading and transport and uptake (Amtmann et al., 2008; Epstein & Bloom, 2005; Zörb et al., 2014). K is considered a fundamental macronutrient for the proper growth, development and sustainable yield (Adnan, 2020; Marschner, 2011). K in the low_{10} model was identified as a primary factor (figure 25) with concentrations in the range of 130-175 mg/kg. The high_{10} model identified K as a secondary variable, with the ideal concentration of 200-260 mg/kg (figure 23). Improving the K concentrations in low yielding regions can also enhance the P uptake since they appear to have a synergistic relation (Epstein & Bloom, 2005; Hussain et al., 2007; Iqbal & Hidayat, 2016). Several works indicate that the critical limit for potassium is in the range of 125-150 mg/kg, depending on the characteristics of the soil (Breker et al., 2019; Mallarino & Higashi, 2009; Van Biljon et al., 2008). These results indicate that a K soil deficiency might exist in this area (figure 26).

Phosphor (P) is the second most limiting nutrient in maize, directly affects growth and yield (Dhillon et al., 2017). P has an important role in the storage and energy transport (ATP) for endergonic processes, photosynthesis, synthesis of nucleic acids and organic compounds, redox reactions, carbohydrate metabolism and active uptake of nutrients (Marschner, 2011; Vance et al., 2003). Despite its importance, P is the least accessible macronutrient, which 80% of the P content is fixed in primary phosphate minerals and as hydroxides, oxides and silicate minerals (Mengel & Kirkby, 2012; Vance et al., 2003; White & Hammond, 2008). Phosphor was identified as a relevant factor in the low_{10} model (figure 25), with concentrations in the region of 120-150 mg/kg. The high_{10} model did not identify P as a relevant variable. This could be since concentrations above 150 mg/kg are enough to maintain the necessary crop nutrient requirements for the high yield regions. Low P concentrations on the soil are known to impact yield (Plénet et al., 2000) directly. The results indicate that the areas with 120-150 mg/kg appear to be deficient in available P. A detailed soil analysis is required to identify the available P in the identified areas correctly (figure 26).

Limestone is mainly use has an soil amendment, and is the most effect managerial practice for reducing high levels of soil pH by neutralizing excessive hydrogen ions in the soil solution (Bolan et al., 2003), reduces the availability of mineral elements (Fe^{2+},Al^{3+}) that are less soluble at higher pH values (Fageria & Baligar, 2008) and improves nutrient availability (Rengel, 2003).LS was identified as a secondary variable in the high$_{10}$ model (figure 23). The average model is not particularly informative since it indicates high ES at all concentrations due to interaction with other variables. The single variable model, however, indicates high ES when LS is above 0.47 g/kg. Even if the levels of LS in the field are low (Hazelton & Murphy, 2016), levels above 0.47 g/kg appear to contribute to the high yielding areas. Although its PI importance is low (0.5%), it is the variable with the 2o highest contribution to model gain. Its contribution is not entirely clear, but since LS is a source of several cations (Ca^{2+}, Mg^{2+}) that are important for crop production (Fageria & Nascente, 2014), this could be having a positive impact on yield.

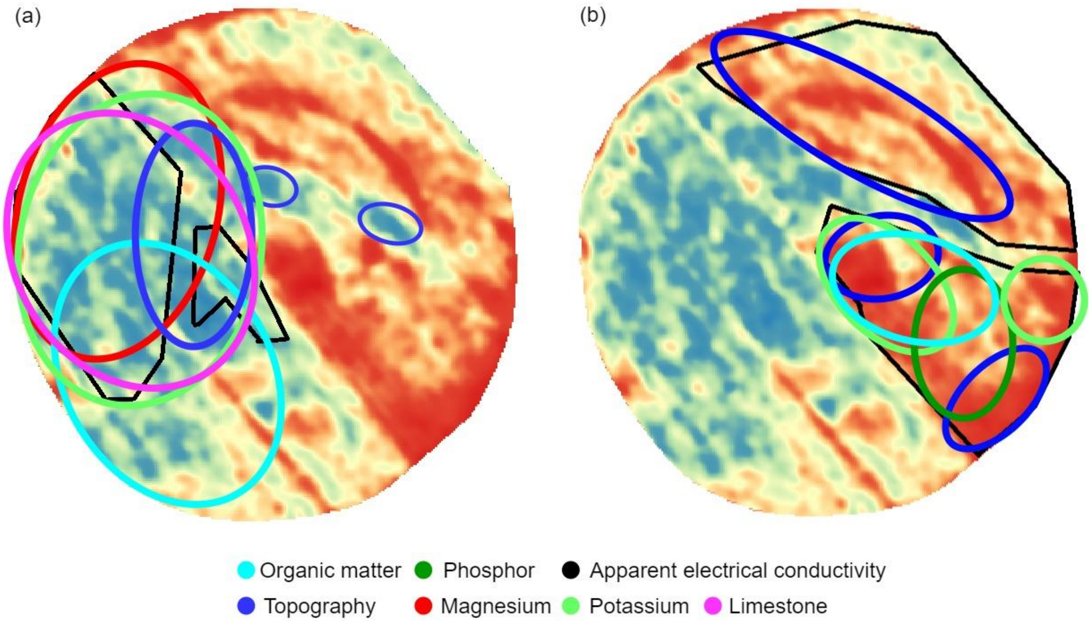


Figure 26: Location of the main factors in Cerca that drive: (a) High yields (b) Low yields

3.2. Lourenço Field

3.2.1. High Yield

The model that was selected to characterize the high yielding areas of Lourenço was the high$_{10}$ model. According to its PC the most contributing variables to model gain are DEM (27%), P (20.7%) and pH (17.4%) (table 12). PI shows that P, pH and DEM are the main contributing factors with 30.3%, 19.9% and 12.3% respectively. These 3 factors explain 62.5% of the high yielding areas.

Table 12: Environmental variables used for yield class model high$_{10}$ in Lourenço.

Variable	DEM	P	pH	OM	TPI	K	East	North	EC$_a$	DFL
Percent Contribution	27	20.7	17.4	7.9	6.4	5.4	4.8	4.6	4.3	1.5
Permutation Importance	12.3	30.3	19.9	9.3	7	6.9	3	6.7	2.5	2.2

DEM = Digital elevation model; P = Phosphorus; pH = Soil acidity; OM = Organic matter; TPI = Topographic position index; ; K = Potassium; East = Eastness; North = Northness; DFL = Distance to flow accumulation lines.

The Jack-knife (figure 27) indicates which secondary impacting factors appear to be relevant. From these, OM , TPI and OM show high model gain when used in isolation, and drop in gain when omitted. The response curves for the high$_{10}$ shows high correlation between several variables (figure 28).

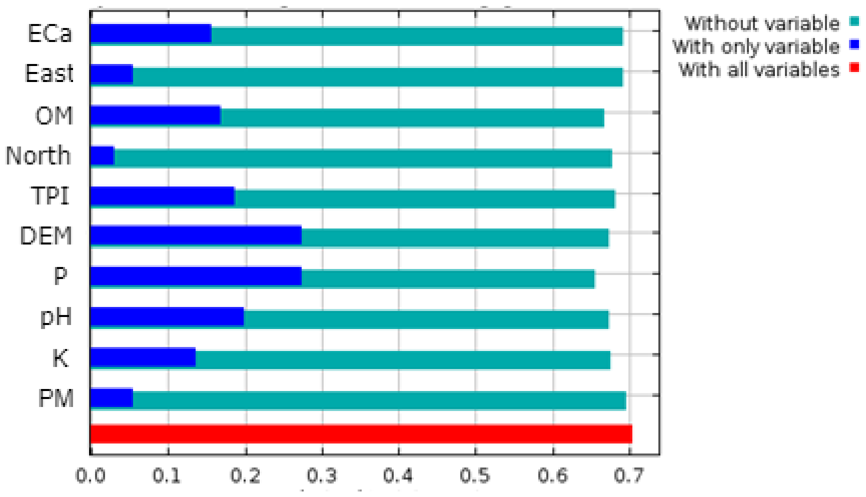


Figure 27: The Jack-knife test for evaluating the relative importance of environmental variables for yield class model high$_{10}$ in Lourenço.

The single and average model for P indicates that high yielding areas are characterized by a total P 95-135 mg/kg. And that below 80 mg/kg, the ES drops drastically.

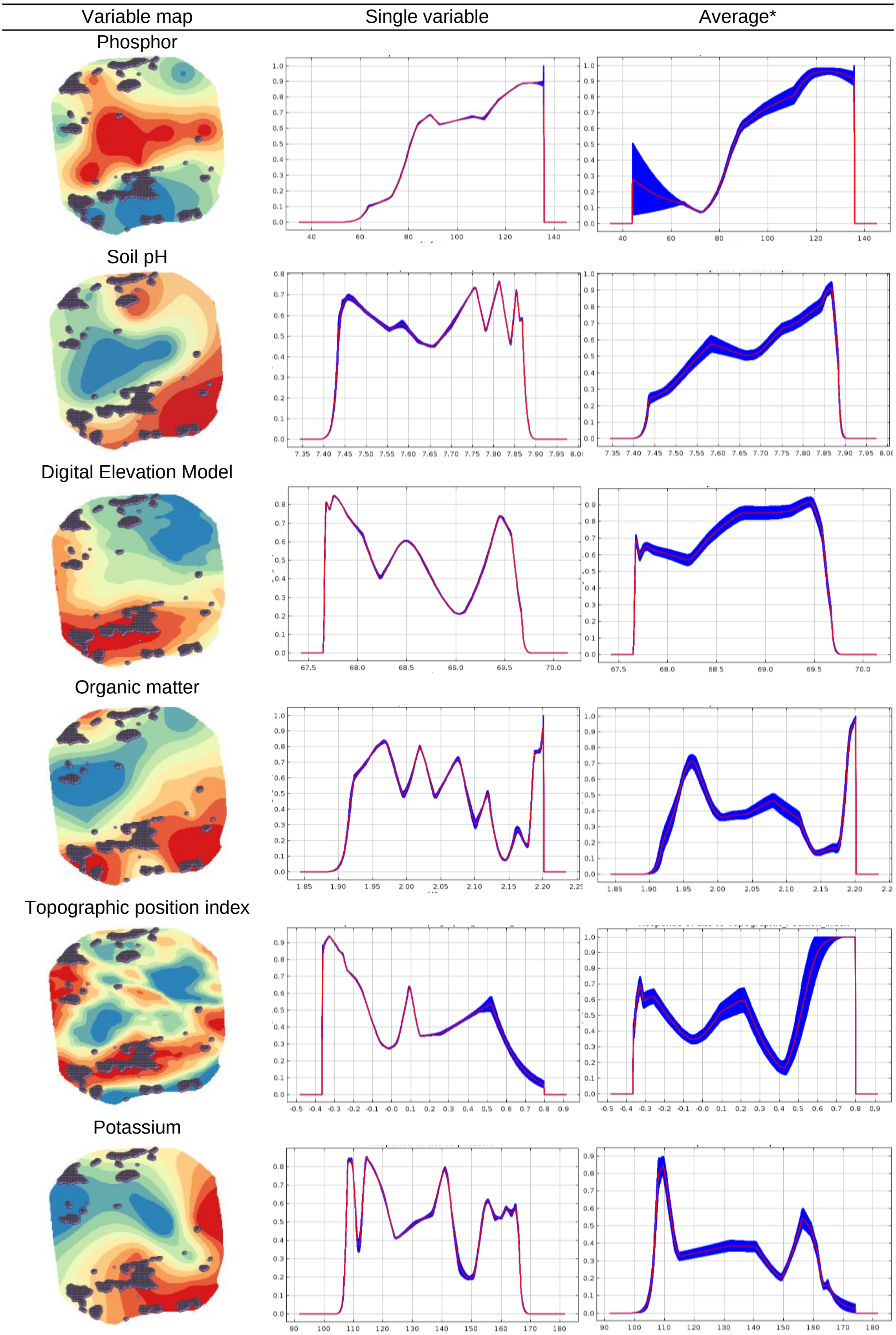


Figure 28: The environmental variables that control the high yield areas in Lourenço. The dots in the variable map column represent the location of the presence points

* Maxent model made with only the corresponding variable

** Maxent model where one variable is made to vary with all the other variables set to their average value

When the single variable model for pH is inspected, the ES indicates a peak in pH of 7.45 and fluctuation of the ES in the ranges of 7.75-7.85 and a ES drop in pH 7.9. The average model indicates that a pH of 7.75-7.85 characterizes the high yield class, and the ES had a drop when the pH reached 7.9. High yielding areas are not present in pH above 7.9.

The average model response curve for the DEM indicates that elevations between 68.5-69.5m have high ES, which are mainly in the northern part of the field. Inspecting the single variable model shows that elevations of 67.7-67.9m, located in the southern region, and elevations of 69.4-69.6m, located in the northern region, have a high ES (figure 28). The average model for the TPI indicates a high ES for regions above 0.55, which are areas with a higher relative elevation with its surroundings, are located in the southern part of the field (figure 28). The single variable model indicates high ES in the range of –(0.4-0.2), which is located in the south part of the field and in the northwest region. And there is also a weaker indication that relatively flat areas (0.1 TPI) in the north influence yield, presenting a small peak of 0.6 of ES.

The single variable and the average model for high yielding areas indicated that OM had a peak in ES in the region of 1.95-1.98%. Afterwards, both models indicates that with increasing percentage of OM, the ES decreases until it reaches the OM value of 2.17%, where a high ES is present.

The K single and the average model showed similar tendencies, with high ES in the 110 mg/kg region and a lower ES peak of 0.6 in the 155-165 mg/kg region. The single model also indicated a high ES in the concentration of 140 mg/kg. Still, the average model did not consider this result due to the existing interactions with the other variables.

3.2.2. Low Yield

The model that best explained the low yielding regions is the low_{20}. According to the PC, the variables that most contribute to model gain are the pH (32.5%), DEM (18.3%), Mg (17.6%), K (11.5%) and P (10.2%) (table 13). The PI shows that Mg is the main contributing factor, with 31.3%, followed by pH, P, DEM and K with 23.5%, 15%, 13.3% and 11.6 %, respectively. These factors cumulatively explain 94.7% of the low yielding regions.

Table 13: Environmental variables used for yield class model low$_{20}$ in Lourenço.

Variable	pH	DEM	Mg	K	P	North	TPI	DFL
Percent Contribution	32.5	18.3	17.6	11.5	10.2	5	2.5	2.3
Permutation Importance	23.5	13.3	31.3	11.6	15	1.7	2.5	1

pH = soil acidity; DEM = Digital elevation model; Mg=Magnesium; K = Potassium ; P = Phosphorus; North = Northness; TPI = Topographic position index; DFL = Distance to flow accumulation lines.

The Jack-knife (figure 29) indicates which secondary impacting factors appear to be relevant. No secondary variables were chosen, since the primary factors have a high explaining power. The response curves for the low$_{20}$ shows high correlation between several variables (figure 30).

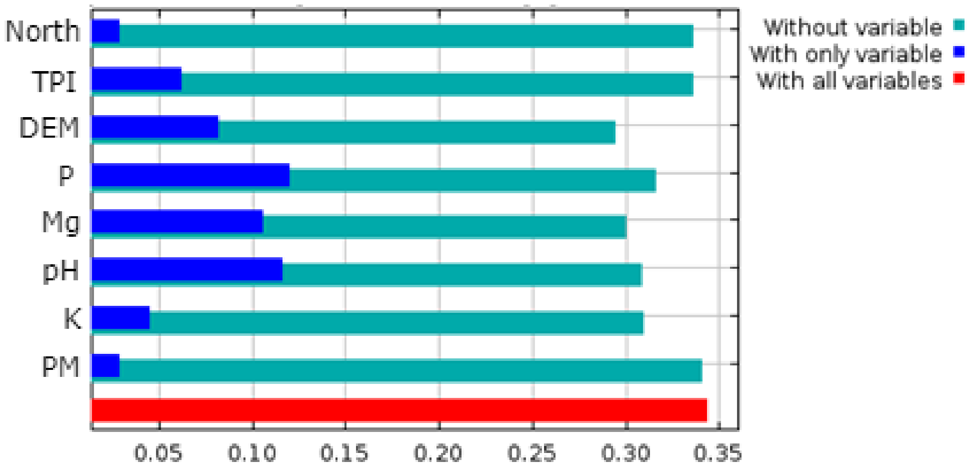


Figure 29: The Jack-knife test for evaluating the relative importance of environmental variables for yield class model low$_{20}$ in Lourenço.

Mg had ES values in the average model of 0.3 in 180-200 mg/kg concentrations for exchangeable magnesium. The single model indicates a tendency for having higher ES with increasing concentrations for the low yielding regions (figure 30).

Near the concentration of 7.9 pH, models made with only pH indicates a high ES. When the model accounting other variables, the pH range increases to 7.4, 7.5-7.7 and 7.9. The 7.9 pH characterizes the centre area of the field (figure 30).

The average P model presents a low and decreasing ES with increasing concentrations of P, having 3 peaks in the region of 43, 70 and 110 mg/kg (figure 30). The single variable model for the low yielding areas indicates that concentrations between 43-70 mg/kg show a high ES.

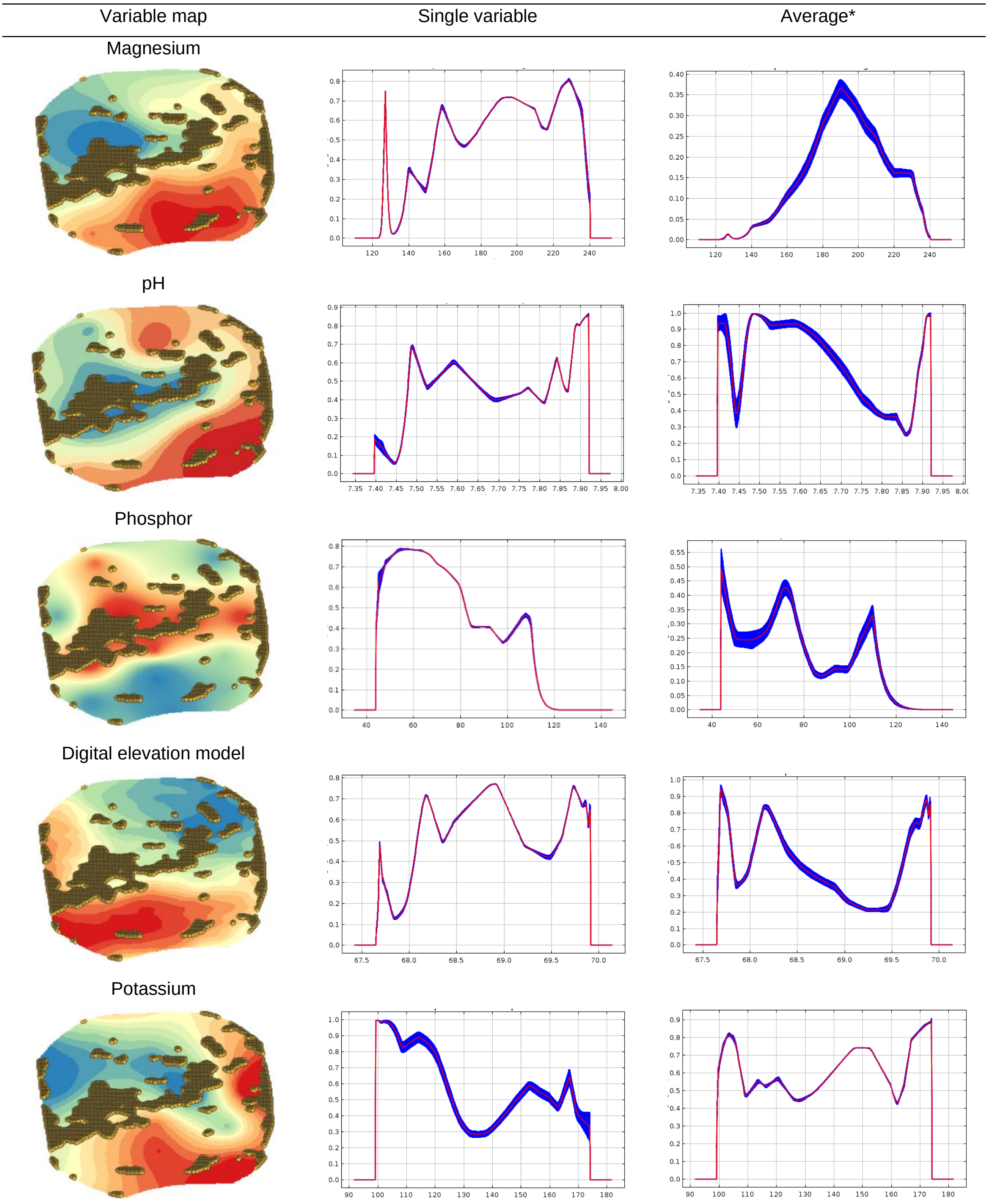


Figure 30: The environmental variables that control the low yield areas in Lourenço. The dots in the variable map column represent the location of the presence points

* Maxent model made with only the corresponding variable

** Maxent model where one variable is made to vary with all the other variables set to their average value

The average model of the DEM identified 3 peaks in ES, with altitudes of 67.8m, 68.1m and 69.7-69.9m. The single variable model also identified 3 peaks of ES above the 0.7 thresholds at 68.2m, 68.7-69m and 69.7m. The low yielding areas of Lourenço are mainly located in the centre of the field and the eastern and western border (figure 30).

For K, the single model indicates 3 ES peaks in the 105 and 150 and 170 mg/kg regions. When the average model is checked, it means high ES in the concentrations of 100-120 mg/kg of K for the low yielding areas (figure 30).

3.2.3. Discussion - Lourenço

In Lourenço, topography plays a significant role in the yield. High yielding areas are located in the north and the south of the field (figure 31). The northern area has an elevation of 68.5-69.5m with a high ES in the average model, and the southern areas have an elevation of 67.7-67.9m with a high ES in the single variable model (figure 28). Although the DEM indicates that elevations in these ranges play a part in influencing the yield, this result is not entirely useful because of the relative nature of the topography since several regions in this field have altitudes in the same range of elevation. But when the relative nature of topography is taken into account with the TPI, it becomes clear that these areas are either lower elevation regions or relatively flat areas. The single variable model indicates that these lower elevation areas are mainly located down south and northwest with a TPI of – (0.4 - 0.2). The flat areas are up north with a TPI of 0.1 in the single variable model (figure 28). But since a 200 m radius was used in the TPI to characterize Lourenço, it is possible that the surrounding areas are strongly influencing the 0.1 TPI value and that using a lower radius might be preferable to characterize the microtopography of a field.

Low yielding areas are mainly located in the centre of the field and in the western border of the field . The centre area of the map has two summit areas with elevations of 68.7- 69m and 69.7m (figure 30). The average model for the low$_{20}$ failed to identify the relative altitude of 68.7 - 69 m, probably due to being better explained by P and pH in that area (figure 31), but the single model did identify that this elevation had a high ES towards low yields (figure 30).

These results show here are in line with previous works about the impact that topography has on crop yield, where high elevation areas are characterized for having low yields and low elevation areas have high yields (Bakhsh, Jaynes, et al., 2000; Hansen et al., 2013; Mishra et al., 2008; Zhu et al., 2015).

Although most of the results align with the existing literature, two areas in Lourenço have been detected that presented conflicting results. A small area in the southern region has a TPI of 0.6 in the average model, a relatively high elevation area, which is characterized for having high yields (figure 28). Typically, high elevations are characterized as being less productive, and the result in this area is contradictory. Since a buffer was applied earlier on, the removal of this information may be inflating the TPI in this area coupled with the 200m radius used for the TPI calculation. These high yielding areas are also located in a region with a high P concentration, impacting the yield in this location. Also, elevations of 68.2m were identified as an impacting factor for the low crop yields in the western border of the field (figure 30). This area is characterized for being having a lower elevation regarding its surroundings. Since no other variables were identified in this region, a possible explanation for the existing low yield in this locale might be soil compaction since the edges of parcels are typically under stress from machine manoeuvres (Alakukku et al., 2003; Hamza & Anderson, 2005). Waterlog could also explain the results in this location, since it is a low elevation area. Waterlogged plants suffer from a reduction in the amount of oxygen available in the cells because oxygen's solubility and diffusion rate are extremely low in water (Voesenek & Bailey-Serres, 2015). However, the model did not identify either TWI or DFL as primary or secondary contributing variables for the low yields, so this explanation is unlikely. Although Maxent searches for the pattern that best describes the existing presence locations, these results indicate that a pattern does exist, but it might not be responsible for the presence locations.

OM for Lourenço was selected only for the high yielding areas, where the $high_{10}$ model indicates that between 1.95-1.98% contributes to the high yield (figure 28). But a ES of 1 above the 2.17% value is present in both the single and average models. This indicates that although this field has a high organic matter content, with an average of 2.07%, it appears that values above 2.17% explain the yield in that specific region (figure 31).

Soil pH is considered the main variable of soil chemistry due to its profound impact on several chemical reactions involving essential plant nutrients. The pH level influences the solubility, biological availability and mobility of the nutrients (McCauley et al., 2009; Penn & Camberato, 2019).

When inspecting the single variable model for high yields, there is a peak of ES in pH of 7.45, a fluctuation of the ES in the ranges of 7.75-7.85 and a ES drop in pH 7.9 (figure 28). The average model indicates that a pH of 7.75-7.85 characterizes the high yield class, and the ES had a drop when the pH reached 7.9. High yielding areas are not present in pH above 7.9. In the low yielding regions, models made with only pH (figure

30) indicates a high ES with a 7.9 pH. With the average model, the pH range increases to 7.4, 7.5-7.7 and 7.9 mainly due to the interaction with other variables. But pH of 7.9 characterized the centre area of the field and was signalled by the high and low model.

When investigating the P response curves, the high yield regions are characterized by a P of 95-135 mg/kg, which influences mainly the southern areas of Lourenço. The northern regions are also affected by P of 80mg/kg. After this threshold, a sharp drop in ES occurs (figure 28). Regions below this threshold barely have any high yielding areas. For the low yielding areas, the average model presents a low and decreasing ES (figure 30) with increasing concentrations of P, having 3 peaks in the region of 43, 70 and 110 mg/kg. Due to the interaction of pH and elevation, P's effect apparently has a low impact on crop yield distribution. But the single variable model for the low yielding areas indicates that concentrations between 43-70 mg/kg show a high ES. This result is biologically plausible since P in low concentrations directly impacts maize growth and yield (Mollier & Pellerin, 1999; Plénet et al., 2000). However, since this work used total P in the analysis, we do not know the exact amount of available P in the soil. So, we can only state that the low concentration of total P in that region impacts the yield**.**

Since P was also identified with pH as a relevant variable for both the $high_{10}$ and low_{20} yield class models, pH could be influencing total P availability in the low yielding areas. Soil pH impact on P is well known, with the highest P biological availability near pH 6.5. Above pH 6.5, phosphorus starts to precipitate with calcium, becoming less available for the plants (Penn & Camberato, 2019; Shen et al., 2011). Hinsinger (2001) showed that in pH values around 8, the solubility of Ca phosphate decreases. In this case, levels above 7.9 pH appear to impact P availability since the areas with pH above 7.9 have low amounts of total P (figure 30). A detailed soil analysis is required to confirm this relation.

Magnesium was identified as the most relevant variable for the low_{20} model. The ES of the average model is 0.3 in the concentrations of 180-200 mg/kg for exchangeable Mg. And the single model indicates a tendency for having higher ES with increasing concentrations for the low yielding regions (figure 30). These results go against the existing bibliography, stating that concentrations above 120 mg/kg are considered relatively sufficient (Wang et al., 2020) for maintaining high yields. Although this range of Mg values explains the low yielding areas in that location, it does not mean that its biologically relevant. These results could be explained because Maxent searches for the distribution that best explains the presence locations of this yield class. But there are more impacting factors that are biologically relevant that exist in the same area.

In the high$_{10}$ model of Lourenço, K was identified as a relevant variable. The northern areas have higher K levels than the southern areas (figure 28). The average model appears to reflect the K distribution in the field better. The northern areas are located in region with K values between 155 mg/kg with a ES of 0.6, below the defined threshold of 0.7, but the single model identified concentrations of 140 mg/kg in the single model. The southern areas are characterized by a K value of 110 mg/kg, and values in these ranges impact yield. Still, since the southern region is located in a depression, the water and nutrients that this area receives probably balances the low level of K that characterizes that area (Franzen et al., 2018; Kravchenko & Bullock, 2000). For the low$_{20}$ model of Lourenço, K was identified as a relevant variable. The single model identified 3 ES peaks in 105, 150 and 170 mg/kg concentrations. The K values of 150 mg/kg and 170 mg/kg are mainly located in the centre of the field, which is also characterized by low levels of P, high levels of pH and high altitude (figure 30). The single model gives a more reliable measure of the influence of K in this case. It indicates that concentrations of 100-120 mg/kg characterize the low yielding areas in the eastern border of Lourenço, and a few areas in the south (figure 31). Several works show that the critical limit for potassium is 125-150 mg/kg, depending on the characteristics of the soil (Breker et al., 2019; Mallarino & Higashi, 2009; Van Biljon et al., 2008). These results indicate that a K soil deficiency might exist in these areas.

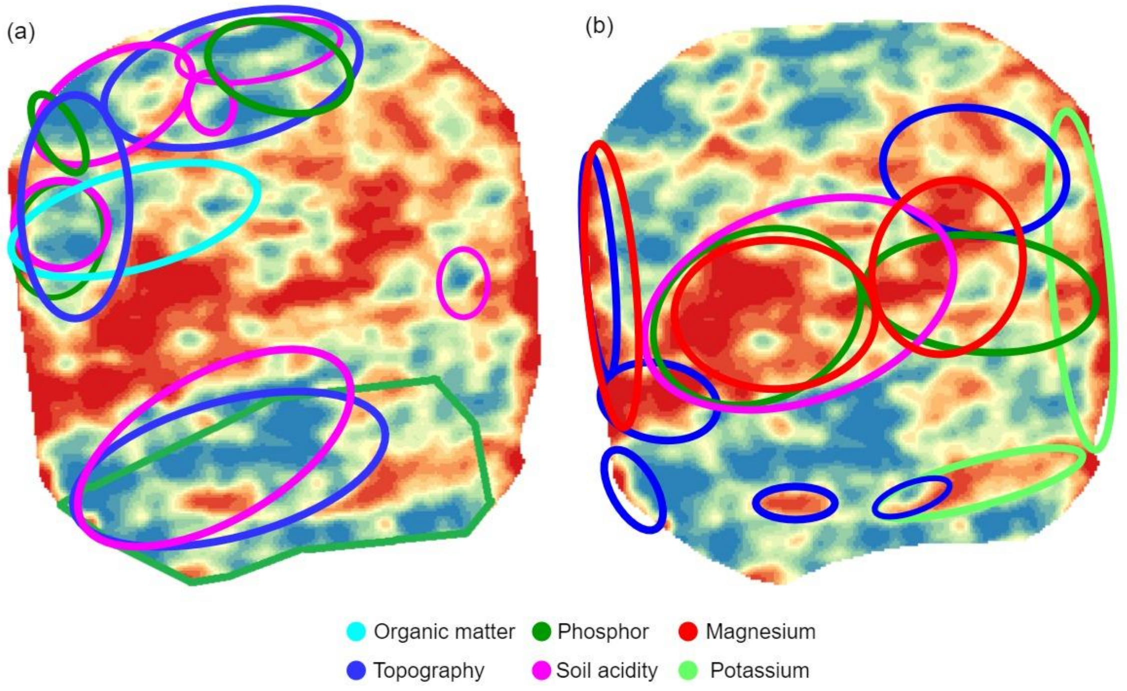


Figure 31: Location of the main factors in Lourenço that drive: (a) High yields (b) Low yields

3.3. Vinha Field

3.3.1. High Yield

The best model to characterize the high yielding regions of Vinha was the high$_{10}$. Following the PC of model gain, TPI (36%), pH (15.8%), Mg (13.5%), and OM (11.7%) have the highest contributions (table 14). The PI shows that Mg is the main contributing factor with 25%, followed by pH and TPI with 23.4% and 17.6% respectively. These three variables cumulatively explain 66% of the high yielding regions.

Table 14: Environmental variables used for yield class model high$_{10}$ in Vinha.

Variable	TPI	pH	Mg	OM	SL	P	EC$_a$	K	North	East
Percent Contribution	36	15.8	13.5	11.7	6.2	4.9	4.2	3.3	2.3	2
Permutation Importance	17.6	23.4	25	8.4	6.5	6.3	7	3.4	1.6	1

TPI = Topographic position index; pH= Soil acidity; Mg = Magnesium; OM = Organic matter; SL = Slope; P = Phosphorus; EC$_a$ = Apparent electrical conductivity; K = Potassium; North = Northness; East = Eastness.

The Jack-knife (figure 32) indicates which secondary impacting factors appear to be relevant. OM shows high model gain when used in isolation, and drop in gain when omitted. The response curves for the high$_{10}$ shows high a correlation between several variables (figure 33).

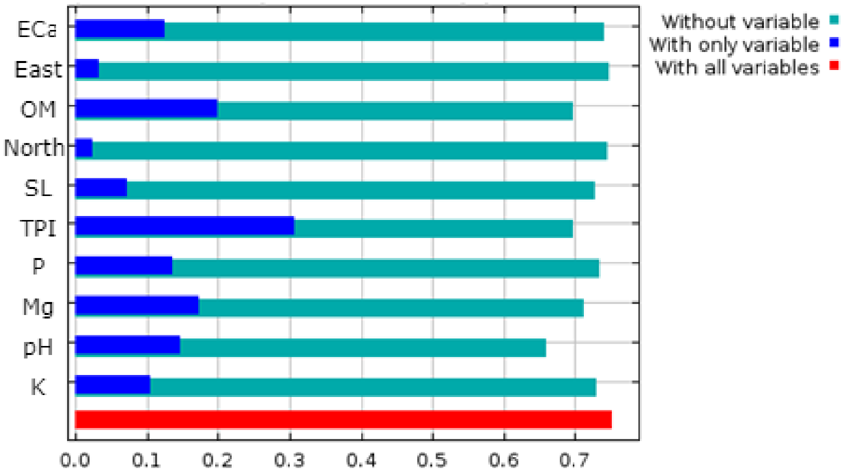


Figure 32: The Jack-knife test for evaluating the relative importance of environmental variables for yield class model high$_{10}$ in Vinha.

The TPI indicates that the region of -0.7 to 0 has a high ES for high yields (figure 33). Negative TPI indicates areas with lower relative altitude concerning its surroundings, mainly located surrounding a higher elevation area situated in the centre of the field.

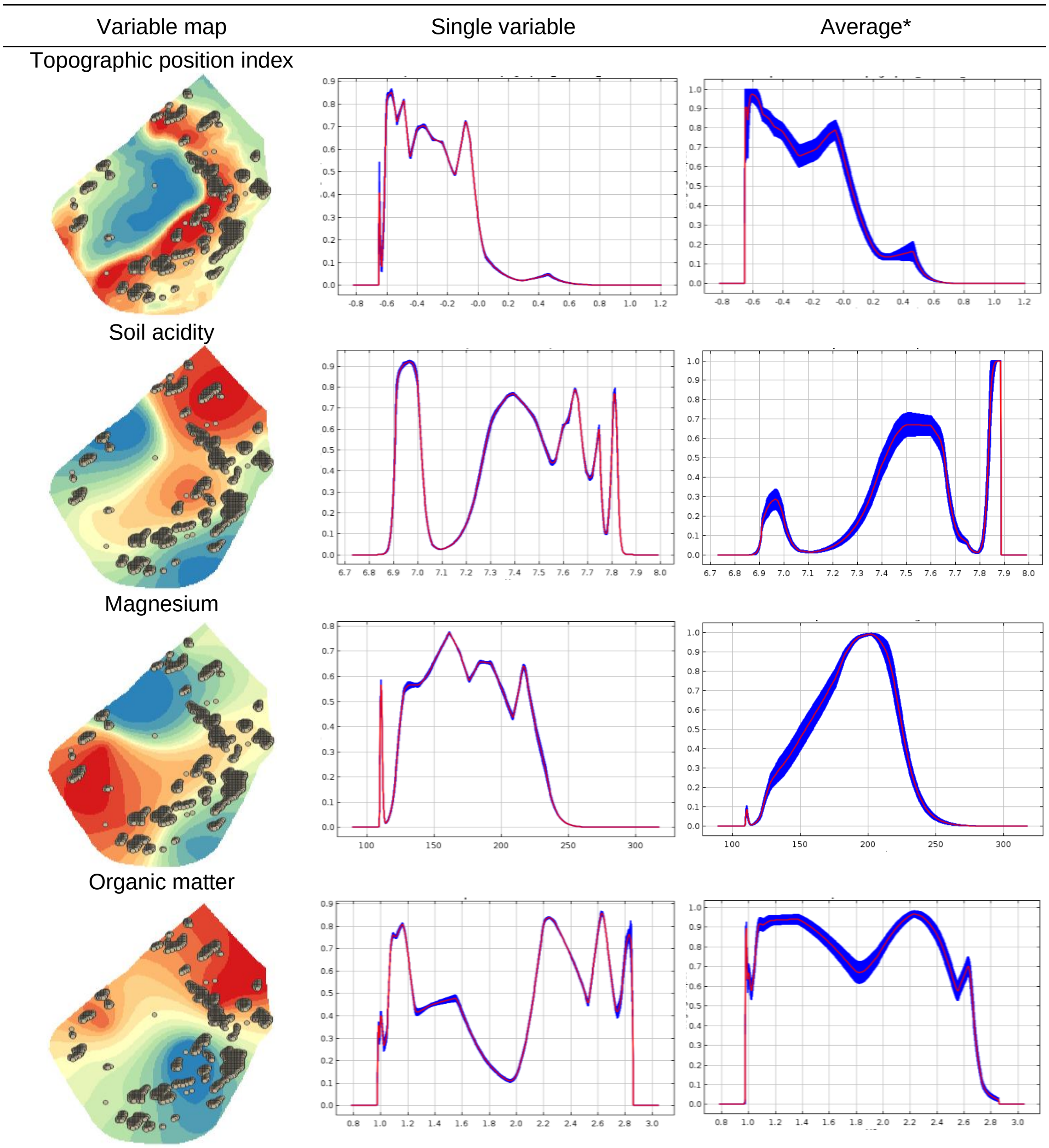


Figure 33: The environmental variables that control the high yield areas in Vinha. The dots in the variable map column represent the location of the presence points.

* Maxent model made with only the corresponding variable

** Maxent model where one variable is made to vary with all the other variables set to their average value

The pH in the single variable model indicates a peak of ES between the concentrations of 6.9-7, and between 7.3-7.8 several fluctuations in ES exist. The average model identified the same concentrations, but only the 7.8 pH range had a ES index above the 0.7 thresholds (figure 33).

The single and average model for Mg showed similar concentrations of ES in the range of 170-220 mg/kg (figure 33).

For the OM, the single variable model identified a high ES in the range of 1.1%, and 2.2-2.8% with fluctuation in the ES. And the average model had a high ES in concentrations of 1-2.4%. But below 1.1% in the single and average model had a high drop in ES.

3.3.2. Low Yield

The best model used to characterize the low yielding regions is the low$_{10}$. According to the PC to model gain, the DEM is the main contributing variable with 54.5%. The PI indicates that DEM is the primary factor with 37%, followed by OM with 12.5% (table 15).

Table 15: Environmental variables used for yield class model low$_{10}$ in Vinha.

Variable	DEM	SL	EC$_a$	DFL	Mg	East	Ph	K	North	PRC	TWI	OM	P
Percent Contribution	54.5	6.4	5	4.9	4.8	4.1	3.9	3.5	3.3	3.1	3.1	1.9	1.4
Permutation Importance	37	7.4	2.5	4.7	5.1	7.2	3.9	5.6	4.3	1.9	2.1	12.5	5.9

DEM = Digital elevation model; SL = Slope; EC$_a$ = Apparent electrical conductivity; DFL = Distance to flow accumulation lines; Mg = Magnesium; East = Eastness; Ph = Soil acidity; K = Potassium; North = Northness; PRC = Profile curvature; TWI = Topographic wetness index; OM = Organic matter; P = Phosphorus.

The Jack-knife (figure 34) indicates which secondary impacting factors appear to be relevant. PRC and TWI have the highest model gain when used in isolation, and DFL and SL present a high drop in model gain when omitted. The response curves for the low$_{10}$ shows high correlation between several variables (figure 35).

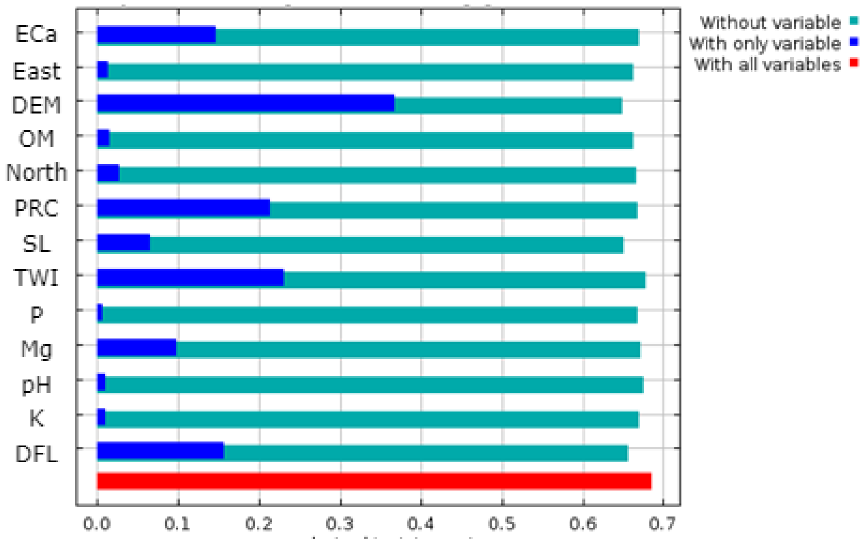


Figure 34: The Jack-knife test for evaluating the relative importance of environmental variables for yield class model low$_{10}$ in Vinha.

The DEM indicates that the centre area of the field is characterized for elevations above 70m, which have a high ES for low yields (figure 35).

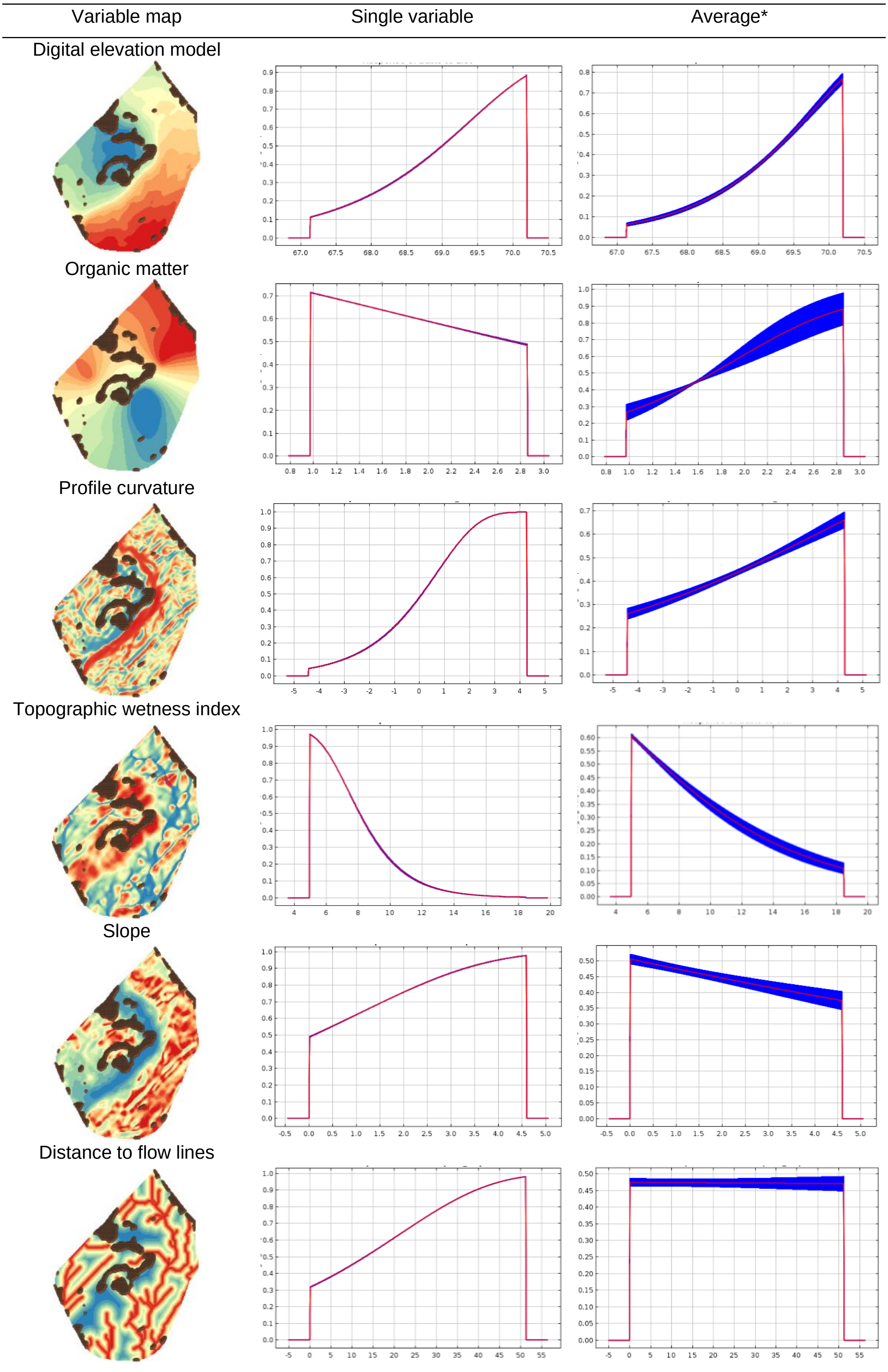


Figure 35: The environmental variables that control the low yield areas in Vinha. The dots in the variable map column represent the location of the presence points.
* Maxent model made with only the corresponding variable
** Maxent model where one variable is made to vary with all the other variables set to their average value

For OM, the single variable model indicates a high ES in 1% of OM concentration, with a decrease in ES with increasing OM values (figure 35). The average model shows different behaviour, that with increasing values of OM, the ES also improves, with values between 2.1 - 2.9% having high ES.

Although correlation is present in PRC, the response curves show similar behaviour. The single and average model indicates that with increasing values of PRC, there is an increase in ES. Only the single variable model indicates high values of ES, which are above 1 °/ m.

For the TWI, the single and average model indicates an increase in ES with decreasing values of TWI. Only the single variable model indicates high values of ES, which are above 5.

The slope single and average models show that the response curves have opposite behaviours due to correlation among variables. The average model indicates a maximum of ES of 0.55 in regions with 0° degrees and a decrease in ES with increasing slope values, while the single variable model indicates high ES with values above 1.5° degrees.

The average model for DFL indicates a steady ES value of 0.47 with increasing distances while the single variable model shows a high ES above 25m.

3.3.3. Discussion - Vinha

Topography was also a primary factor in affecting the high and low yielding areas in Vinha (figure 36). High yields are characterized for having a TPI of -0.7 to 0, which indicates lower relative elevation concerning its surroundings (figure 33). The low yield areas are located in a clear summit, with an elevation above 70m that dominates the centre area of the field (figure 35). High yields are typically characterized for being in lower elevations, and lower yields are typically in summit regions or backslope areas, and the results for Vinha indicate the same (Kumhálová et al., 2011; Muñoz et al., 2014; Zhu et al., 2015).

However, an issue was detected. Since the TPI and the DEM variables were correlated, the $high_{10}$ and low_{10} class yield models chose different variables to characterize the topography during the variable selection process. The $high_{10}$ used the TPI and identified the north region of the field as a high elevation area (figure 35). Since the low_{10} model chose the DEM to characterize the topography, it failed to identify that same area as a high elevation zone regarding its surroundings (figure 33). This is relevant because, without the proper information regarding the relative topography, the

Maxent analysis for the low yields failed to characterize that specific region since it did not have the right information.

Slope and profile curvature are low contributing variables for the low$_{10}$ model, so their impact on the yield class distribution is small but not underrated. The profile curvature average model does not show any values above the designated threshold of 0.7, but the single model indicates a high ES above 1 °/ m. The slope average model has no values above the 0.7 thresholds and with increasing slope values, there is a continuous decrease in ES. But the single model indicates high ES above values of above 1.5º (or 2.6%), which indicates that regions above 2.6% inclination are characterized for being low yielding locales. Iqbal et al. (2005) showed that in good growing conditions, regions with slopes under 2% had high yields, and steeper slopes have more erosion and lower water infiltration rate, which leads to lower productivities (Jiang & Thelen, 2004). Kaspar et al. (2003) and Jaynes et al. (2003) also show that curvature, along with other topographical information, were significant attributes in predicting corn yields for dry years. Meaning that convex areas are associated to low yielding areas, with our results indicating similar findings.

Analysing the OM, the high yields of Vinha in the average model are characterized for having an OM content between 1.1% and 2.8% overall. The single model shows a dip in ES in the region between 1.2 – 2.1%, with the average model indicating a similar but smaller dip in ES in the same range of values. This range of values characterizes the centre of the field, which is also a high elevation area that was shown to impact the yields. The high ES demonstrated in the average model for 1.1% of OM is explained due to being in a low elevation area (figure 33), which was also shown to impact the yields positively. There is an indication that the high crop yields are impacted below 1.1% of OM (figure 35), but the average model doesn't follow the same trend. The low$_{10}$ yield model had contradicting response curves. The single variable model indicates a high ES in 1% of OM. Increasing concentrations of OM shows a decrease in the ES (figure 35). Although the average model shows the opposite relation, this is due to the DEM interaction, which is the highest contributing variable to model gain. The area in the centre of the field where high ES is found, is located in a high elevation area which is skewing the average model results. So, taking into account the results from the single variable model, the areas with OM below 1%, mainly located in the northern region of the field, appears to have an impact on the yield in that location (figure 35). Since concentrations below 2% have been associated with lower yields (Lal, 2020; Oldfield et al., 2019), an increase of OM in the identified area is suggested.

Mg was identified as a variable with a high PI in the high$_{10}$ yield model but not in the low$_{10}$ yield model. The single variable and the average model showed similar concentrations of ES in the range of 170-220 mg/kg (figure 33). Above 220 mg/kg, the sharp drop in ES is provoked by the high elevation areas where high concentrations of Mg are located. High elevations were previously identified as a main contributing factor for the low yield. Wang et al. (2020) reported that values above 120 mg/kg for Mg are enough to maintain a high level of yield, and this result agrees with the author.

The low$_{10}$ model revealed that water availability plays a secondary role in the low yielding areas (figure 36). There is a significant model gain then TWI and DFL are used in isolation, and DFL also presents a significant drop in model gain when this variable is omitted. The average models for the TWI and DFL present lower ES due to the existing interactions from other higher contributing variables to the model gain, such as the DEM. Because they are low contributing variables, their impact on the overall ES is reduced but should not be overlooked because of the high importance of water in the crop yield. DFL has an ES value of 0.45-0.55 in all the distances (figure 35), while TWI presents a maximum value of 0.6 ES in the range of 5-7. When the single variable models are analysed for TWI and DFL, the low$_{10}$ indicates that TWI in the range of 5-7 have a high ES, and DFL in the 25-50m distance have a high ES (figure 35). These results are similar to those of Maestrini and Basso (2018) and Kumhálová et al. (2014), where a low TWI characterizes low yielding areas due to being drier than the rest of the field. Following the DFL, Da Silva and Silva (2008a) reported an increase in the average yield until 17.5 m of distance to flow lines. After this mark, there is a continuous decrease in average yield when longer distances occur. Our results indicate similar results, with the low yielding areas mainly situated in the ranges of 25-50m, indicating lower water availability.

The pH was identified as a primary variable for the high$_{10}$ yield classes models, but was not identified as a contributing variable for the low yield areas. In the single model, the ES indicates a peak of pH between the values of 6.9-7 and 7.3-7.8 with fluctuations in the ES (figure 33). The average model identified the same range of values, but only the 7.8 pH range had a ES index above the 0.7 threshold. The results from the average model do not appear to have a biological basis since maize has a strong preference for neutral pH (Fernández & Hoeft, 2009; Islam et al., 1980). However, the single model indicates the highest ES in the range of 6.9 pH (figure 33) indicating that lower pH values have a positive impact on yield in that specific region.

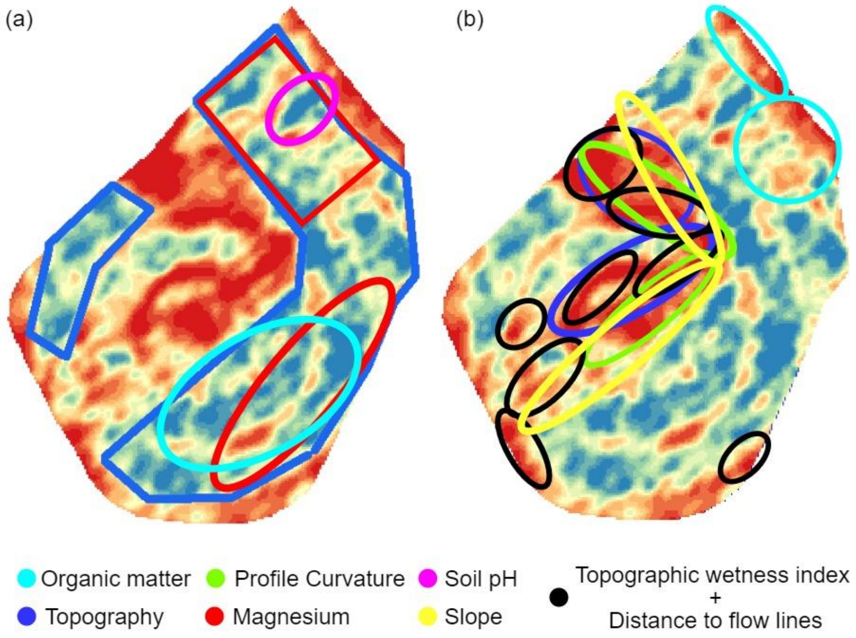


Figure 36: Location of the main factors in Vinha that drive: (a) High yields (b) Low yields

3.4. General discussion

For each field, 1530 models were built to characterize the different yield impacting factors. The best selected models correctly modelled the relationship between different variables, and multiple yield impacting factors were examined and determined.

Cerca was the only field where EC_a is identified as the main driving force, influencing both the high and low yielding areas. Lourenço showed a particular relation between pH and total P in the centre area of the field, where the area with pH near 8 is characterized for having low total P, which could be due to the decrease of the solubility of calcium phosphate due to high pH. The water-related variables, TWI and DFL, were seldom selected except for Vinha. The low yield model for Vinha indicates that water is a secondary factor also coincides with an area of high elevation. Although Quinta da Cholda has an efficient water management system, low TWI combined with high DFL in high elevation areas is an indication that this area is has less water availability, which can drive the yields down.

The topographical attributes were consistently selected as the main contributing variables in the three fields. This highlights the impact that topography has on high and low yielding areas, as previously mentioned by several authors. Organic matter was also selected for all fields, but its effects were less impactful and more focused. Areas whose values were above 1.5% OM had a positive impact on yields and areas below 1% OM had a negative impact on yields.

4. Using Maxent approach to fill the gap

Following the success of this approach in identifying the main yield impact factors, using Cerca field as an example, we propose a set of specific recommendations to help reduce the existing yield gap. The results are compiled in table 16 and figure 26 for Cerca. The maximum yield gap in this field is 44% between the high and low yields. Following on the variable importance, EC_a is the most relevant factor for the low yields (23.5 %). We suggest (if possible) increasing the EC_a between 10-14 mS/m. But since this might be texture related, this recommendation might not be realistic or feasible. The next variable to be corrected should be total P (22.1 %). The model indicates a lack of P in the low yielding areas, and a correction should be made. But since Total P was used, it is necessary to realize a soil analysis to determine the available P. After determining the available P, the identified areas by the model should be corrected. K is the third highest contributing variable to low yields (20.3 %), where a correction of 25 – 75 mg / kg is recommended.

Table 16: The main yield gap factors and the proposed solution to reduce the existing yield gap in the Cerca.

		Highest yield		Lowest yield		
Variables	Units	Values	importance	Values	importance	Yield gap
Yield	%	116	-	72	-	44
Yield	Ton./ha	20.1	-	12.6	-	7.5
ECa	mS/m	17-26	43.6 %	7-12	23.5%	+10-14
Topography	-	x	18.2%	x	12.3%	x
Mg	mg/kg	100-130	18.1%	80-90	6.5%	+20-40
OM	%	1.1-1.5	14.6%	0.65-0.8	4.5%	+0.5-0.7
LS	g/kg	0.5	0.6%	0	6.1%	+0.5
K*	mg/kg	>200	1.7%	130-175	20.3%	+25-70
Total P*	mg/kg	>350	-	120-150	22.1%	+200-230

EC_a = Apparent electrical conductivity; Topography = Digital elevation model + Topographical position index; Mg= Magnesium; OM = Organic matter; LS = Limestone; K = Potassium; P = Phosphorus.
Bold indicates the main variables to be corrected.
X indicates structural variables, not viable to be corrected.
*Possible interaction between K and P.

5. Conclusion and perspectives

A maximum entropy approach combined with PA technologies was used to determine the main factors responsible for the existing yield gaps in three different fields. Agriculture is a complex system where the yield pattern is driven by an intricate interaction between several environmental components. Following on the hypothesis that an existing yield pattern of a species (maize) can be interpreted as an ecological niche, we analysed these complex interactions using the Maxent algorithm based on a dataset consisting of yield maps of several years, topographical information and fertility maps.

Since Maxent can model complex, existing non-linear relationships, the results obtained here showed that complex relations in yield can be modelled using this approach. The main crop impacting factors at the farm level were determined, where they are located, and a detailed agronomic recommendation can be prescribed for narrowing the actual yield gap. Although Maxent managed to identify several yields impacting factors, caution is required when interpreting the results because the identified patterns by the maximum entropy approach sometimes are not agronomically relevant.

Because this study managed to answer the proposed research questions by modelling a highly complex environment, using such an innovative approach holds the potential to support smart PA solutions.

Perspectives

During the development of this work, a few issues were identified that in the future could be improved. Variable selection should contemplate removing low contributing variables to model gain through the jack-knife combined with the permutation importance. The jack-knife allows to inspect the amount of information available, and small contributing variables might contain information others do not. Using both metrics will ensure that the model possesses the variables with the most information.

When Maxent identifies a variable responsible for a specific pattern, but the pattern reveals to be a high importance variable but agronomically irrelevant, the variable should be removed, and the model rerun from the beginning of the analysis.

The radius used in the topographical position index (200m) sometimes failed to characterize the local topography, so adding a layer with a smaller radius (25 - 50 m) might improve the model performance.

We suggest an existing tool for the Maxent algorithm to visually check the variables that negatively influences model prediction the most. The automatic compilation of the limiting factors for the high and low yields can drive more insight in the interpretation and the analysis of the model output, instead of manually building the yield characterizing factors map by hand, as was done in this work. We suggest the mapping tool implemented in the package "*rmaxent*" (Baumgartner et al., 2017).

A calculation tool can be developed to calculate the profit margin from correcting the yield liming factors, to determine which factors are worth correcting first through a cost/benefit analysis and the level of yield gap that can be closed by correcting a limiting variable.

6. Bibliographic References

Adamchuk, V. I., Hummel, J. W., Morgan, M., & Upadhyaya, S. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture, 44*(1), 71-91.

Adamchuk, V. I., Rossel, R. V., Sudduth, K. A., & Lammers, P. S. (2011). Sensor fusion for precision agriculture. *Sensor Fusion-Foundation and Applications. InTech, Rijeka, Croatia*, 27-40.

Adnan, M. (2020). Role of potassium in maize production: A review. *Op Acc J Bio Sci Res, 3*(5), 1-4.

Alakukku, L., Weisskopf, P., Chamen, W., Tijink, F., Van Der Linden, J., Pires, S., . . . Spoor, G. (2003). Prevention strategies for field traffic-induced subsoil compaction: a review: Part 1. Machine/soil interactions. *Soil and Tillage Research, 73*(1-2), 145-160.

Almeida, C., Mendonça, J., Jesus, M., & Gomes, A. (2000). Sistemas aquíferos de Portugal continental. *Centro de Geologia da Fac. Ciências Univ. Lisboa, Instituto da Água, 3*.

Amtmann, A., Troufflard, S., & Armengaud, P. (2008). The effect of potassium nutrition on pest and disease resistance in plants. *Physiologia plantarum, 133*(4), 682-691.

Anderson, R. P., & Gonzalez Jr, I. (2011). Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. *Ecological Modelling, 222*(15), 2796-2811.

Anderson, R. P., Lew, D., & Peterson, A. T. (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling, 162*(3), 211-232.

Araujo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography, 33*(10), 1677-1688.

Arslan, S., & Colvin, T. S. (2002). Grain yield mapping: Yield sensing, yield reconstruction, and errors. *Precision agriculture, 3*(2), 135-154.

Bakhsh, A., Colvin, T. S., Jaynes, D. B., Kanwar, R. S., & Tim, U. S. (2000). Using soil attributes and GIS for interpretation of spatial variability in yield. *Transactions of the ASAE, 43*(4), 819.

Bakhsh, A., Jaynes, D. B., Colvin, T. S., & Kanwar, R. S. (2000). Spatio-temporal analysis of yield variability for a corn-soybean field in Iowa. *Transactions of the ASAE, 43*(1), 31.

Balafoutis, A., Beck, B., Fountas, S., Vangeyte, J., Wal, T. V. d., Soto, I., . . . Eory, V. (2017). Precision agriculture technologies positively contributing to GHG emissions mitigation, farm productivity and economics. *Sustainability, 9*(8), 1339.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., . . . Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling, 222*(11), 1810-1819.

Basso, B., Fiorentino, C., Cammarano, D., Cafiero, G., & Dardanelli, J. (2012). Analysis of rainfall distribution on spatial and temporal patterns of wheat yield in Mediterranean environment. *European journal of agronomy, 41*, 52-65.

Beza, E., Silva, J. V., Kooistra, L., & Reidsma, P. (2017). Review of yield gap explaining factors and opportunities for alternative data collection approaches. *European journal of agronomy, 82*, 206-222.

Bishop, T., & McBratney, A. (2002). Creating field extent digital elevation models for precision agriculture. *Precision agriculture, 3*(1), 37-46.

Blackmore, S. (1999). Remedial correction of yield map data. *Precision agriculture, 1*(1), 53-66.

Blackmore, S. (2000). The interpretation of trends from multiple yield maps. *Computers and Electronics in Agriculture, 26*(1), 37-51.

Blackmore, S., Godwin, R. J., & Fountas, S. (2003). The analysis of spatial and temporal trends in yield map data over six years. *Biosystems engineering, 84*(4), 455-466.

Bolan, N. S., Adriano, D. C., & Curtin, D. (2003). Soil acidification and liming interactions with nutrient and heavy metal transformation and bioavailability. *Advances in Agronomy, 78*(21), 5-272.

Bongiovanni, R., & Lowenberg-DeBoer, J. (2004). Precision agriculture and sustainability. *Precision agriculture, 5*(4), 359-387.

Braga, R., & Pinto, P. A. (2011). Agricultura de precisão: adopção & principais obstáculos. *AGROTEC, 1(1)*, 84–88.

Bramley, R., & Hamilton, R. (2004). Understanding variability in winegrape production systems: 1. Within vineyard variation in yield over several vintages. *Australian Journal of Grape and Wine Research, 10*(1), 32-45.

Breker, J., DeSutter, T., Rakkar, M., Chatterjee, A., Sharma, L., & Franzen, D. (2019). Potassium requirements for corn in North Dakota: Influence of clay mineralogy. *Soil Science Society of America Journal, 83*(2), 429-436.

Brevik, E. C., Fenton, T. E., & Lazari, A. (2006). Soil electrical conductivity as a function of soil water content and implications for soil mapping. *Precision agriculture, 7*(6), 393-404.

Cahn, M., Hummel, J., & Brouer, B. (1994). Spatial analysis of soil fertility for site-specific crop management. *Soil Science Society of America Journal, 58*(4), 1240-1248.

Cakmak, I., & Kirkby, E. A. (2008). Role of magnesium in carbon partitioning and alleviating photooxidative damage. *Physiologia plantarum, 133*(4), 692-704.

Cakmak, I., & Yazici, A. M. (2010). Magnesium: a forgotten element in crop production. *Better crops, 94*(2), 23-25.

Cambardella, C., & Karlen, D. (1999). Spatial analysis of soil fertility parameters. *Precision agriculture, 1*(1), 5-14.

Cambouris, A. N., Zebarth, B. J., Ziadi, N., & Perron, I. (2014). Precision agriculture in potato production. *Potato Research, 57*(3-4), 249-262.

Cardoso, J. d. C. (1965). *Os solos de Portugal: sua classificação, caracterização e génese; 1-A sul do rio Tejo*. Lisbon: General-Directorate for Agricultural Services.

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.

Chang, J., Clay, D. E., Carlson, C. G., Clay, S. A., Malo, D. D., Berg, R., . . . Wiebold, W. (2003). Different techniques to identify management zones impact nitrogen and phosphorus sampling variability. *Agronomy journal, 95*(6), 1550-1559.

Clay, D. E., Kitchen, N. R., Byamukama, E., & Bruggeman, S. A. (2017). Calculations supporting management zones. *Practical mathematics for precision farming*, 123-135.

Cobos, M. E., Peterson, A. T., Barve, N., & Osorio-Olvera, L. (2019). kuenm: an R package for detailed development of ecological niche models using Maxent. *PeerJ, 7*, e6281.

Condon, A. G., Richards, R., Rebetzke, G., & Farquhar, G. (2002). Improving intrinsic water-use efficiency and crop yield. *Crop science, 42*(1), 122-131.

Corwin, D., & Lesch, S. (2003). Application of soil electrical conductivity to precision agriculture: theory, principles, and guidelines. *Agronomy journal, 95*(3), 455-471.

Corwin, D. L., & Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture, 46*(1-3), 11-43.

Cunningham, S. A., Attwood, S. J., Bawa, K. S., Benton, T. G., Broadhurst, L. M., Didham, R. K., . . . Tscharntke, T. (2013). To close the yield-gap while saving biodiversity will require multiple locally relevant strategies. *Agriculture, Ecosystems & Environment, 173*, 20-27.

Da Silva, J. M., & Silva, L. L. (2006). Relationship between distance to flow accumulation lines and spatial variability of irrigated maize grain yield and moisture content at harvest. *Biosystems engineering, 94(4)*, 525-533.

Da Silva, J. M., & Silva, L. L. (2008a). Evaluation of the relationship between maize yield spatial and temporal variability and different topographic attributes. *Biosystems engineering, 101*(2), 183-190.

Da Silva, J. M., & Silva, L. L. (2008b). The yield pattern considering the distance to flow accumulation lines. *European journal of agronomy, 28*(4), 551-558.

Dhillon, J., Torres, G., Driver, E., Figueiredo, B., & Raun, W. R. (2017). World phosphorus use efficiency in cereal crops. *Agronomy journal, 109*(4), 1670-1677.

Diker, K., Heermann, D., & Brodahl, M. (2004). Frequency analysis of yield for delineating yield response zones. *Precision agriculture, 5*(5), 435-444.

Dobermann, A., Ping, J., Adamchuk, V., Simbahan, G., & Ferguson, R. (2003). Classification of crop yield variability in irrigated production fields. *Agronomy journal, 95*(5), 1105-1120.

Doerge, T. (1999). Defining management zones for precision farming. *Crop Insights, 8*(21), 1-5.

Doran, J. W., & Zeiss, M. R. (2000). Soil health and sustainability: managing the biotic component of soil quality. *Applied soil ecology, 15*(1), 3-11.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., . . . Leitao, P. J. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27-46.

Dudik, M., Phillips, S. J., & Schapire, R. E. (2004). *Performance guarantees for regularized maximum entropy density estimation.* Paper presented at the International Conference on Computational Learning Theory, Berlin.

Edwards, J. L. (2004). Research and societal benefits of the Global Biodiversity Information Facility. *BioScience, 54*(6), 485-486.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., . . . Lehmann, A. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29*(2), 129-151.

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution, 1*(4), 330-342. doi:10.1111/j.2041-210X.2010.00036.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics, 40*(1), 677-697. doi:10.1146/annurev.ecolsys.110308.120159

Epstein, E., & Bloom, A. (2005). Mineral nutrition of plants: principles and perspectives. Sinauer Associates. *Inc. Sunderland, Mass*.

European Commission. (2019). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: The European Green Deal, COM (2019) 640 final, Brussels. In.

Evans, L., & Fischer, R. (1999). Yield potential: its definition, measurement, and significance. *Crop science, 39*(6), 1544-1551.

Fageria, N., & Baligar, V. (2008). Ameliorating soil acidity of tropical Oxisols by liming for sustainable crop production. *Advances in Agronomy, 99*, 345-399.

Fageria, N. K., & Nascente, A. S. (2014). Management of soil acidity of South American soils for sustainable crop production. *Advances in Agronomy, 128*, 221-275.

Farahani, H., & Buchleiter, G. (2004). Temporal stability of soil electrical conductivity in irrigated sandy fields in Colorado. *Transactions of the ASAE, 47*(1), 79.

Farhat, N., Elkhouni, A., Zorrig, W., Smaoui, A., Abdelly, C., & Rabhi, M. (2016). Effects of magnesium deficiency on photosynthesis and carbohydrate partitioning. *Acta physiologiae plantarum, 38*(6), 145.

Feng, X., Park, D. S., Liang, Y., Pandey, R., & Papeş, M. (2019). Collinearity in ecological niche modeling: Confusions and challenges. *Ecology and Evolution, 9*(18), 10365-10376.

Ferguson, R. B., & Hergert, G. W. (2009). EC00-154 Precision Agriculture: Soil Sampling for Precision Agriculture. *Historical Materials from University of Nebraska-Lincoln Extension, 154*, 708.

Ferguson, R. B., Luck, J. D., & Stevens, R. (2017). Developing prescriptive soil nutrient maps. *Practical mathematics for precision farming*, 149-166.

Fernández, F. G., & Hoeft, R. G. (2009). Managing soil pH and crop nutrients. *Illinois agronomy handbook, 24*, 91-112.

Figueira, T. (1997). *A cultura do milho nas regiões da Golegã e Chaves: Estudo dos sistemas de produção e análise económica e social da cultura.* Dissertação de mestrado, ISA-UTL, Lisboa,

Fischer, R. (2015). Definitions and determination of crop yield, yield gaps, and of rates of change. *Field Crops Research, 182*, 9-18.

Flowers, M., Weisz, R., & White, J. G. (2005). Yield-based management zones and grid sampling strategies: Describing soil test and nutrient variability. *Agronomy journal, 97*, 968–982.

Foley, J. A., Ramankutty, N., Brauman, K. A., Cassidy, E. S., Gerber, J. S., Johnston, M., . . . West, P. C. (2011). Solutions for a cultivated planet. *Nature, 478*(7369), 337-342.

Fraisse, C., Sudduth, K., & Kitchen, N. (2001). Delineation of site-specific management zones by unsupervised classification of topographic attributes and soil electrical conductivity. *Transactions of the ASAE, 44*(1), 155.

Franklin, J. (2010). *Mapping species distributions: spatial inference and prediction*. UK: Cambridge University Press.

Franzen, D., Casey, F., & Derby, N. (2008). Site Specific Farming 3: Yield Mapping and Use of Yield Map Data. *NDSU Extension*. Retrieved from https://www.sbreb.org/wp-content/uploads/2018/05/Yield-Mapping.pdf

Franzen, D., Shannon, D., Clay, D., & Kitchen, N. (2018). Soil variability and fertility management. *Precision agriculture basics*(precisionagbasics), 79-92.

Franzen, D. W., & Peck, T. R. (1995). Field soil sampling density for variable rate fertilization. *Journal of Production Agriculture, 8*(4), 568-574.

Fulton, J., Hawkins, E., Taylor, R., Franzen, A., Shannon, D., Clay, D., & Kitchen, N. (2018). Yield monitoring and mapping. *Precision Agriculture Basics; Shannon, DK, Clay, DE, Kitchen, NR, Eds*, 63-78.

Gąsiorek, P., Vončina, K., Zając, K., & Michalczyk, Ł. (2021). Phylogeography and morphological evolution of Pseudechiniscus (Heterotardigrada: Echiniscidae). *Scientific reports, 11*(1), 1-16.

Gebbers, R., & Adamchuk, V. I. (2010). Precision agriculture and food security. *Science, 327*(5967), 828-831.

Gerendás, J., & Führs, H. (2013). The significance of magnesium for crop quality. *Plant and Soil, 368*(1), 101-128.

Giller, K. E., Rowe, E. C., de Ridder, N., & van Keulen, H. (2006). Resource use dynamics and interactions in the tropics: Scaling up in space and time. *Agricultural Systems, 88*(1), 8-27.

Godwin, R., & Miller, P. (2003). A review of the technologies for mapping within-field variability. *Biosystems engineering, 84*(4), 393-407.

González-García, J., Swenson, R. L., & Gómez-Espinosa, A. (2020). Real-time kinematics applied at unmanned aerial vehicles positioning for orthophotography in precision agriculture. *Computers and Electronics in Agriculture, 177*, 105695.

Gotway, C. A., Ferguson, R. B., Hergert, G. W., & Peterson, T. A. (1996). Comparison of kriging and inverse-distance methods for mapping soil parameters. *Soil Science Society of America Journal, 60*(4), 1237-1247.

Grassini, P., Cassman, K. G., & van Ittersum, M. (2017). Exploring Maize Intensification with the Global Yield Gap Atlas. *Better crops with plant food, 101*(2), 7-9.

Grassini, P., van Bussel, L. G., Van Wart, J., Wolf, J., Claessens, L., Yang, H., . . . Cassman, K. G. (2015). How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. *Field Crops Research, 177*, 49-63.

Grinnell, J. (1917). The niche-relationships of the California Thrasher. *The Auk, 34*(4), 427-433.

Grisso, R. D., Alley, M. M., Holshouser, D. L., & Thomason, W. E. (2005). Precision farming tools. soil electrical conductivity. *Virginia Cooperative Extension*(442–508). Retrieved from https://vtechworks.lib.vt.edu/handle/10919/51377

Gruber, S., & Peckham, S. (2009). Land-surface parameters and objects in hydrology. *Developments in Soil Science, 33*, 171-194.

Guignard, M. S., Leitch, A. R., Acquisti, C., Eizaguirre, C., Elser, J. J., Hessen, D. O., . . . Soltis, P. S. (2017). Impacts of nitrogen and phosphorus: from genomes to natural ecosystems and agriculture. *Frontiers in Ecology and Evolution, 5*, 70.

Haghverdi, A., Leib, B. G., Washington-Allen, R. A., Ayers, P. D., & Buschermohle, M. J. (2015). Perspectives on delineating management zones for variable rate irrigation. *Computers and Electronics in Agriculture, 117*, 154-167.

Hampe, A. (2004). Bioclimate envelope models: what they detect and what they hide. *Global Ecology and Biogeography, 13*(5), 469-471.

Hamza, M., & Anderson, W. (2005). Soil compaction in cropping systems: A review of the nature, causes and possible solutions. *Soil and Tillage Research, 82*(2), 121-145.

Hansen, S., Clay, S. A., Clay, D. E., Carlson, C. G., Reicks, G., Jarachi, Y., & Horvath, D. (2013). Landscape features impact on soil available water, corn biomass, and gene expression during the late vegetative stage. *The Plant Genome, 6*(2).

Hardie, M. (2020). Review of novel and emerging proximal soil moisture sensors for use in agriculture. *Sensors, 20*(23), 6934.

Hazelton, P., & Murphy, B. (2016). *Interpreting soil test results: What do all the numbers mean?* : CSIRO publishing.

He, Y., Hou, L., Wang, H., Hu, K., & McConkey, B. (2014). A modelling approach to evaluate the long-term effect of soil texture on spring wheat productivity under a rain-fed condition. *Scientific reports, 4*(1), 1-10.

Heege, H. J. (2013). Sensing of natural soil properties. In *Precision in crop farming* (pp. 51-102): Springer.

Hijmans, R. J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., . . . Shortridge, A. (2015). Package 'raster'. *R package, 734*.

Hinsinger, P. (2001). Bioavailability of soil inorganic P in the rhizosphere as affected by root-induced chemical changes: a review. *Plant and Soil, 237*(2), 173-195.

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.

Hussain, N., Khan, A. Z., Akbar, H., Bangash, N. G., Hayat, Z., & Idrees, M. (2007). Response of maize varieties to phosphorus and potassium levels. *Sarhad Journal of Agriculture, 23*(4), 881.

Hutchinson, G. E. (1957). Cold spring harbor symposium on quantitative biology. *Concluding remarks, 22*, 415-427.

Iqbal, A., & Hidayat, Z. (2016). Potassium management for improving growth and grain yield of maize (Zea mays L.) under moisture stress condition. *Scientific reports, 6*(1), 1-12.

Iqbal, J., Thomasson, J. A., Jenkins, J. N., Owens, P. R., & Whisler, F. D. (2005). Spatial variability analysis of soil physical properties of alluvial soils. *Soil Science Society of America Journal, 69*(4), 1338-1350.

Islam, A., Edwards, D., & Asher, C. (1980). pH optima for crop growth. *Plant and Soil, 54*(3), 339-357.

IUSS Working Group, W. (2006). World reference base for soil resources. *World Soil Resources Report, 103*.

Jalota, S., Singh, S., Chahal, G., Ray, S., Panigraghy, S., & Singh, K. (2010). Soil texture, climate and management effects on plant growth, grain yield and water use by rainfed maize–wheat cropping system: Field and simulation study. *Agricultural Water Management, 97*(1), 83-90.

Jaynes, D., Kaspar, T., Colvin, T., & James, D. (2003). Cluster analysis of spatiotemporal corn yield patterns in an Iowa field. *Agronomy journal, 95*(3), 574-586.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review, 106*(4), 620.

Jenson, S. K., & Domingue, J. O. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric engineering and remote sensing, 54*(11), 1593-1600.

Jiang, P., & Thelen, K. (2004). Effect of soil and topographic properties on crop yield in a North-Central corn–soybean cropping system. *Agronomy journal, 96*(1), 252-258.

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography, 21*(4), 498-507.

Kaky, E., Nolan, V., Alatawi, A., & Gilbert, F. (2020). A comparison between Ensemble and MaxEnt species distribution modelling approaches for conservation: A case study with Egyptian medicinal plants. *Ecological Informatics, 60*, 101150.

Kaspar, T. C., Colvin, T. S., Jaynes, D. B., Karlen, D. L., James, D. E., Meek, D. W., . . . Butler, H. (2003). Relationship between six years of corn yields and terrain attributes. *Precision agriculture, 4*(1), 87-101.

Kaspar, T. C., Pulido, D., Fenton, T., Colvin, T., Karlen, D., Jaynes, D., & Meek, D. (2004). Relationship of corn and soybean yield to soil and terrain properties. *Agronomy journal, 96*(3), 700-709.

Kaur, T., Brar, B., & Dhillon, N. (2008). Soil organic matter dynamics as affected by long-term use of organic and inorganic fertilizers under maize–wheat cropping system. *Nutrient Cycling in Agroecosystems, 81*(1), 59-69.

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology letters, 12*(4), 334-350.

Khosla, R., Westfall, D., Reich, R., Mahal, J., & Gangloff, W. (2010). Spatial variation and site-specific management zones. In *Geostatistical applications for precision agriculture* (pp. 195-219): Springer.

Kitchen, N., Sudduth, K., Myers, D., Drummond, S., & Hong, S. (2005). Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. *Computers and Electronics in Agriculture, 46*(1-3), 285-308.

Koning, N., & van Ittersum, M. K. (2009). Will the world have enough to eat? *Current Opinion in Environmental Sustainability, 1*(1), 77-82.

Kopecký, M., Macek, M., & Wild, J. (2021). Topographic Wetness Index calculation guidelines based on measured soil moisture and plant species composition. *Science of the Total Environment, 757*, 143785.

Koppen, W. (1936). Das geographische system der klimat. *Handbuch der klimatologie*, 46.

Kravchenko, A., & Bullock, D. (2000). Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy journal, 92*(1), 75-83.

Kravchenko, A., & Bullock, D. (2002). Spatial variability of soybean quality data as a function of field topography: I. Spatial data analysis. *Crop science, 42*(3), 804-815.

Kravchenko, A., Bullock, D., & Boast, C. (2000). Joint multifractal analysis of crop yield and terrain slope. *Agronomy journal, 92*(6), 1279-1290.

Kravchenko, A., Robertson, G., Thelen, K., & Harwood, R. (2005). Management, topographical, and weather effects on spatial variability of crop grain yields. *Agronomy journal, 97*(2), 514-523.

Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Henten, E. J. (2012). Sensing soil properties in the laboratory, in situ, and on-line: a review. *Advances in Agronomy, 114*, 155-223.

Kumhálová, J., Kumhála, F., Kroulík, M., & Matějková, Š. (2011). The impact of topography on soil properties and yield and the effects of weather conditions. *Precision agriculture, 12*(6), 813-830.

Kumhálová, J., Zemek, F., Novák, P., Brovkina, O., & Mayerová, M. (2014). Use of Landsat images for yield evaluation within a small plot. *Plant, Soil and Environment, 60*(11), 501-506.

Laborte, A. G., de Bie, K. C., Smaling, E. M., Moya, P. F., Boling, A. A., & Van Ittersum, M. K. (2012). Rice yields and yield gaps in Southeast Asia: past trends and future outlook. *European journal of agronomy, 36*(1), 9-20.

Lal, R. (2016). Soil health and carbon management. *Food and Energy Security, 5*(4), 212-222.

Lal, R. (2020). Soil organic matter content and crop yield. *Journal of Soil and Water Conservation, 75*(2), 27A-32A.

Liang, W.-l., Carberry, P., Wang, G.-y., Lü, R.-h., Lü, H.-z., & Xia, A.-p. (2011). Quantifying the yield gap in wheat–maize cropping systems of the Hebei Plain, China. *Field Crops Research, 124*(2), 180-185.

Licker, R., Johnston, M., Foley, J. A., Barford, C., Kucharik, C. J., Monfreda, C., & Ramankutty, N. (2010). Mind the gap: how do climate and agricultural management explain the 'yield gap'of croplands around the world? *Global Ecology and Biogeography, 19*(6), 769-782.

Lidberg, W., Nilsson, M., Lundmark, T., & Ågren, A. M. (2017). Evaluating preprocessing methods of digital elevation models for hydrological modelling. *Hydrological processes, 31*(26), 4660-4668.

Lindsay, J. B. (2016). Efficient hybrid breaching-filling sink removal methods for flow path enforcement in digital elevation models. *Hydrological processes, 30*(6), 846-857.

Lindsay, J. B. (2018). WhiteboxTools user manual. *Geomorphometry and Hydrogeomatics Research Group, University of Guelph, Guelph, Canada, 20*.

Lindsay, J. B. (2020). *Pit-centric depression removal methods.* Paper presented at the GEOMORPHOMETRY 2020, Perugia, Italy.

Lobell, D. B., Cassman, K. G., & Field, C. B. (2009). Crop yield gaps: their importance, magnitudes, and causes. *Annual review of environment and resources, 34*, 179-204.

Lobo, J. M., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography, 33*(1), 103-114.

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17*(2), 145-151.

Lyle, G., Bryan, B. A., & Ostendorf, B. (2014). Post-processing methods to eliminate erroneous grain yield measurements: review and directions for future development. *Precision agriculture, 15*(4), 377-402.

Machado, S., Bynum, E., Archer, T., Bordovsky, J., Rosenow, D., Peterson, C., . . . Wilson, L. (2002). Spatial and temporal variability of sorghum grain yield: Influence of soil, water, pests, and diseases relationships. *Precision agriculture, 3*(4), 389-406.

Maestrini, B., & Basso, B. (2018). Drivers of within-field spatial and temporal variability of crop yield across the US Midwest. *Scientific reports, 8*(1), 1-9.

Maidment, D. R., & Morehouse, S. (2002). *Arc Hydro: GIS for water resources*: ESRI, Inc.

Majumder, B., Mandal, B., & Bandyopadhyay, P. (2008). Soil organic carbon pools and productivity in relation to nutrient management in a 20-year-old rice–berseem agroecosystem. *Biology and Fertility of Soils, 44*(3), 451-461.

Mallarino, A. P., & Higashi, S. (2009). Assessment of potassium supply for corn by analysis of plant parts. *Soil Science Society of America Journal, 73*(6), 2177-2183.

Marschner, H. (2011). *Marschner's mineral nutrition of higher plants* (5th ed.). Dordrecht, Netherlands: Academic press.

McCauley, A., Jones, C., & Jacobsen, J. (2009). Soil pH and organic matter. *Nutrient management module, 8*(2), 1-12.

Melo-Merino, S. M., Reyes-Bonilla, H., & Lira-Noriega, A. (2020). Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling, 415*, 108837.

Mengel, K., & Kirkby, E. A. (2012). *Principles of plant nutrition*: Springer Science & Business Media.

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography, 36*(10), 1058-1069. doi:10.1111/j.1600-0587.2013.07872.x

Michael Mertens, F., Pätzold, S., & Welp, G. (2008). Spatial heterogeneity of soil properties and its mapping with apparent electrical conductivity. *Journal of Plant Nutrition and Soil Science, 171*(2), 146-154.

Mieza, M. S., Cravero, W. R., Kovac, F. D., & Bargiano, P. G. (2016). Delineation of site-specific management units for operational applications using the topographic position index in La Pampa, Argentina. *Computers and Electronics in Agriculture, 127*, 158-167.

Mishra, U., Clay, D., Trooien, T., Dalsted, K., Malo, D., & Carlson, C. (2008). Assessing the value of using a remote sensing-based evapotranspiration map in site-specific management. *Journal of plant nutrition, 31*(7), 1188-1202.

Mollier, A., & Pellerin, S. (1999). Maize root system growth and development as influenced by phosphorus deficiency. *Journal of experimental botany, 50*(333), 487-497.

Moore, I. D., Grayson, R., & Ladson, A. (1991). Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrological processes, 5*(1), 3-30.

Morales, N. S., Fernandez, I. C., & Baca-Gonzalez, V. (2017). MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. *PeerJ, 5*, e3093. doi:10.7717/peerj.3093

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.

Mueller, T., Pierce, F., Schabenberger, O., & Warncke, D. (2001). Map quality for site-specific fertility management. *Soil Science Society of America Journal, 65*(5), 1547-1558.

Mulla, D., & Khosla, R. (2016). Historical evolution and recent advances in precision farming. *Soil-specific farming precision agriculture*, 1-35.

Muñoz, J. D., Steibel, J. P., Snapp, S., & Kravchenko, A. N. (2014). Cover crop effect on corn growth and yield as influenced by topography. *Agriculture, Ecosystems & Environment, 189*, 229-239.

Naimi, B., Hamm, N. A., Groen, T. A., Skidmore, A. K., & Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography, 37*(2), 191-203.

Nanni, M. R., Povh, F. P., Demattê, J. A. M., Oliveira, R. B. d., Chicati, M. L., & Cezar, E. (2011). Optimum size in grid soil sampling for variable rate application in site-specific management. *Scientia Agricola, 68*(3), 386-392.

Nawar, S., Corstanje, R., Halcro, G., Mulla, D., & Mouazen, A. M. (2017). Delineation of soil management zones for variable-rate fertilization: A review. *Advances in Agronomy, 143*, 175-245.

Neumann, K., Verburg, P. H., Stehfest, E., & Müller, C. (2010). The yield gap of global grain production: A spatial analysis. *Agricultural Systems, 103*(5), 316-326.

Olaya, V. (2004). A gentle introduction to SAGA GIS. *The SAGA User Group eV, Gottingen, Germany, 208*.

Oldfield, E. E., Bradford, M. A., & Wood, S. A. (2019). Global meta-analysis of the relationship between soil organic matter and crop yields. *Soil, 5*(1), 15-32.

Oldfield, E. E., Wood, S. A., & Bradford, M. A. (2018). Direct effects of soil organic matter on productivity mirror those observed with organic amendments. *Plant and Soil, 423*(1), 363-373.

Oliver, M. A. (2010). An overview of geostatistics and precision agriculture. *Geostatistical applications for precision agriculture*, 1-34.

Ozulu, I. M., & Gökgöz, T. (2018). Examining the stream threshold approaches used in hydrologic analysis. *ISPRS International Journal of Geo-Information, 7*(6), 201.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., . . . Dasgupta, P. (2014). *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*: Ipcc.

Pearson, R. G. (2007). Species' distribution modeling for conservation educators and practitioners. *Synthesis. American Museum of Natural History, 50*, 54-89.

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Townsend Peterson, A. (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography, 34*(1), 102-117.

Pedersen, S. M., & Lind, K. M. (2017). *Precision Agriculture: Technology and Economic Perspectives*. Switzerland: Springer.

Penn, C. J., & Camberato, J. J. (2019). A critical review on soil chemical processes that control how soil pH affects phosphorus availability to plants. *Agriculture, 9*(6), 120.

Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling, 213*(1), 63-72.

Peterson, A. T., Papeş, M., & Soberón, J. (2015). Mechanistic and correlative models of ecological niches. *European Journal of Ecology, 1*(2), 28-38.

Peterson, A. T., & Soberón, J. (2012). Species distribution modeling and ecological niche modeling: getting the concepts right. *Natureza & Conservação, 10*(2), 102-107.

Philip Robertson, G., Gross, K. L., Hamilton, S. K., Landis, D. A., Schmidt, T. M., Snapp, S. S., & Swinton, S. M. (2014). Farming for ecosystem services: An ecological approach to production agriculture. *BioScience, 64*(5), 404-415.

Phillips, S. J. (2005). A brief tutorial on Maxent. *AT&T Research, 190*(4), 231-259.

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: an open-source release of Maxent. *Ecography, 40*(7), 887-893. doi:10.1111/ecog.03049

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling, 190*(3-4), 231-259. doi:10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography, 31*(2), 161-175.

Phillips, S. J., Dudík, M., & Schapire, R. E. (2017). Maxent software for modeling species niches and distributions (Version 3.4.4). *Biodiversity Informatics*. Retrieved from Available from: http://biodiversityinformatics.amnh.org/open_source/maxent/

Ping, J., & Dobermann, A. (2005). Processing of yield map data. *Precision agriculture, 6*(2), 193-212.

Planchon, O., & Darboux, F. (2002). A fast, simple and versatile algorithm to fill the depressions of digital elevation models. *Catena, 46*(2-3), 159-176.

Plénet, D., Mollier, A., & Pellerin, S. (2000). Growth analysis of maize field crops under phosphorus deficiency. II. Radiation-use efficiency, biomass accumulation and yield components. *Plant and Soil, 224*(2), 259-272.

Porter, M. E., & Heppelmann, J. E. (2014). How smart, connected products are transforming competition. *Harvard business review, 92*(11), 64-88.

Pringle, M., McBratney, A., Whelan, B., & Taylor, J. (2003). A preliminary approach to assessing the opportunity for site-specific crop management in a field, using yield monitor data. *Agricultural Systems, 76*(1), 273-292.

Qin, C., Zhu, A. X., Pei, T., Li, B., Zhou, C., & Yang, L. (2007). An adaptive approach to selecting a flow-partition exponent for a multiple-flow-direction algorithm. *International journal of geographical information science, 21*(4), 443-458.

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography, 41*(4), 629-643.

Raghavan, R., Barker, S., Cobos, M. E., Barker, D., Teo, E., Foley, D., . . . Peterson, A. T. (2019). Potential spatial distribution of the newly introduced long-horned tick, Haemaphysalis longicornis in North America. *Scientific reports, 9*(1), 1-8.

Recanati, F., Maughan, C., Pedrotti, M., Dembska, K., & Antonelli, M. (2019). Assessing the role of CAP for more sustainable and healthier food systems in Europe: A literature review. *Science of the Total Environment, 653*, 908-919.

Rengel, Z. (2003). *Handbook of soil acidity* (Vol. 94): CRC Press.

Reyns, P., Missotten, B., Ramon, H., & De Baerdemaeker, J. (2002). A review of combine sensors for precision farming. *Precision agriculture, 3*(2), 169-182.

Rossel, R. A. V., & Bouma, J. (2016). Soil sensing: A new paradigm for agriculture. *Agricultural Systems, 148*, 71-74.

Sachs, J. D., Binagwaho, A., Birdsall, N., Broekmans, J., Chowdhury, M., Garau, P., . . . Navarro, Y. K. (2019). *Investing in Development A Practical Plan to Achieve the Millennium Development Goals: Overview*: Routledge.

Sadler, E. J., Bauer, P. J., Busscher, W. J., & Millen, J. A. (2000). Site-specific analysis of a droughted corn crop: II. Water use and stress. *Agronomy journal, 92*(3), 403-410.

Saupe, E., Barve, V., Myers, C., Soberón, J., Barve, N., Hensz, C., . . . Lira-Noriega, A. (2012). Variation in niche and distribution model performance: the need for a priori assessment of key causal factors. *Ecological Modelling, 237*, 11-22.

Schils, R., Olesen, J. E., Kersebaum, K.-C., Rijk, B., Oberforster, M., Kalyada, V., . . . Manolova, V. (2018). Cereal yield gaps across Europe. *European journal of agronomy, 101*, 109-120.

Searcy, C. A., & Shaffer, H. B. (2016). Do ecological niche models accurately identify climatic determinants of species ranges? *The American Naturalist, 187*(4), 423-435.

Senbayram, M., Gransee, A., Wahle, V., & Thiel, H. (2015). Role of magnesium fertilisers in agriculture: plant–soil continuum. *Crop and Pasture Science, 66*(12), 1219-1229.

Shannon, D. K., Clay, D. E., & Kitchen, N. R. (2020). *Precision agriculture basics* (Vol. 176): John Wiley & Sons.

Shen, J., Yuan, L., Zhang, J., Li, H., Bai, Z., Chen, X., . . . Zhang, F. (2011). Phosphorus dynamics: from soil to plant. *Plant physiology, 156*(3), 997-1005.

Sillero, N. (2011). What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. *Ecological Modelling, 222*(8), 1343-1346.

Sillero, N., Arenas-Castro, S., Enriquez-Urzelai, U., Vale, C. G., Sousa-Guedes, D., Martínez-Freiría, F., . . . Barbosa, A. M. (2021). Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. *Ecological Modelling, 456*, 109671.

Sillero, N., & Barbosa, A. M. (2021). Common mistakes in ecological niche models. *International journal of geographical information science, 35*(2), 213-226.

Simbahan, G., Dobermann, A., & Ping, J. (2004). Screening yield monitor data improves grain yield maps. *Agronomy journal, 96*(4), 1091-1102.

Sishodia, R. P., Ray, R. L., & Singh, S. K. (2020). Applications of remote sensing in precision agriculture: A review. *Remote Sensing, 12*(19), 3136.

Soberón, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics, 2*, 1-10.

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359*(1444), 689-698.

Srinivasan, A. (2006). *Handbook of precision agriculture: principles and applications*: CRC press.

Stafford, J. V. (2000). Implementing precision agriculture in the 21st century. *Journal of agricultural engineering research, 76*(3), 267-275.

Stojanovic, N., & Stojanovic, D. (2019). Parallelizing multiple flow accumulation algorithm using cuda and openacc. *ISPRS International Journal of Geo-Information, 8*(9), 386.

Sudduth, K., Kitchen, N., Wiebold, W., Batchelor, W., Bollero, G., Bullock, D., . . . Schuler, R. (2005). Relating apparent electrical conductivity to soil properties across the north-central USA. *Computers and Electronics in Agriculture, 46*(1-3), 263-283.

Sudduth, K. A., & Drummond, S. T. (2007). Yield editor: Software for removing errors from crop yield maps. *Agronomy journal, 99*(6), 1471-1482.

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS One, 8*(2), e55158.

Taechatanasat, P., & Armstrong, L. (2014). *Decision support system data for farmer decision making*.

Tarboton, D. G., Bras, R. L., & Rodriguez-Iturbe, I. (1991). On the extraction of channel networks from digital elevation data. *Hydrological processes, 5*(1), 81-100.

Taylor, J., McBratney, A., & Whelan, B. (2007). Establishing management classes for broadacre agricultural production. *Agronomy journal, 99*(5), 1366-1376.

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical review, 38*(1), 55-94.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

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.

Townsend Peterson, A., Papeş, M., & Eaton, M. (2007). Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography, 30*(4), 550-560.

Tribe, A. (1992). Automated recognition of valley lines and drainage networks from grid digital elevation models: a review and a new method. *Journal of Hydrology, 139*(1-4), 263-293.

Turcotte, R., Fortin, J.-P., Rousseau, A. N., Massicotte, S., & Villeneuve, J.-P. (2001). Determination of the drainage structure of a watershed using a digital elevation model and a digital river and lake network. *Journal of Hydrology, 240*(3-4), 225-242.

Van Biljon, J., Fouche, D., & Botha, A. (2008). Threshold values and sufficiency levels for potassium in maize producing sandy soils of South Africa. *South African Journal of Plant and Soil, 25*(2), 65-70.

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.

Vance, C. P., Uhde-Stone, C., & Allan, D. L. (2003). Phosphorus acquisition and use: critical adaptations by plants for securing a nonrenewable resource. *New phytologist, 157*(3), 423-447.

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography, 36*(12), 2290-2299.

Verity, G., & Anderson, D. (1990). Soil erosion effects on soil quality and yield. *Canadian Journal of Soil Science, 70*(3), 471-484.

Viscarra Rossel, R., McBratney, A., & Minasny, B. (2010). *Proximal soil sensing*: Springer Science & Business Media.

Voesenek, L. A., & Bailey-Serres, J. (2015). Flood adaptive traits and processes: an overview. *New phytologist, 206*(1), 57-73.

Wang, L., & Liu, H. (2006). An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *International journal of geographical information science, 20*(2), 193-213.

Wang, Y.-J., Qin, C.-Z., & Zhu, A.-X. (2019). Review on algorithms of dealing with depressions in grid DEM. *Annals of GIS, 25*(2), 83-97.

Wang, Z., Hassan, M. U., Nadeem, F., Wu, L., Zhang, F., & Li, X. (2020). Magnesium fertilization improves crop yield in most production systems: A meta-analysis. *Frontiers in plant science, 10*, 1727.

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological applications, 21*(2), 335-342.

Weiss, A. (2001). *Topographic position and landforms analysis.* Paper presented at the Poster presentation, ESRI user conference, San Diego, CA.

Whelan, B., & McBratney, A. (2000). The “null hypothesis” of precision agriculture management. *Precision agriculture, 2*(3), 265-279.

Whelan, B., & Taylor, J. (2013). *Precision agriculture for grain production systems*: Csiro publishing.

White, P. J., & Hammond, J. P. (2008). Phosphorus nutrition of terrestrial plants. In *The ecophysiology of plant-phosphorus interactions* (pp. 51-81): Springer.

Wibawa, W. D., Dludlu, D. L., Swenson, L. J., Hopkins, D. G., & Dahnke, W. C. (1993). Variable fertilizer application based on yield goal, soil fertility, and soil map unit. *Journal of Production Agriculture, 6*(2), 255-261.

Wilson, J. P., & Gallant, J. C. (2000a). Digital terrain analysis. *Terrain analysis: Principles and applications, 6*(12), 1-27.

Wilson, J. P., & Gallant, J. C. (2000b). *Terrain analysis: principles and applications*: John Wiley & Sons.

Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M.-J. (2017). Big data in smart farming–a review. *Agricultural Systems, 153*, 69-80.

Zarco-Tejada, P. J., Hubbard, N., & Loudjani, P. (2014). Precision agriculture: An opportunity for EU farmers—Potential support with the CAP 2014-2020. *Joint Research Centre (JRC) of the European Commission*.

Zevenbergen, L. W., & Thorne, C. R. (1987). Quantitative analysis of land surface topography. *Earth surface processes and landforms, 12*(1), 47-56.

Zhai, Z., Martínez, J. F., Beltran, V., & Martínez, N. L. (2020). Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture, 170*, 105256.

Zhang, N., Wang, M., & Wang, N. (2002). Precision agriculture—a worldwide overview. *Computers and Electronics in Agriculture, 36*(2-3), 113-132.

Zhang, Q. (2016). *Precision agriculture technology for crop farming* (1 ed.). Boca Raton: Taylor & Francis.

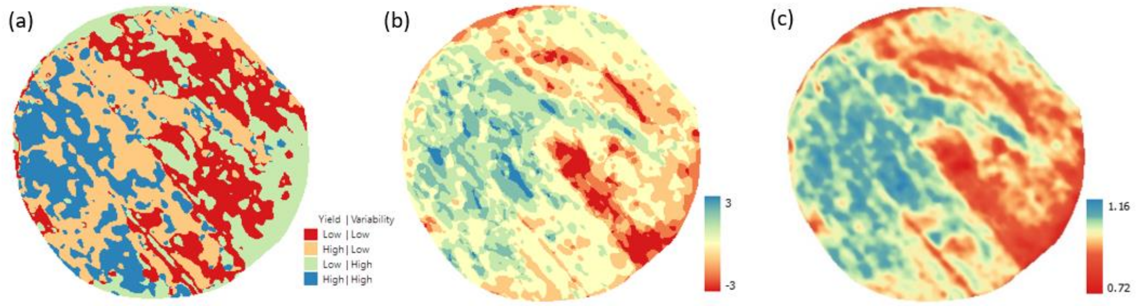
Zhu, Q., Schmidt, J. P., & Bryant, R. B. (2015). Maize (Zea mays L.) yield response to nitrogen as influenced by spatio-temporal variations of soil–water-topography dynamics. *Soil and Tillage Research, 146*, 174-183.

Zimmermann, N. E., Edwards Jr, T. C., Graham, C. H., Pearman, P. B., & Svenning, J. C. (2010). New trends in species distribution modelling. *Ecography, 33*(6), 985-989.

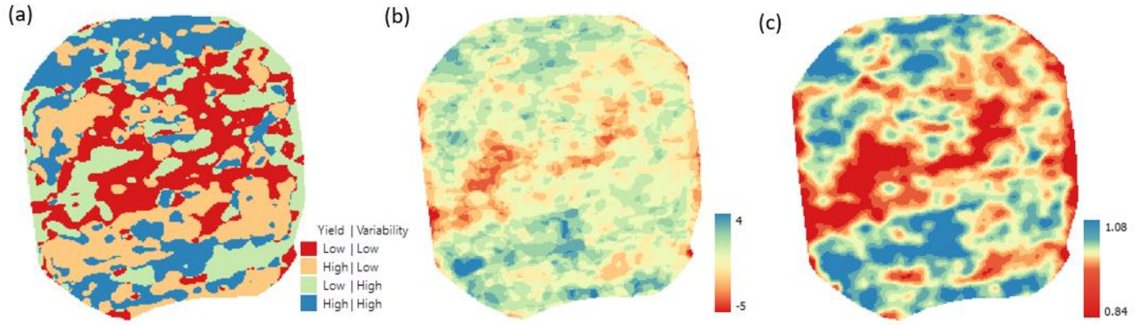
Zörb, C., Senbayram, M., & Peiter, E. (2014). Potassium in agriculture–status and perspectives. *Journal of plant physiology, 171*(9), 656-669.

Annex

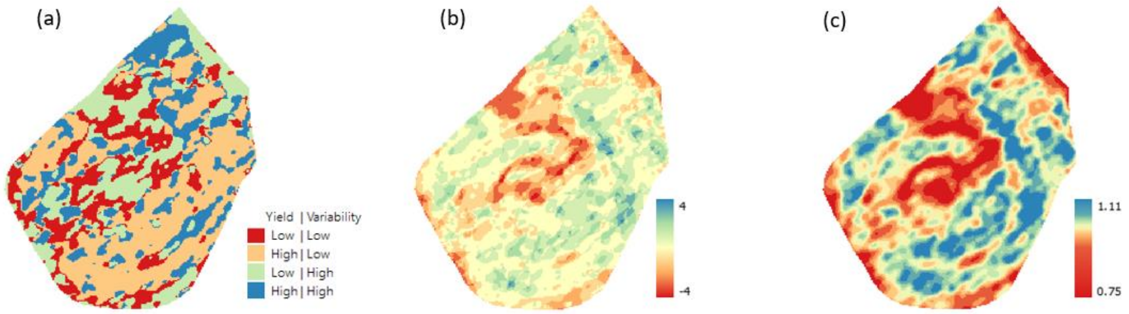
Annex I – Different yield maps produced



Multiyear Yield maps for Cerca: (a) Temporal Yield map (b) Yield frequency map. Number of years that are above or below the average yield within 1 standard deviation (c) Relative yield map



Multiyear Yield maps for Lourenço: (a) Temporal Yield map (b) Yield frequency map. Number of years that are above or below the average yield within 1 standard deviation (c)Relative yield map



Multiyear Yield maps for Vinha (a) Temporal Yield map (b) Yield frequency map. Number of years that are above or below the average yield within 1 standard deviation (c) Relative yield map

Annex II – Model selection and goodness of fit of Lourenço and Vinha

Model selection process of Lourenço for the high and low yield classes.

Percentile	TCM	SSM	MOr	MAIC	SSM + MOr	SSM + MAIC	SSM + Mor + Maic	Selected model
High yield models								
10	255	255	213	1	213	1	1	RM_0.1_F_lqph
15	255	255	134	1	134	1	1	RM_0.1_F_ph
20	255	255	93	2	39	2	1	RM_2_F_lh
Low yield models								
10	255	255	1	1	1	1	1	RM_5_F_lh
15	255	255	194	3	194	3	1	RM_0.1_F_qh
20	255	255	191	1	191	1	1	RM_0.1_F_qh

Statistics of the best selected models (goodness of fit) for Lourenço

Yield Class	Selected model	Mean AUC ratio	OR	AICc	AUC$_{Train}$	Stdev$_{train}$	AUC$_{Test}$	Stdev$_{test}$	AUC$_{Diff}$
High yield models									
10	RM_0.1_F_lqph	1.479	0.049	70192.42	0.8195	0.00065	0.814	0.00598	0.0055
15	RM_0.1_F_ph	1.394	0.049	108694	0.7588	0.00115	0.7539	0.00661	0.0049
20	RM_2_F_lh	1.269	0.049	148927.4	0.6958	0.00079	0.6943	0.01100	0.0015
Low yield models									
10	RM_5_F_lh	1.406	0.049	72022.74	0.7983	0.00075	0.7971	0.00663	0.0012
15	RM_0.1_F_qh	1.404	0.05	107147.8	0.7703	0.00067	0.767	0.00528	0.0033
20	RM_0.1_F_qh	1.358	0.045	146250.6	0.7255	0.00071	0.7213	0.00699	0.0042

Model selection process of Vinha for the high and low yield classes.

Percentile	TCM	SSM	MOr	MAIC	SSM + MOr	SSM + MAIC	SSM + Mor + Maic	Selected model
High yield models								
10	255	255	181	1	181	1	1	RM_0.1_F_lph
15	255	255	4	1	4	1	1	RM_10_F_lh
20	255	255	22	1	22	1	1	RM_0.1_F_lph
Low yield models								
10	255	255	20	1	20	1	1	RM_0.4_F_lp
15	255	255	21	1	21	1	1	RM_0.5_F_h
20	255	255	17	1	17	1	1	RM_0.2_F_qph

Statistics of the best selected models (goodness of fit) for Vinha

Yield Class	Selected model	Mean AUC ratio	OR	AICc	AUC$_{Train}$	Stdev$_{train}$	AUC$_{Test}$	Stdev$_{test}$	AUC$_{Diff}$
High yield models									
10	RM_0.1_F_lph	1.462	0.049	48940.75	0.8313	0.00139	0.8213	0.00042	0.01
15	RM_10_F_lh	1.237	0.05	77471.87	0.7093	0.000704	0.7072	0.00986	0.0021
20	RM_0.1_F_lph	1.361	0.049	100980.7	0.7453	0.00119	0.7374	0.00828	0.0079
Low yield models									
10	RM_0.4_F_lp	1.382	0.048	48589.19	0.8244	0.00120	0.8231	0.01289	0.0013
15	RM_0.5_F_h	1.464	0.05	73539.95	0.8093	0.00057	0.8057	0.00965	0.0036
20	RM_0.2_F_qph	1.489	0.049	98578.73	0.78	0.00094	0.7746	0.00506	0.0054