Geo-information Science and Remote Sensing

Thesis Report GIRS-2018-37

Analysis of soil and crop patterns to delineate potential management zones for precision agriculture in arable cropping

Tom Hardy

August 2018

WAGENINGEN
UNIVERSITY & RESEARCH

Analysis of soil and crop patterns to delineate potential management zones for precision agriculture in arable cropping

Tom Hardy

Registration number 841204-305-110

Supervisors:

Dr. ir. Lammert Kooistra
Dr. Harm Bartholomeus

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University & Research,
The Netherlands.

August 2018
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2018-37
Wageningen University & Research
Laboratory of Geo-Information Science and Remote Sensing

Acknowledgements

First of all, from the department Geo-Information Science and Remote Sensing at Wageningen University and Research, I want to thank my thesis supervisors Lammert Kooistra and Harm Bartholomeus, my study advisor Willy Ten Haaf, as well as other staff members from the GRS department and from Wageningen University in general. Their teaching methods, supervision, support, and advice were of great benefit during my thesis research and during the MGI programme in general.

I also would like to thank my dean Ruur Boersma for her help and advice with respect to my internship and thesis reports. Moreover, I want to thank Willem Dragt for critically reviewing and giving me feedback on those reports.

Besides this, I want to thank Jacob Van Den Borne, as well as the staff of Aurea Imaging for providing the data of the study field, which I was allowed to use in my thesis research.

Last but not least, a special sign of gratitude goes to my parents, friends, and my therapist Dona Dijkema, whose encouragement, patience, and support were of great value for me to get this far with my thesis, and to finally graduate for my MSc in Geo-Information Science at Wageningen University.

Abstract

Precision agriculture (PA) for arable cropping is an application of technologies and principles to manage spatial and temporal variability, related with all aspects of agricultural production. These aspects include sampling, mapping, analysis and the management of production areas, and have the purpose to increase crop yield while keeping constant or improving the quality of the environment. Remote and proximal sensing data are often used as input for methods to delineate potential management zones (MZs) to support PA practices. The objective of this research was to define and investigate a method to determine how soil variation related to crop yield to delineate potentially homogeneous management zones for precision agriculture. Remote and proximal sensing data from a potato parcel from a farm in Reusel in the south of the Netherlands were obtained for this purpose, which included spectral indices of bare soil (NDRG and SUMVIS), apparent electric conductivity (ECa), elevation, soil depth of the top horizon, soil organic matter (SOM), Total SOM, potato crop yield, and vegetation indices (NDVI and WDVI). Those data were used to define and investigate a stepwise method to delineate potentially homogeneous MZs for the potato field. The first step consisted of descriptive and visual assessments of data, including a correlation analysis. Moderate to strong correlations were found between NDRG, ECa, elevation, soil depth, and Total SOM. Also significant correlations were found between these variables and crop yield, although the correlations between soil depth and Total SOM on the one hand and crop yield on the other hand were lower. These eight soil and elevation variables were selected as input for a MULTISPATI-PCA algorithm. Four spatial principal components (sPCs) were derived explaining 94.08% of the total variance in the dataset, which were used as input for a k-means clustering algorithm. This algorithm led to four potentially homogeneous MZs, from which the classification results were smoothed by means of a non-linear spatial filter. For three selected areas in the field, validation of the delineated MZs was performed by fitting a number of statistical models on stratified random samples of crop yield per area, selecting the most optimal model, and testing two hypotheses based on that model to examine the differences in expected crop yield between the four MZs. For each of the three areas, significant differences ($p < 0.05$), or at least practical differences (9 ton/ha or more) in expected crop yield were found between all MZs, except between MZ2 and MZ3, indicating that these two MZs possibly could have been merged into one MZ. A method like this could be incorporated in a decision support system (DSS), in order to support management practices in precision agriculture for arable cropping.

Keywords: Precision agriculture, management zones, remote sensing, proximal sensing, MULTISPATI-PCA, k-means clustering, potato crop yield.

List of acronyms

AIC	Akaike's Information Criterion
ANOVA	Analysis Of Variance
DSS	Decision Support System
EC, ECa	Electric Conductivity, Apparent Electric Conductivity
EMI	Electromagnetic Induction
FCM, FKM	Fuzzy c-means Clustering, Fuzzy k-means Clustering
GeoTIFF	Geometric Tagged Image File Format
GNSS	Global Navigation Satellite System
IR, NIR	Infrared, Near Infrared
MLM	Mixed Linear Model
MULTISPATI	Multivariate Spatial Analysis based on Moran's index I
MZ	Management Zone
N	Nitrogen
NDRG	Normalized Difference Red minus Green (index)
NDVI	Normalized Difference Vegetation Index
PA, PF	Precision Agriculture, Precision Farming
PC, sPC	Principal Component, Spatial Principal Component
PCA	Principal Component Analysis
RD (New)	Rijksdriehoeksstelsel (New)
RGB	Red, Green, Blue (Colour Bands)
RPAS	Remotely Piloted Aerial System
SOM	Soil Organic Matter
SPDF	Spatial Points Data Frame
SSFM	Site-Specific Field Management
SUMVIS	Sum of visible (RGB) colour bands
UAV	Unmanned Aerial Vehicle
VRT	Variable Rate Technology
WDVI	Weighted Difference Vegetation Index
WGS	World Geodetic System
WSS	Within Sum of Squares

Table of contents

1. Introduction

This introduction starts with a background and problem definition about site-specific field management in relation to precision agriculture, followed by the objective and research questions of this thesis research and the outline of this report.

1.1 Background

During the last centuries, the demand for new and larger amounts of agricultural products has increased rapidly, because of a growing wealth and an expanding world population. Hence, efficient agricultural production has been developing extensively worldwide. For many years, this has been done in a traditional way. One of the persistent assumptions in traditional farming is that agricultural fields have homogeneous features, encouraging a farmer to apply whole-field management strategies (El Nahry, Ali, and Baroudy 2011). However, it is well known that soils contain heterogeneous spatial patterns, which could be relevant information for optimizing the yield of agricultural crops. Therefore, during the last couple of decades, traditional farming has increasingly been considered to be less efficient, and precision farming (PF) or precision agriculture (PA) has been applied instead (Johnson et al. 2003). Precision agriculture is the application of technologies and principles to manage spatial and temporal variability, related with all aspects of agricultural production (El Nahry et al. 2011; Pierce and Nowak 1999). Those technologies and principles include sampling, mapping, analysis and the management of production areas, and have the purpose to increase crop yields while keeping constant or even improving the quality of the environment (Weiss 1996).

An important strategy in precision farming is the application of site-specific field management (SSFM). This strategy aims to optimize farmer's input such as nutrient and fertilizer rates, based on site-specific crop requirements (Kooistra et al. 2011). Numerous variable rate technologies (VRTs) have been applied to identify and measure spatial variability within a field. An example of a VRT is the development of management zone maps for farmers to use for managing their fields in a more efficient way (Zhang et al. 2010). A management zone (MZ) is a sub-region in an agricultural field containing one or more similar quantitative and yield-limiting soil features, such as ground water level, mineral composition, soil organic matter (SOM), various local landscape attributes, and similar crop input needs (Haghverdi et al. 2015; Vrindts et al. 2005). These features could be relevant information to optimize farmer's input on an agricultural field, in order to increase crop yields (Fridgen et al. 2004).

The acquisition and management of soil and crop data, and the development of suitable decision criteria are challenging, yet important steps to realizing a VRT program (Kitchen et al. 1996). Remote and proximal sensing techniques are increasingly being used to capture data for the delineation of management zones, since they are appropriate for collecting information about soil properties and crop characteristics (Moran, Inoue, and Barnes 1997). Remote sensing is defined as the recording of data from a distant location without physical interaction, with help of platforms such as satellites or sensors and cameras on board of (un)manned aerial vehicles that measure reflected electro-magnetic radiation (Lillesand, Kiefer, and Chipman 2008). In proximal sensing, ground based sensors are used to measure for example spectral reflectance (SR) or electric conductivity (EC) of the soil, which have proven to relate closely to many soil properties that often determine a field's productivity (Lund, Christy, and Drummond 1999). Those properties involve among others texture, water holding capacity, porosity, salinity, and temperature (Grisso et al. 2009). More about remote and proximal sensing techniques will be discussed in section 2.2.

1.2 Problem statement

As stated in the previous section, soil and crop patterns are potential information to use as a basis for site-specific field management (SSFM), as a strategy in precision farming (PF). For example, yield maps have often been used as a measure of crop productivity to support delineation of management zones (MZs) (Whelan and McBratney 2003). To be able to make use of PF techniques effectively, a good understanding of soil and crop patterns and other physical and biological factors is vital. However, there is no such thing as one universal type of dataset that is applicable as standard information for PF practices at any place in the world (Kitchen et al. 2005). Reasons for this are that crop yield, but also other variables, are greatly influenced by local variations in soil and landscape, changing weather conditions, different approaches to field management, and hazards such as pests and diseases (Vitharana et al. 2008). Two papers suggested that crop yield data of at least five years should be used for the purpose of delineating stable management zones (Boydell and McBratney 2002; Stoorvogel, Kooistra, and Bouma 2015). Besides this, another paper recommended the use of historical remote sensing images in combination with real-time remote sensing data in high spectral and spatial resolutions for improved delineation of management zones (Mulla 2012).

Additionally, it is important to define and investigate methods to identify key soil and topographic variables (or yield-limiting factors) for specific regions, agricultural crops and systems, especially because required soil and crop data are increasingly becoming available for farmers to use in site-specific field management (SSFM). A number of studies have made attempts to identify these factors as a basis to delineate potential management zones for precision farming. For instance, one study investigated field in a loess area near Brussels, Belgium by collecting apparent soil electric conductivity (ECa) measurements in addition to soil samples, and performing a principal component analysis (PCA) on the derived soil variables (Vitharana et al. 2008). It turned out that ECa, elevation, and pH were the indicators representing the largest amount of variation responsible for the soil patterns. A few years later, a similar study was performed, but in a different study area in Belgium containing sandy soil (Van Meirvenne et al. 2012). In addition, a larger soil sample was drawn, and gamma ray measurements were included as well. Surprisingly, the same yield-limiting factors were identified, despite the large differences in soil and landscape development between the two areas in Belgium. Regardless of these outcomes, each agricultural area is different. Therefore, potentially yield-limiting factors should be measured and evaluated effectively for each farm to characterize spatial field variation, and to decide which of those factors to include in the creating of potential MZs (Nawar et al. 2017).

In addition to the wide range of available soil and crop variables, another important question is which sources and sensors to use for acquiring these kinds of data. A widely used and accurate approach to determining and analysing soil properties is intensive grid sampling, but this is economically not always favourable because of the high costs and the time consuming process (El Nahry et al. 2011). For that reason, other approaches for delineating management zones have been proposed, such as remote and proximal sensing techniques, which have been widely adopted in the fields of precision farming and crop monitoring, and for effectively supporting the crop production chain (Johnson et al. 2003; Thessler et al. 2011). However, it was argued that challenges existed to develop sensing technology, measurement services, and management tools that could be of added value for improving crop yields, and decreasing input costs and production risks. Propositions have been made to tackle those challenges, such as developing simple protocols for calibrating sensors, standardizing output formats for different types of sensor systems, writing more extensive product specifications, and making those specifications more widely available to end-users (Kooistra 2011).

Because of the large range of soil and landscape properties, using data from one single sensor is not always sufficient (Castrignanò et al. 2012), and therefore several multi-sensor based approaches have been proposed. For example, one investigated approach was the integrated use of EMI data, gamma-ray emission, and GPS measurements for improved soil and landscape characterization, by means of geostatistical methods and the estimation of relationships between the different sources of data (De Benedetto et al. 2013; Castrignanò et al. 2012). In addition, it has also been proposed to develop ways for integrating data from both remote and proximal sensing technology, for instance to overcome the lack of data availability from satellites caused by cloudy weather conditions (Kooistra 2011). This suggestion was implemented in a study that revealed good relations between vegetation indices derived from multiple sources of sensing data, and showed significant differences in crop development for a number of parcels throughout a growing season (Kooistra et al. 2012). Another proposition to overcome the lack of availability from satellite images was to use sensors and multispectral cameras on board of (un)manned aerial vehicles to acquire remotely sensed images. Because of the low costs, high spatial and temporal resolutions, and high flexibility in image acquisition, data recorded in that way could be a good alternative to satellite images (Zhang and Kovacs 2012).

When an appropriate dataset with different kinds of soil and crop data coming from multiple types of sensing platforms has been acquired, the question remains which descriptive and statistical techniques are most suitable for deriving stable management zones to support decision making for precision agriculture. The most widely applied approach for delineating management zones is cluster analysis, which is a collection of unsupervised learning algorithms to classify values of input variables into one or more given clusters, based on metrics such as Euclidian distance (James et al. 2013). For instance, one study compared the results of cluster analysis based on the correlation of soil and crop parameters and cluster analysis based on soil parameters only, and found promising results for the first type of clustering, which was considered valuable information for site-specific field management (SSFM) (Vrindts et al. 2005). Another research performed a correlation analysis to select soil variables from a dataset that was collected for experiments in Brazil, and normalized the variables as input for a clustering algorithm to delineate potential MZs, showing that normalization was required if variables were measured in different measurement scales (Schenatto et al. 2017). Principal component analysis (PCA) is another widely used approach for multivariate image analysis and as pre-processing step for cluster analysis (Ding and He 2004; Geladi et al. 1989). An important purpose of PCA is dimensionality reduction, in order to retain only the first few principal components (PCs) that explain the largest amount of variation in a dataset (Wold, Esbensen, and Geladi 1987). This could be applied for instance to store relevant image information in the first number of components, and background noise in the other PCs (Geladi et al. 1989). Several papers have used PCA as a basis for cluster analysis, and showed that using PCs as input for classification algorithms significantly improved clustering accuracy, in comparison to using the original variables of a dataset (Ben-Hur and Guyon 2003; Ding and He 2004).

The challenges for this thesis research are to investigate what kinds of soil and crop data from which types of sensing platforms are available, and to find out which statistical learning methods are appropriate for analysing those data, in order to come up with a potential management zone classification and validation.

1.3 Objective and research questions

Consequently, the main objective of this research is to define and investigate a method to determine how soil variation relates to crop yield, in order to delineate potentially homogeneous management zones for precision agriculture, based on data obtained from remote and proximal sensing technology. This leads to the following main research question:

How does soil variation relate to crop yield, in order to delineate potential management zones for precision agriculture in arable cropping?

This main research question is addressed by the following four sub-research questions:

1. What kinds of geographical soil and crop datasets and methods are available for delineating potential management zones?
2. Which spatial patterns and relationships are observed, both within and between those geographical soil and crop datasets?
3. What method can provide representative delineation of potential management zones based on these spatial patterns and relationships?
4. What is the validity of those potential management zones?

1.4 Outline thesis report

The coming chapter contains a review of literature on developments, data collection, and methods to delineate potential management zones for precision agriculture. Chapter three describes the materials and methods applied for this thesis project, starting with a description of the study area. Next, it explains the methods for pre-processing the used soil and crop datasets, calculating descriptive statistics, and for delineating and validating potential management zones. Chapter four gives interpretations on the results of these analyses, followed by a discussion of those results in chapter five, and chapter six provides conclusions and recommendations for this research. References and appendices are included at the end of this report.

2. Review of literature

In past studies, lots of data and methods have been applied to determine soil and yield variation in order to delineate potential management zones for precision agriculture (PA). This literature review starts with a description of current developments, and continues with an overview of soil and vegetation data acquisition techniques for PA. After that, approaches used as a basis for precision farming will be discussed, including the investigation of soil patterns, crop patterns, and the relationship between soil and crop patterns as a basis for site-specific field management for PA.

2.1 Developments in precision farming for arable cropping

As stated in the introduction, precision agriculture (PA) or precision farming (PF) for arable cropping is the application of technologies and methods to manage spatial and temporal variability, related with all aspects of agricultural production (El Nahry et al. 2011). The main purpose is to optimize crop yields without repressing the environment (Weiss 1996). PA emerged in the 1980s, but has only been practiced commercially since the 1990s, and approaches and technologies for PA have been developing ever since (Mulla 2012). The first step in PA for arable cropping is assessing a field's spatial variability, because crop and soil parameters are subject to both spatial and temporal variation. This variation can be assessed in a number of ways, including soil surveying, soil sampling and spatial interpolation of samples, high resolution sensing, and (statistical) modelling (Pierce and Nowak 1999). Information about spatial soil and crop variability is valuable for site-specific field management (SSFM), which is an important approach for arable farm management (Whelan and McBratney 2003). SSFM has been developing since the late 1990s and has proven to be useful as a basis for soil sampling and the use of variable rate technologies (VRTs) for optimizing nutrient application on agricultural fields (Barnes et al. 2003; El Nahry et al. 2011).

Delineation of management zones (MZs) is an important strategy in SSFM, which has been investigated since the early 2000s (McBratney et al. 2005; Zhang et al. 2010), and can be achieved in a number of ways (Nawar et al. 2017). First of all, creating MZ maps based on farmer knowledge could be an effective approach, since farmers have a broad knowledge about an agricultural field's properties, spatial variation, and the past production history of that field. More on this will be described in section 2.3.1. A second approach is investigating the local geomorphology and using topographical variation and landform properties as a basis for delineating MZs. Other ways proposed to create MZ maps are based on analysis of soil chemical properties or soil classes and based on yield maps and crop coverage (Nawar et al. 2017). Many techniques to obtain information about these properties and spatial variation are available, such as soil sampling, remote sensing, and proximal sensing (Kooistra 2011). More about these techniques will be discussed in sections 2.2.1 and 2.2.2.

A relatively new development in precision farming is the adoption of big data technologies, that are expected to cause large effects on the directions of smart and precision farming (Wolfert et al. 2017). Big data is a concept to describe data with sizes that exceed the capacities of common tools and software to store, analyse, visualize and manage data with respect to time and memory, and is often explained according to the amount of data (volume), speed of data processing and transaction (velocity), different data types and sources (variety), and data reliability (veracity) (Mohanty, Bhuyan, and Chenthati 2015). Examples of information technologies in big data are cloud computing, which is a technology to generate access to shared resources over a large network (such as the internet) for storing and analysing big amounts of data, and the internet of things (IoT), which is a network of physical appliances, devices and vehicles provided with electronics, software and sensors to be able to communicate with each other (Mohanty et al. 2015). For instance, devices such as sensors and robots are nowadays capable to capture real-time sensing data, in addition to high-resolution images and videos that could be used for decision making and to manage arable fields in a fast and concurrent way. Besides this, more and more stakeholders start to play a role in big data for precision farming, such as technology companies and suppliers to provide farmers with the latest technologies, platforms and solutions for efficient monitoring and management of farms (Wolfert et al. 2017). In addition, many start-up companies emerge that offer IT solutions for analysing and visualizing data, and for proposing decision support to farmers. However, these developments also raise questions about data privacy, security and ownership. On the one hand, there is a desire for closed proprietary data architectures, but other stakeholders encourage more freely accessible open-source systems on the other hand, so the discussion about these issues is expected to be an ongoing business analogous to the developments in big data and precision farming (Mohanty et al. 2015; Wolfert et al. 2017).

All described technologies are promising for the developments in precision agriculture. However, it is not the proposed technologies or data itself, rather the methods to transform these data into useful information, the interpretation of that information, and management practices based on those interpretations that may lead to profitable economic and environmental outcomes conceding from precision farming (Pierce and Nowak 1999; Weiss 1996). Back in 2005, a number of strategies were proposed to support developments in precision farming, improvement of management practices, and increased awareness about PA, which still apply today (McBratney et al. 2005). First of all, it was proposed to develop ongoing new equipment and technologies for farmers to support arable management of their fields, such as improved crop and soil sensors, and instruments for seed bed preparation and mechanical weed control. Another suggested approach was to develop a decision support system (DSS) for setting up standardized ways to produce crop yield and soil maps, to develop robust methods for soil and crop data integration, analysis and management zone delineation, and to propose tools and software for farmers, researchers and other stakeholders to use in a user-friendly way. Third, it was proposed to integrate technologies to support whole-farm SSFM practices, rather than investigating each arable field separately, for example by performing cost-benefit analyses for the whole farm (Weiss 1996). And lastly, to raise awareness about all aspects related to precision farming, consumers should be informed about the environmental impact, quality assurance, and product supply chain of production systems that have PA oriented approaches. This can be achieved by linking farmers, students and researchers to exchange knowledge about precision farming, by organizing information meetings at schools and companies, and by media coverage about PA practices and technologies (McBratney et al. 2005).

2.2 Data acquisition for farm and crop management

In the problem statement it was suggested that sensor technology and the integration of data from a large range of remote and proximal sensors are able to show good relations between different sources of data, and that it has a lot of potential for crop monitoring and production over time, related to many applications in precision agriculture (Kooistra et al. 2012; Thessler et al. 2011). This section elaborates more on those applications and data acquisition by means of remote and proximal sensing techniques.

2.2.1 Remote sensing

As stated in the introduction, remote sensing is a set of technologies for collecting data from a remote location without physical interaction, by means of platforms such as satellites or sensors and cameras on board of (un)manned aerial vehicles that measure reflected electro-magnetic radiation (Lillesand et al. 2008). Besides applications in precision agriculture, remote sensing has also been used in research areas such as soil mapping and land use classification (Paul Obade and Lal 2013), detection of regional water leakages (Hadjimitsis et al. 2013), obtaining information to support archaeological applications (Masini and Lasaponara 2007; Papadopoulos and Sarris 2006; Parcak 2009), and wildlife monitoring (McDermid et al. 2009; Raizman et al. 2013). Satellite remote sensing has had applications in agriculture since the 1970s, when the first Landsat (Landsat-1) satellite was launched, for example to classify agricultural landscapes in the US into maize or soybean fields (Bauer and Cipra 1973; Mulla 2012). Another research used Landsat images taken from bare soil in addition to soil samples, in order to estimate spatial patterns in soil organic matter (SOM), soil organic carbon (SOC), phosphorus, and wheat grain yield (Bhatti, Mulla, and Frazier 1991). Data was analysed with the use of classical statistics and geostatistics, and results gave strong evidence for non-random spatial patterns in soil properties and crop yields. Nowadays, the Landsat-8 is used worldwide for earth observation, and the Landsat-9 is expected to be launched in 2020 (USGS 2018). Applications of recent Landsat data are for instance found in crop monitoring and modelling (Roy and Yan 2018), and accurately estimating cropland presence in South America based on time series data (Graesser and Ramankutty 2017). Other widely used earth observation devices to acquire data for land cover mapping and agricultural applications are MODIS and SPOT satellites. Data from these satellites have for instance been used to quantify changes in wetlands by means of landcover classification and creating monthly flood maps (Di Vittorio and Georgakakos 2018), crop yield prediction by regressing crop yield against NDVI time series derived from MODIS data (Nagy, Fehér, and Tamás 2018), and crop classification based on SPOT-5 multispectral data and statistical learning techniques, such as maximum likelihood and support vector machines (Yang, Everitt, and Murden 2011).

Another rapidly developing technology in the field of remote sensing and photogrammetry is the use of multispectral cameras and sensors on board of remotely piloted aerial systems (RPAS), unmanned aerial vehicles (UAVs) or with a popular term *drones*, which is currently innovated enough to be used in the development of GIS products, services and applications (Colomina and Molina 2014). One application is to support precision agriculture by capturing information on soil and plant radiation with help of UAVs, such as the fixed-wing system of eBee shown in Figure 2.1 (SenseFly 2018). UAVs could be a good alternative for data acquisition in comparison to satellites, because of their low cost, high spatial and temporal resolution, and high flexibility in image acquisition (Zhang and Kovacs 2012). Moreover, satellite images are not always available, could be hindered by cloud cover, or have often too coarse resolutions (Kooistra 2011; Zhang and Kovacs 2012). Just as for all types of geospatial datasets, paying attention to data quality also applies to UAV-recorded imagery. One way to improving the measurement accuracy of UAV imagery is to record images supported by ground control points (GPCs), which was for instance applied in the research by Van der Voort (Van der Voort 2016). GPCs are large marks on the ground from which the coordinates are measured with a GNSS device, used to achieve accurately georeferenced images in relation to the real world (Wang et al. 2012).

Satellite sensors and multispectral cameras are capable of sensing reflected soil and plant radiation in different spectral bands, such as the visible bands red, green and blue (RGB) that have wavelengths between approximately 400nm and 700nm, and several infra-red (IR) bands, containing wavelengths of 700nm and higher (Mulla 2012). The reflectance values of these bands can further be used as input to calculate various spectral indices. For instance, two studies calculated spectral indices of bare soil to predict soil properties determined from soil samples. The first research used linear regression on indices derived from publicly available aerial photographs from three study areas in the Netherlands with different soil types to assess soil organic carbon (SOC), and concluded that SOC explained most of the differences between the indices (Bartholomeus and Kooistra 2012). The second study used UAV-recorded RGB images in combination with elevation data to estimate soil organic matter content (SOM), finding that the largest part of the variation in SOM could be explained by RGB derived soil indices (Bartholomeus, Suomalainen, and Kooistra 2014).

Figure 2.1 – eBee for recording multispectral aerial images in four bands: Green, Red, Red-Edge, NIR (SenseFly 2018)

Figure 2.2 – Aerialtronics Altura AT8 Octocopter mounted with Hyperspectral Mapping System (Suomalainen et al. 2014)

Besides spectral indices from bare soil, also many vegetation indices are broadly used, such as the NDVI and WDVI. The NDVI (Normalized Difference Vegetation Index) is an index that is used to detect and determine diverse vegetation properties. The idea behind it is that vegetation absorbs a large portion of visible light (RGB) for photosynthesis, whereas near infrared (NIR) light is mostly reflected, since it is hardly used for photosynthesis (Lillesand et al. 2008). The values range between -1 and 1, for which a value of for example 0.2 indicates poor vegetation and 0.8 shows healthy, abundant vegetation. Hence, the higher the value, the healthier or more developed the vegetation is. The differences in reflectance also depend on the amount of available sunlight, which is the reason for dividing (normalizing) the difference between NIR and Red by the total incoming light (NIR + Red) (Lillesand et al. 2008). In this (and many other) applications, the red band is chosen as visible light band, but the blue or green band can be used as well. The WDVI (Weighted Difference Vegetation Index) is another index for the assessment of vegetation properties. This index is developed for the estimation of the thickness of vegetation's canopy, or leaf area index (LAI). It is calculated by subtracting the contribution of the soil's reflectance from the vegetation's reflectance, with the assumption that the ratio between infrared and red reflectance of bare soil is constant (Clevers 1991). Since the soil's reflectance is filtered from the signal, the differences between scarce and abundant vegetation are more prominent for the WDVI than for the NDVI, which is beneficial to assessing the differences in the development of vegetation (Clevers 1991). One example to use the WDVI (among other vegetation indices) was to assess the status of potato crops at a study field of Van den Borne potato farm as an application in PA, based on Sentinel-2 satellite data (Clevers, Kooistra, and van den Brande 2017). The WDVI and other indices proved to be of good predictive power to estimate LAI of the potato crops. In addition to these bands and indices, another special spectral band, called the red-edge position (REP), is frequently used as well. The wavelengths between the upper limits of the red band (670nm) and the lower limits of the NIR band (780 nm) indicate a sudden increase in vegetation's reflectance, causing the REP's wavelength to be around 720 nm (Clevers et al. 2001). Whereas the NDVI and WDVI focus more on LAI, the REP has also proven to be an important parameter to derive the chlorophyll content of vegetation, which is another indicator for the condition of plants (Adamczyk and Osberger 2015). The REP and related spectral indices have found their way in several studies monitoring the development of many different crops. For instance, one paper used the REP to estimate the LAI from nine different crops on a test area in Spain, and discovered a strong correlation between the index and LAI, in comparison to a lower correlation between NDVI and LAI (Delegido et al. 2013). Also good correlations were found between the normalized difference red edge (NDRE) index and N uptake in a study to estimate nitrogen concentration from maize on an experimental field in Quzhou County, China (Li et al. 2014).

Imaging spectroscopy, or hyperspectral imaging, is another advancing technology in the field of remote sensing. Image spectroscopy is the simultaneous acquisition of co-registered images (images with exactly the same spatial resolution, extents and coordinate system) in a large range of narrow and contiguous spectral bands by means of remote sensing techniques (Schaepman et al. 2009). Because of the many available spectral bands, the technology is for instance very suitable for accurately measuring soil processes and properties, such as salinity, erosion, soil formation, soil organic matter and contamination (Ben-Dor et al. 2009), and the assessment of chlorophyll and nitrogen content in vegetation (Clevers and Kooistra 2012). At the Laboratory of Geo-Information Science and Remote Sensing at Wageningen University, a hyperspectral mapping system (HYMSY) was developed especially to be mounted on an Altura AT8 Octocopter, which is a rotor-based UAV as shown in Figure 2.2. The system contains a *pushbroom* spectrometer and a photogrammetric camera, intended for research and applications to support characterization of crop variation and monitoring potato crops on agricultural parcels (Suomalainen et al. 2014). During the past years, other research projects have been using data from UAV-based hyperspectral systems for investigation various agricultural applications as well. First of all, one paper described promising applications in calculating parcel sizes of paddy field areas, palm tree mapping, and sugar cane estimation in Indonesia, based on orthophotos and 3D digital elevation models (DEMs) collected with UAV-based cameras and sensors (Rokhmana 2015). Similar methods were applied in two case studies monitoring the decline of chestnut trees and the development of vines in Portugal based on data recorded with and eBee (SenseFly 2018), and it was suggested that those methods could be suitable to support decision making in relation to site-specific field management (Pádua et al. 2017). Other papers used hyperspectral imagery to predict soil properties of agricultural fields in the USA by means of soil sampling and using partial least squares (PLS) to predict soil properties based on the image's spectral bands, stating that high resolution soil maps could be a suitable tool to support site-specific field management of farmlands (Hively et al. 2011), and to derive red-edge vegetation indices for estimating nitrogen (N) uptake of winter wheat in China, showing moderate to strong correlations between those indices and N uptake (Feng et al. 2015).

Besides the use of optical remote sensing, it was argued that UAV-based thermal remote sensing is another promising technology, for instance in the fields of drought monitoring, plant disease detection, soil mapping, crop monitoring and crop yield estimation, although several limitations such as calibration, atmospheric absorption and disruption by changing weather conditions have to be solved first before large-scale use of this technology is possible (Khanal, Fulton, and Shearer 2017).

2.2.2 Proximal sensing

As a complement to remote sensing, spatial patterns are also measured by means of proximal sensing technology as a basis for precision agriculture. Proximal (or close) sensing is a range of techniques to measure for instance chemical, physical and biological soil properties and crop yield with help of sensors at no more than about 2 meters above ground level (Bartholomeus et al. 2011; Stoorvogel et al. 2015). Soil electric conductivity (EC) is one proximally sensed variable that is able to detect local differences and variation in the soil (McNeill 1980), making it potentially suitable information to explain other soil properties determining an agricultural field's productivity, such as texture, water holding capacity, porosity, salinity, and temperature (Grisso et al. 2009). For instance, one study investigated the potential of apparent soil electric conductivity (ECa) to predict the spatial distribution of clay and pH as a basis for lime application, and confirmed that soil ECa was indeed suitable to serve as an indicator for other soil properties, such as clay content and pH (Sanches et al. 2018). Moreover, EC also proved to be closely correlated with wheat and corn yield as a basis for SSFM (Johnson et al. 2003). Therefore, electromagnetic induction (EMI) sensors measuring soil EC are increasingly applied as potential information to derive soil properties for arable management practices (Corwin and Lesch 2005). The coming paragraphs discuss the acquisition of soil and crop variables such as EC, soil organic matter (SOM), pH, and nitrogen (N) content by means of EMI and other proximal sensing techniques.

Figure 2.3 shows a schematic view from the principle of measuring soil EC. In general, a sensor contains one transmitter and one or more receiver coils. The transmitter sends different magnitudes of currents into the soil, generating a primary magnetic field (*quadrature* (or *quad*) phase, or H_p) that induces current loops as displayed in Figure 2.3 (Corwin and Lesch 2003). Consequently, those current loops generate a second magnetic field (*in* phase or H_i), which is measured by the receiver coil or coils. For instance, an EM38-MK2 soil sensor as shown in Figure 2.4 and on the image on the front page (Van den Borne 2018a) is equipped with three coils: one transmitter and two receivers that are placed at 0.5 and 1 meters horizontally from the transmitter. For that reason, an EM38-MK2 is able to measure EC at both 0.5 meters and 1.0 meters soil depth, since the horizontal distances of the coils are proportional to the vertical soil depths. Similarly, a Veris sensing platform measures EC at 0.3 meters and 0.9 meters soil depth. The ratio between the quad phase and in phase is proportional to the electric conductivity of the soil (McNeill 1980). From the quad-phase data, the apparent soil electric conductivity (ECa) is calculated by multiplying EC by a temperature correction factor (Corwin and Lesch 2005; Ma et al. 2011). Apparent electric conductivity is expressed as mean conductivity of the soil's volume in mS/m (Saey et al. 2013). However, sensors such as the EM38-MK2 are often calibrated in a way that under low induction the quad phase output is directly stored as ECa (Geonics Ltd. 2008).

In addition to ECa, devices such as Veris sensing platforms are also equipped to measure variables like soil organic matter (SOM) and soil pH. These variables are recorded with separate sensors on a mobile sensor platform that could for instance be mounted on a tractor. Those sensors measure soil variables on-the-go, and link the measurements to spatial coordinates by means of GNSS (Vantage Agrometius 2018). For instance, SOM is measured by sending light pulses into the soil and measuring the reflection of the incoming red and near infrared (NIR) wavebands of light (Schans and Berg 2013). The reflection data including geolocation is send to a spectrophotometer and transformed into suitable values to create a SOM map. In addition, pH is measured by two ion-selective pH electrodes that are brought into contact with the soil particles. Every recording is the average of the two electrodes, in order to validate the measurements and to filter out possible measurement errors (Schirrmann et al. 2011).

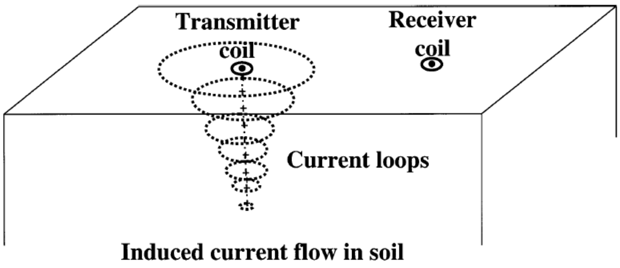


Figure 2.3 – Operation principle of the electromagnetic soil conductivity meter (Corwin and Lesch 2003)

Figure 2.4 – EM38-MK2 Soil sensor (Van den Borne 2018a)

Similar to hyperspectral imaging in remote sensing, proximal soil spectroscopy has increasingly been used in the field of soil science, since those techniques are able to measure multiple soil properties from just a single scan in a relatively inexpensive, non-destructive way, and with a minimum amount of preparation compared to soil sampling (Viscarra Rossel 2011). In the past decades, various studies have been conducted with respect to soil spectroscopy. Back in 1998, the same author evaluated if a potential soil sensor could be used for proximal measurements in the field, in order to formulate management decisions for precision agriculture, by taking soil samples from the surface horizon of an area in New South Wales, Australia, and taking spectral measurements from the samples with a PIMA II spectrometer in different wavelengths ranging from 1300 to 2500 nm (Viscarra Rossel and McBratney 1998). After performing statistical analysis on the measurements, reflectance values showed significant results to clay and water content, but not to organic matter. Therefore, the method was considered to be useful for agricultural management practices, although further refinement of the sensor for organic matter content would be necessary. A more recent study measured hyperspectral reflectance data from various locations across the United States with an on-the-go spectrometer pulled behind a tractor, for the purpose of developing partial least square (PLS) models to predict chemical properties of the soil (Christy 2008). The models showed significant results and were used to calculate prediction maps of the chemical properties of the soil, which were suggested to be a substantial source of information for management practices in precision agriculture.

In addition to collecting soil measurements, proximal sensing also has applications in crop yield data acquisition. Two ways to collect crop yield data with ground sensors are by determining the yield directly, and by measuring the electromagnetic reflection of crops. Direct crop yield (in ton per hectare) is being measured during harvest by a sensing system on board of a tractor, such as the *Yield Master Pro* by Probotiq that combines information of load cells, speed sensors and GNSS measurements to produce crop yield maps (Van den Borne 2018b). Optical reflection from crops could for instance be measured by two Fritzmeier ISARIA sensors parallelly attached at the front side of a tractor (Figure 2.5 and Figure 2.6), that are able to measure reflectance values of soil and vegetation in several spectral bands (Van den Borne 2018b; Fritzmeier-Umwelttechnik 2016). These values are consecutively used to calculate nitrogen (IRM) and biomass (IBI) vegetation indices by means of a black box operation, for which the input values of the four bands are not known. Just as for direct crop measurements, GNSS information is collected as well and linked to the crop measurements, to be able to produce N and biomass maps based on the determined vegetation indices.

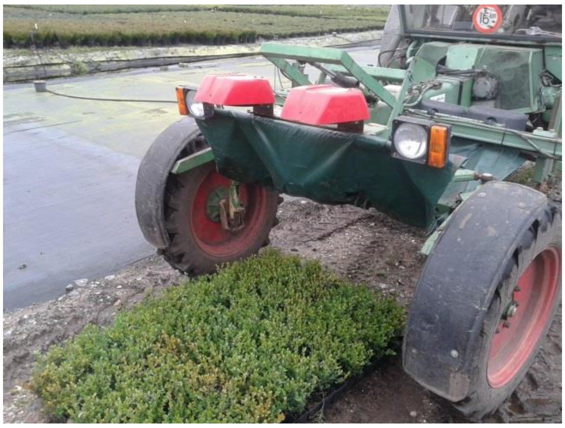


Figure 2.5 – Fritzmeier sensors attached on the front side of a tractor, measuring crop samples

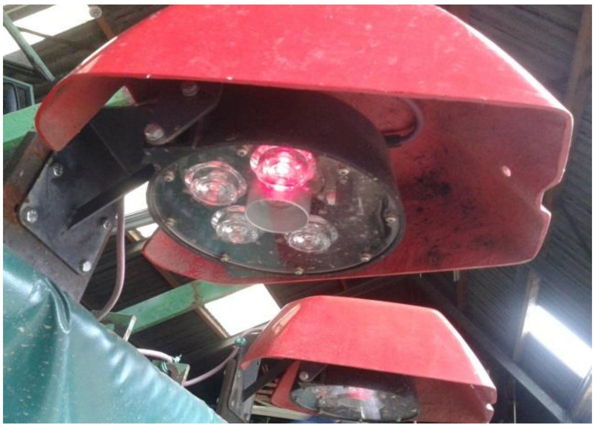


Figure 2.6 – Close-up from Fritzmeier sensor, the four sensors to measure IBI and IRM are visible

Another study measured NDVI from growing canola crops with a hand-held optical sensor to use in experiments conducted on five agricultural fields across Canada between 2004 to 2007, for the purpose of discovering potential relations between the NDVI of the crops and various seed and nitrogen (N) rate inputs (Holzapfel et al. 2009). Although some of the experiments were disturbed by bad weather conditions, regression analyses between NDVI, seed and N input showed moderate correlations, suggesting that optical sensors could be a valuable tool to obtain crop yield information as a basis for applying variable N input rates to support site-specific field management for PA.

2.3 Approaches to delineating potential management zones

Investigating the spatial patterns of soil and crop variables from arable fields is an important first step for potential management zone delineation to support in precision agriculture, because those patterns have a significant impact on the modelling and optimization of crop yield (Kuang et al. 2012). Therefore, this section discusses approaches that have been used in previous studies for delineating potentially homogeneous management zones, such as investigating within-field knowledge of farmers and using unsupervised classification techniques on the detected soil and crop yield patterns.

2.3.1 Farmer’s knowledge about spatial patterns on agricultural fields

Precision farming is often technology-driven, with a focus on new developments of sensing and measuring techniques, but the primary emphasis should be on farmers who are responsible for decisions with respect to their farm management (Stoorvogel et al. 2015). In addition, a research regarding precision agriculture in cotton farming in the US already suggested that younger, more educated cotton farmers exploiting large farms were more likely to implement SSFM techniques than other farmers (Roberts et al. 2004). Evaluating spatial within-field knowledge of farmers could be an effective approach, since farmer’s expert knowledge is a valuable source of information for precision farming (Fleming et al. 2000). For example, one study underpinned this by investigating the spatial knowledge of farmers to determine within-field soil variation with the help of aerial photographs (Heijting, De Bruin, and Bregt 2011). Semi-structured interviewing techniques were applied together with soil sampling to determine and validate the within-field knowledge of farmers. It turned out that the farmers had substantial spatial knowledge about their fields, which they used intuitively for the site-specific field management of their lands. To make this spatial knowledge more concrete, it could for example be formalized into a decision support system (DSS) determining how to apply farm management based on local circumstances in the field. In addition, a paper evaluating a combination of farmer knowledge, PA tools and crop simulation modelling also concluded that farmers had a good understanding of spatial patterns and poor crop producing regions within their arable fields (Oliver, Robertson, and Wong 2010). Another study compared MZs derived from bare soil colour on aerial images assessed by farmers with MZs delineated based on spatial patterns from ECa, and concluded that both methods indeed seemed to identify homogeneous zones within arable fields, although the EC method was slightly more effective for one of the fields than the farmer based method (Fleming, Heermann, and Westfall 2004).

2.3.2 Management zone delineation based on crop yield information

In order to delineate potential MZs, crop (and soil) variables are often used as input for cluster analysis algorithms, which are a collection of unsupervised classification techniques (James et al. 2013). Two widely used types of cluster analyses for delineating potential MZs are k-means and fuzzy c-means (FCM) clustering (sometimes also referred to as fuzzy k-means or FKM). The aim of k-means clustering is to allocate the values of given input variables into one of the *k* number of clusters, in such a way that the sum of squares within each cluster is as low as possible (Hartigan and Wong 1979). The method of FCM (or FKM) clustering is very similar, with the adaptation that all values are classified into all clusters, provided with a weighting factor that determines the maximum likelihood for a value to belong to a certain cluster (Bezdek, Ehrlich, and Full 1984).

One way to delineate potentially homogeneous MZs with cluster analysis is based on (past) crop yield information. As argued before, crop yield data of multiple years were suggested to use for the purpose of delineating stable management zones (Boydell and McBratney 2002; Mulla 2012; Stoorvogel et al. 2015). For instance, one study investigated management zone delineation by applying three different clustering techniques on past crop yield data and comparing the results with yield responses to variable N rate inputs (Milne et al. 2012). The three clustering algorithms included two 'hard' k-means techniques based on (dis)similarities between crop yield, and one 'soft' classification technique that was based on computing membership to fuzzy classes. Figure 2.7 shows the classification results containing five crop yield classes, with (a) and (b) being the 'hard' means algorithms and (c) the 'soft' membership classification. The third one is clearly the most practical classification map of the three.

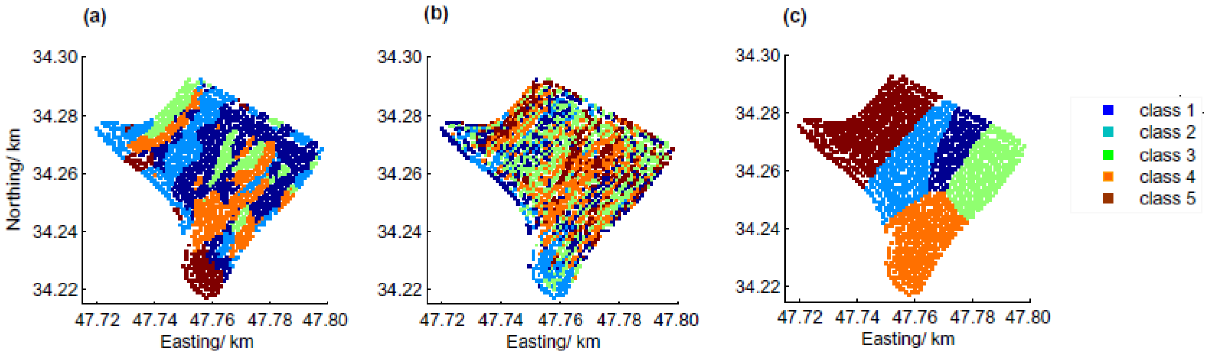


Figure 2.7 – Three different fuzzy classification results (Milne et al. 2012)

Another paper used estimated cotton yield maps of multiple fields in New South Wales, Australia from multiple years ranging from 1988 to 1998 as a basis for FKM cluster analysis to delineate potential MZs (Boydell and McBratney 2002). After optimizing the number of clusters and identifying maps of years having possibly undesirable influences on the outcomes, the authors concluded that reasonably stable yield zones could be derived from multi-year estimates with data from 5 years (± 2 years), that data should be distinguished based on proposed water management, and that the used approach could allow potential MZ delineation to support field investigations for the Australian cotton industry. A different study used historical crop yield maps from multiple years in addition to the NDVI, the red spectral band, and plant surface temperature derived from remotely sensed images from multiple fields across the Midwest of the USA, and examined the correlation between these covariates to discover the best predictors for within-field crop yield information (Maestrini and Basso 2018). Results showed that historical yield maps provide the best prediction for spatial variability over time, while the remotely sensed images were better predictors for within-field spatial variability.

2.3.3 Management zone delineation based on combined soil and crop yield information

Other methods to support management zone delineation are either based on soil information solely, or on a combination of soil and yield information. For instance, one paper studied the relationships between soil properties and grain yield on three different farms (representing a high-, medium- and low-yielding zone) in Sweden, and evaluated whether hydraulic conductivity of saturated fields could be an indicator of variable crop yield for the purpose of site-specific field management (Keller et al. 2012). Soil properties were derived from soil samples, field saturated hydraulic conductivity was measured, and yields were recorded by sensors attached on crop harvesters. Results based on regression analysis showed that on average low-yielding zones were characterized by higher bulk density and lower hydraulic conductivity than medium and high-yielding zones, indicating that soil structure is a valuable source to consider for precision farming practices.

As input for a FCM cluster analysis to delineate and validate potential MZs, one study conducted a correlation analysis for selecting and normalizing soil variables such as elevation, slope, soil penetration resistance (SPR), clay content, and soil organic matter (SOM) (Schenatto et al. 2017). The study was conducted for three experimental fields in Brazil between 2010 and 2014, and results showed that if the clustering algorithm used multiple variables with different measurement scales, normalization of the variables was required to get reliable classification results. Another study collected soil samples in addition to spectral reflection measurements of winter wheat from an agricultural parcel in Belgium to use as input in FCM cluster analysis (Vrindts et al. 2005), and concluded that cluster analysis based on the correlation between soil and crop information was a more effective way to delineate potential management zones compared to cluster analysis based on soil information solely. For another research to an agricultural field in Belgium, apparent electric conductivity (ECa) measurements were collected, in addition to soil samples for analysing topsoil and subsoil clay content (Vitharana et al. 2006). Good correlations were found between ECa and subsoil clay content, but weaker correlations for topsoil clay content. Therefore, subsoil clay content in addition to gravimetric water content were chosen to create a potentially homogeneous MZ map from (with k-means cluster analysis), and this map was compared to an aerial image of sugar beet crops, resembling the spatial patterns and confirming the reliability of the results.

From a study area in Venice in Italy that mostly contained silt-clay soil (left side of Figure 2.8), EC measurements were collected in combination with soil sample data and NDVI values derived from bare soil reflectance (Scudiero et al. 2013). Goals were to discover possible correlations between these covariates, and to apply this information to delineate potentially homogeneous management zones using FCM classification. Results from an analysis of variance (ANOVA) indicated that five management zones were most suitable (right side of Figure 2.8), since the least amount of within-unit variance of the investigated variables remained for each of the five zones.

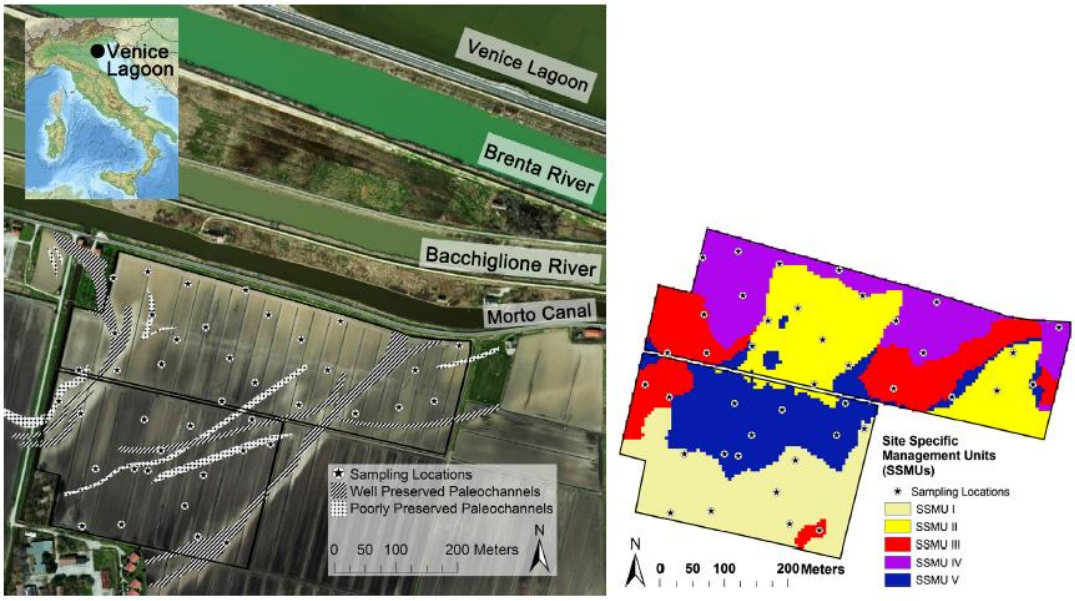


Figure 2.8 – Study area Italy (left) and classified management zone map of the area (Scudiero et al. 2013)

For several experimental fields in Colorado, USA, relationships between EC measurements collected with a Veris soil sensor, and maize yield measured with a grain yield monitor and a GPS device were investigated (Johnson et al. 2003). This study also used EC-based management zone maps and crop yield maps to compare with soil properties determined from grid sampling. Correlation and regression analyses were conducted, and most of the results showed strong correlations between the different soil and crop variables, concluding that EC determination and the use of management zone maps provide valuable information for site-specific field management in precision agriculture. Directed soil sampling was also used in a paper by (Whelan and McBratney 2003) to validate delineated management zones based on differences between wheat yield, soil EC and elevation for a number of arable fields in New South Wales, Australia. Significant differences between zones were found for some of the derived soil properties, and for the purpose of effective field management suggestions were made to take a soil's water holding capacity in combination with early seasonal indicators into account, to discard areas of low yield potential, and to change land use or reduce various farming inputs.

Another paper combined very high-resolution vegetation images from two years obtained by the QuickBird satellite and digital maps derived from soil sampling to derive leaf area index (LAI) maps of a commercial maize field in Huesca, Spain (López-Lozano, Casterad, and Herrero 2010). These maps included mineral concentrations extracted from soil samples and EMI-measurements, elevation, and a grain yield map from 2004. Significant spatial relations were found between these covariates. Based on those relations, different site-specific management zones were proposed, see Figure 2.9 for a possible management zone solution. A comparable study was done by (Haghverdi et al. 2015) for an agricultural field for cultivating cotton in Tennessee, USA, although more than one clustering algorithm (including k-means, Gaussian and integer linear programming) was applied and compared to delineate potential management zones. Similar to other articles, the authors of both papers suggested that the delineation of such management zones is an important step in the application of variable rate technologies (VRT) for as a basis for site-specific field management (SSFM).

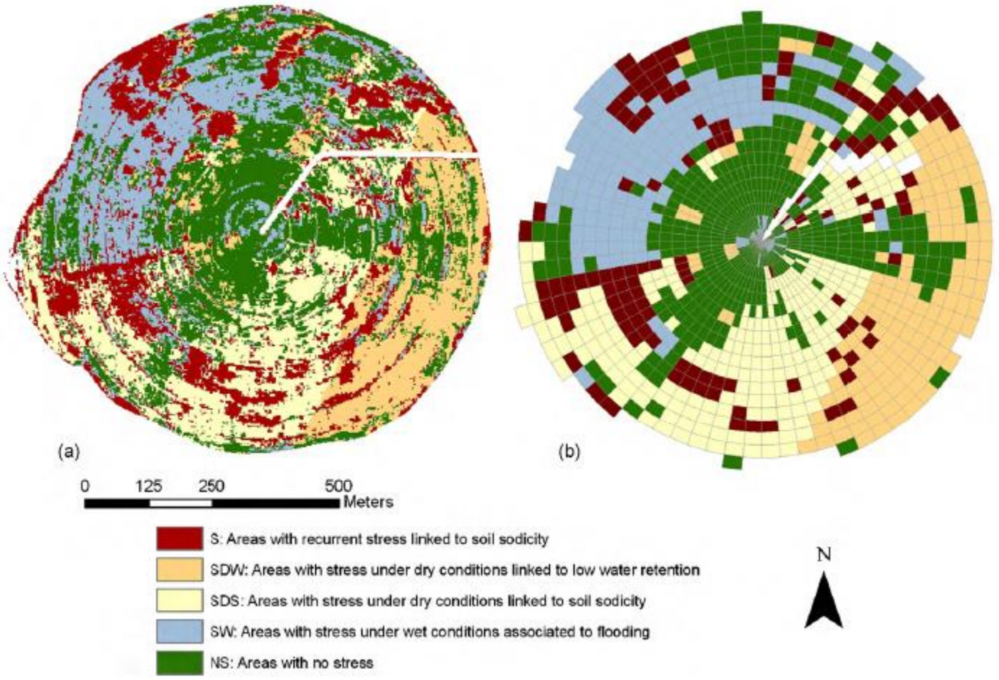

(a)
(b)
0
125
250
500
Meters
S: Areas with recurrent stress linked to soil sodicity
SDW: Areas with stress under dry conditions linked to low water retention
SDS: Areas with stress under dry conditions linked to soil sodicity
SW: Areas with stress under wet conditions associated to flooding
NS: Areas with no stress
N

Figure 2.9 – Site-specific management units derived from LAI and soil maps (left) and homogeneous rectangular delineated zones for applying variable rate technologies (VRT) (López-Lozano et al. 2010)

2.3.4 Management zone delineation based on principal component analysis

Besides using soil and crop variables directly as input for cluster analysis, principal component analysis (PCA) is another frequently used approach as a pre-processing step for cluster analysis (Ding and He 2004). The principle of PCA is to transform a given set of variables into a new uncorrelated set of variables, called the principal components or PCs, ordered by a descending amount of explained variation from the total dataset (Jolliffe 2014). One purpose of PCA is dimensionality reduction, retaining only the first few PCs explaining the largest amount of variation in a dataset (Wold et al. 1987). A few decades ago, one study performed a multivariate image analysis based on PCA, in order to store relevant image information in the first number of components and background noise in the other components, and already stated that the method was useful for detecting outliers, and for setting up a better strategy for image analysis (Geladi et al. 1989). As described in section 1.2, two papers attempted to find key soil and topographic variables for delineating potentially homogeneous management zones by collecting apparent soil electric conductivity (ECa) measurements in addition to soil samples and gamma ray measurements for study sites across Belgium, and performing principal component analyses on the derived soil variables (Van Meirvenne et al. 2012; Vitharana et al. 2008). Covariates representing the largest amount of variation in the first PCs (and therefore being the largest contributors responsible for the soil patterns in the fields) were soil ECa, elevation, and soil pH.

In addition, some papers suggested that using PCs as input for classification algorithms significantly improved clustering accuracy and stability compared to using the original variables of a dataset as input for cluster analysis (Ben-Hur and Guyon 2003; Ding and He 2004). One paper used interpolated ECa measurements collected with a Veris 3100 sensor (Vantage Agrometius 2018) from a study field in the state of Paraná, Brazil, in addition to eleven other soil parameters derived from soil sampling such as pH, SOM, phosphorus, and cation exchange capacity as input for a PCA analysis, and consecutively used two PCs explaining 80% of the total variation of the dataset in a FKM cluster algorithm (Molin and Castro 2008). This resulted in a management zone map containing three homogeneous zones that were validated using a one-way ANOVA on all variables, indicating that significant differences occurred between zones for most of the soil variables. A similar study was conducted for an arable field in Badajoz in the southwest of Spain (Moral, Terrón, and Silva 2010), for which a management zone map with two clusters was derived based on two PCs explaining about 99% of the total variance, and validated by a visual comparison between the spatial distribution from both the MZs and the soil properties. Another paper describes the delineation of management zones by deriving soil properties such as clay content, pH, CEC, soil organic carbon (SOC), an nitrogen content from soil samples collected in an arable region in the north of Iran that contained approximately 24,000 ha of paddy fields (left side of Figure 2.10) (Davatgar, Neishabouri, and Sepaskhah 2012). After interpolation by means of kriging, the variables were used in a PCA, and the first three PCs best representing the paddy field properties were used in a fuzzy c-means classification to divide the field into four MZs (right side of Figure 2.10) based on the within-group variability of soil fertility.

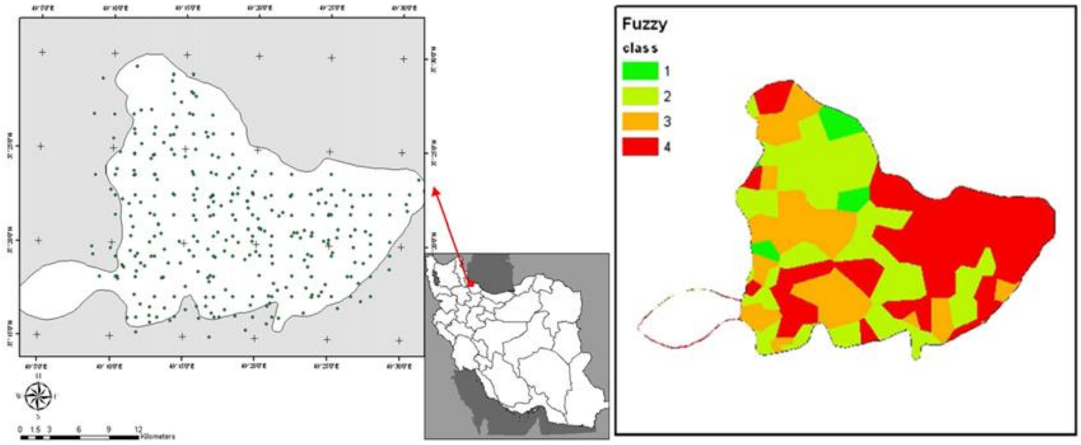


Figure 2.10 – Study area Iran (left) and classification result of the area with 4 MZs (Davatgar et al. 2012)

A recently developed extension to PCA, called MULTISPATI-PCA, is aimed at integrating spatial information before computing the actual spatial principal components (sPCs), which is achieved by calculating a parameter called Moran's index that determines spatial dependence (or autocorrelation) between local observations in relation to their neighbouring observations, and providing those values with spatial weighs signifying the magnitude of the spatial dependence (Dray, Saïd, and Débias 2008). During the past years, a number of papers have incorporated this approach in their researches. For instance, one study compared the performance of ordinary PCA with MULTISPATI PCA on data collected from an agricultural field in the southeast of Argentina (Costa 2012), and discovered that the extended PCA detected correlations within the data that were not discovered by ordinary PCA, suggesting that MULTISPATI-PCA would be a promising tool to map spatial variability within agricultural fields to support management zone delineation for precision agriculture. Another study that collected data from arable fields in the southeast in Argentina (Cordoba et al. 2013), compared management zones delineated by k-means cluster analysis based on the original soil variables, on the first PCs of ordinary PCA, and on the first sPCs of MULTISPATI-PCA, and concluded that the latter approach notably improved the delineations in comparison to ordinary PCA. A similar study on comparing three unsupervised methods was conducted for an experimental field in Brazil (Gili et al. 2017), and concluded that the methods incorporating spatial structure in soil data in general performed better, but that the choice of the method depends on the objectives of crop management, the main yield-limiting factors, and the agro-ecological conditions of the field.

3. Materials and methods

The different analysis stages were performed based on the research questions defined in section 1.3. The first sub-research question was covered by an extensive literature review included in chapter 2 on data and methods that have been applied in past research projects in relation to precision agriculture. The second sub-research question was attended by performing descriptive and visual data analyses, and the third sub-research question was focused on finding a method to delineate potential management zones (MZ) by performing principal component analysis (PCA), and by delineating MZs represented by clusters based on the extracted spatial principal components (sPCs). The final sub-research question aimed at detecting and validating potato crop yield variation, both between and within the delineated MZs, by fitting a number of linear models. In addition to a description of the study area, this chapter will elaborate on the applied methodology for data pre-processing and data analysis, for the purpose of addressing those research questions.

3.1 Description of the study area

The study area for this research is focused on the potato farm of Van Den Borne, which is located near Reusel, south-west of Eindhoven in the province of Noord Brabant, on the Dutch-Belgian border (Figure 3.1). In addition, Van Den Borne farm is very near to the *Maatschap Gebroeders Laarakker* farm in Reusel. For that farm, an investigation to the region's soil quality and hydrology was conducted in 1980, from which the results still apply today. The soil consists of poorly loamy sand that is deposited by wind, and has grain sizes of about 160 µm (Dekkers 1981). On many locations, the soil structure consists of a large A-horizon on top of a shallow B horizon. These horizons are deposited on a C horizon consisting of densely packed layers of sand (Stoorvogel et al. 2015). The sand belongs to the Formation of Sterksel, a river deposition from the Early and Middle Pleistocene epochs (Naturalis 2018). The soil also contains some gravel, larger rocks, and some small layers of humus. However, these layers are rather scattered and therefore not easily quantifiable (Dekkers 1981).

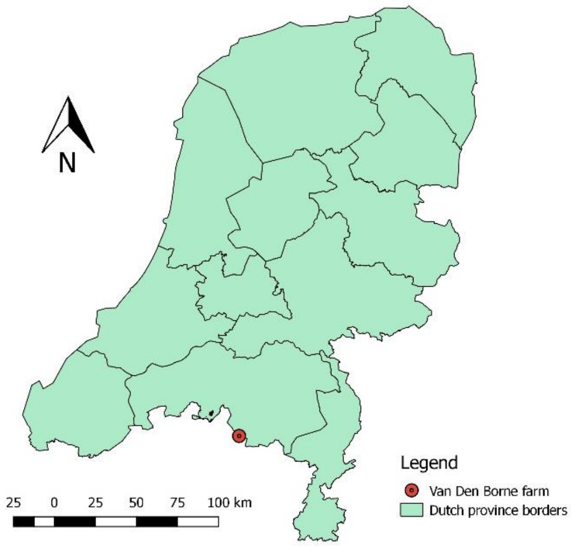


Figure 3.1 – Location of Van Den Borne Potato farm in the province of Noord Brabant, The Netherlands

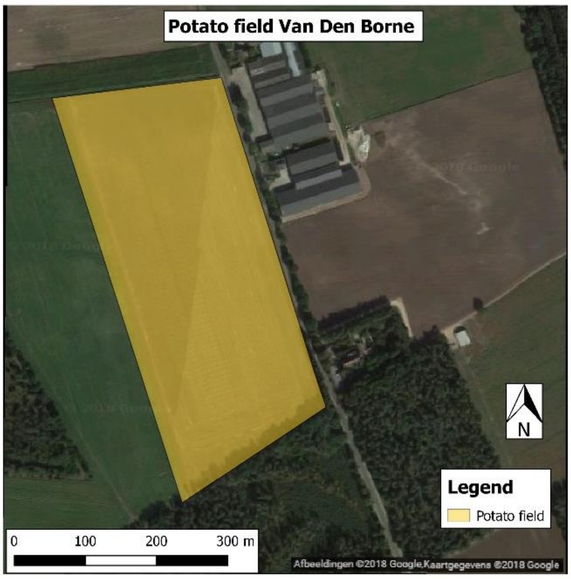

Potato field Van Den Borne
N
Legend
Potato field
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

Figure 3.2 – Parcel of Van Den Borne farm used for this research

Van Den Borne farm is specialized in arable farming, mainly for cultivating potatoes (*Solanum tuberosum*), but also sugar beet (*Beta vulgaris*) and corn (*Zea mays*) are produced (Van den Borne 2018a). On several parcels, many vegetation properties have been measured from different kinds of crops. Also soil electric conductivity has been measured regularly, so lots of soil and crop data of multiple growing seasons are available, for instance to use as a basis for site-specific field management (Kooistra et al. 2011). The focus of this research lies on one of the parcels at the farm, which is located across the farm buildings (Figure 3.2). The parcel has a size of approximately 12.7 ha and its mid-field coordinates are 51°19’03.4”N and 5°10’23.4”E (WGS84).

For many parcels on the farm a crop rotation cycle of four years is applied, meaning that one specific crop is cultivated once every four years, and during the growing season of 2015 the parcel was used to cultivate potatoes (Van den Brande 2015). Topographic maps, such as the one from 1980 extracted from the *Topotijdreis* website (Kadaster 2018c) that was also included in the study by (Dekkers 1981) (Figure 3.3), show that the parcel is located along the Postelse Dijk on the Dutch-Belgian border (represented by the yellow-black line), and that it contains a ditch for water run-off (indicated with a blue line) at the north of the field. This situation has not been changed much; more recent maps and images show very similar patterns.

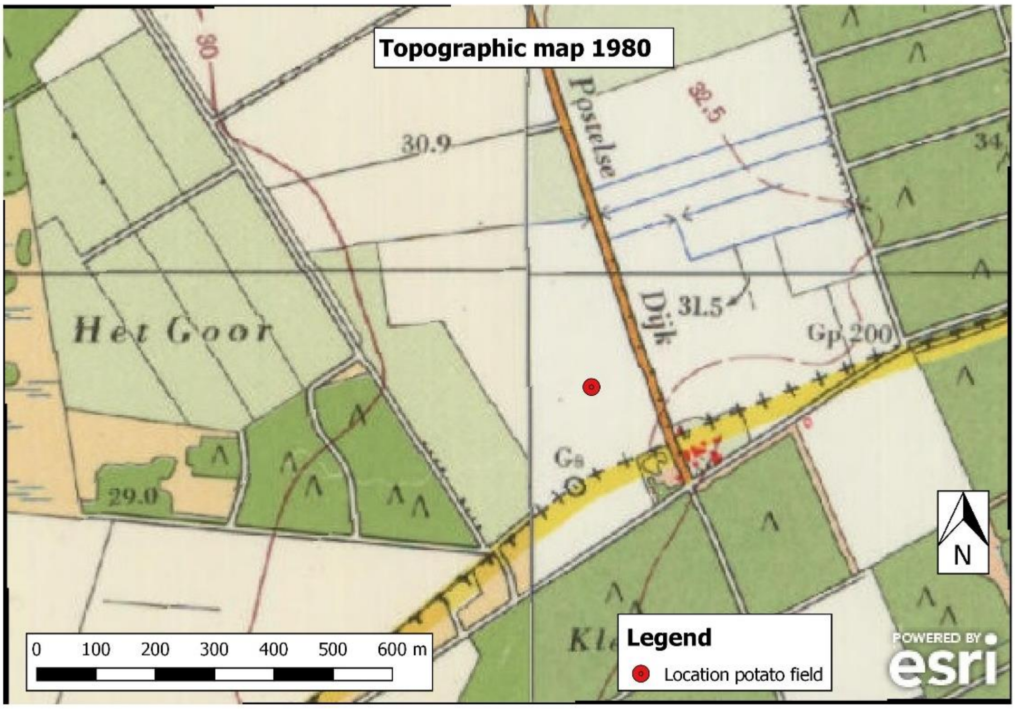


Figure 3.3 – Topographic map from the location of Van Den Borne farm from 1980 (Kadaster 2018c)

3.2 Description and pre-processing of datasets

In this section, the datasets used for this research are described, and the data pre-processing are explained. Data pre-processing was performed in R, an open-source programming language to perform extensive statistical analysis and visualizations on large (geospatial) datasets (Bivand et al. 2008). R was accessed via RStudio, an open source integrated development environment (IDE) that interacts with R, making it appropriate for data analysis and visualization in a user-friendly way (RStudio 2018).

3.2.1 Available datasets and terminology

For this research, a number of initial datasets were available to use for pre-processing and data analysis, including aerial images of bare soil and potato crops, apparent electric conductivity (ECa) at four soil depths, potato crop yield, elevation, soil depth of the A-horizon, and soil organic matter (SOM). Characteristics and specifications of these datasets are included in Appendix 1. The pre-processing steps applied on all soil and potato crop datasets are summarized in Figure 3.4, and will be further explained in the coming subsections.

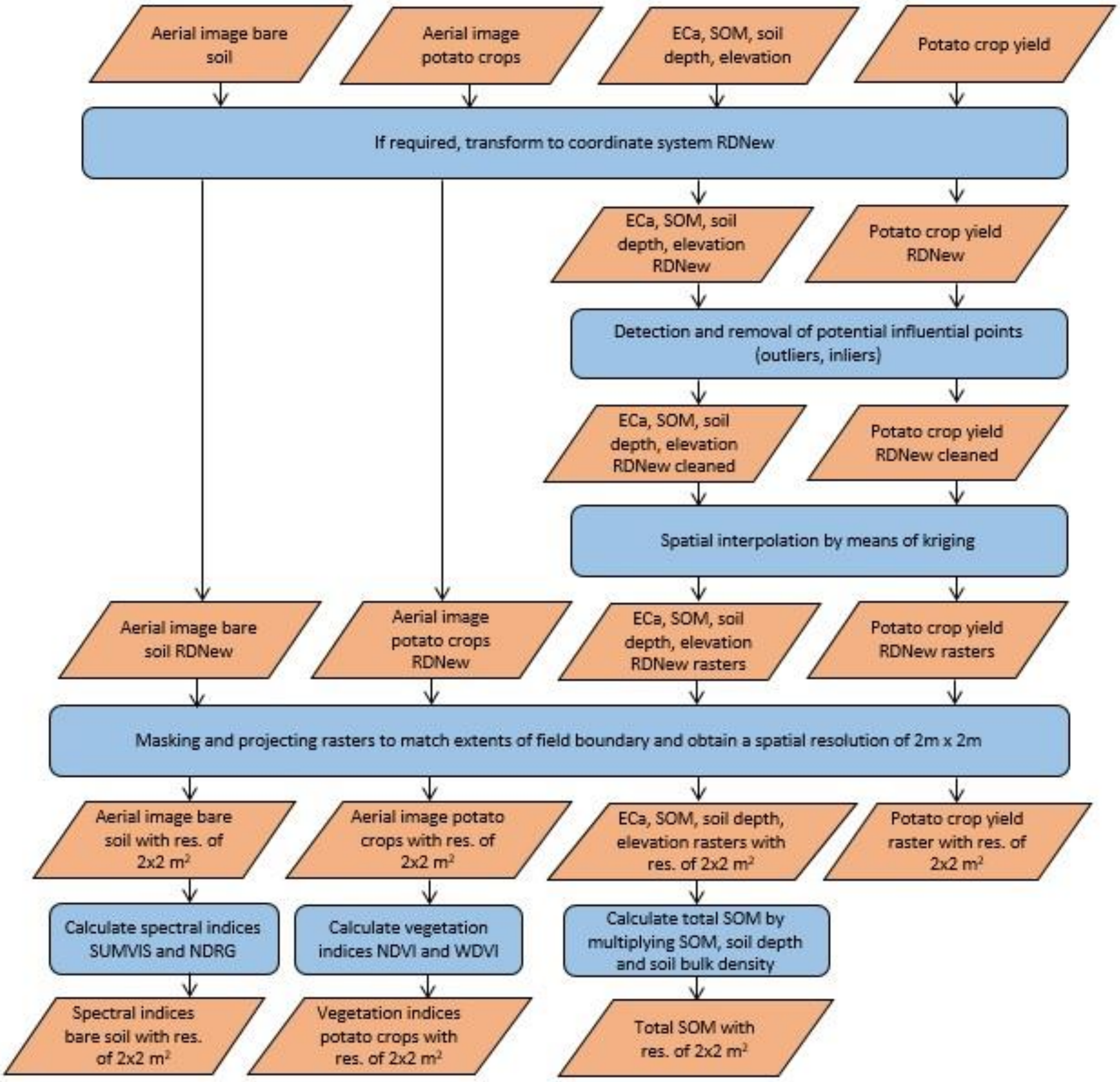

Aerial image bare soil
Aerial image potato crops
ECa, SOM, soil depth, elevation
Potato crop yield
If required, transform to coordinate system RDNew
ECa, SOM, soil depth, elevation RDNew
Potato crop yield RDNew
Detection and removal of potential influential points (outliers, inliers)
ECa, SOM, soil depth, elevation RDNew cleaned
Potato crop yield RDNew cleaned
Spatial interpolation by means of kriging
Aerial image bare soil RDNew
Aerial image potato crops RDNew
ECa, SOM, soil depth, elevation RDNew rasters
Potato crop yield RDNew rasters
Masking and projecting rasters to match extents of field boundary and obtain a spatial resolution of 2m x 2m
Aerial image bare soil with res. of 2x2 m²
Aerial image potato crops with res. of 2x2 m²
ECa, SOM, soil depth, elevation rasters with res. of 2x2 m²
Potato crop yield raster with res. of 2x2 m²
Calculate spectral indices SUMVIS and NDRG
Calculate vegetation indices NDVI and WDVI
Calculate total SOM by multiplying SOM, soil depth and soil bulk density
Spectral indices bare soil with res. of 2x2 m²
Vegetation indices potato crops with res. of 2x2 m²
Total SOM with res. of 2x2 m²

Figure 3.4 - Flow chart data pre-processing, describing all the steps to create matching raster layers for all soil and crop variables at a spatial resolution of 2m x 2m

The following sections and chapters often refer to specific data formats and coordinate systems, from which the terminology is briefly explained in this section. Initial datasets were provided either as vector data or as raster data. Vector data embody spatial features as points, lines or polygons that are spatially related to each other, and are allowed to contain multiple attributes per geometric location (Chang 2015). On the other hand, raster datasets are gridded spatial datasets with a certain number of rows and columns containing cells with a particular size, which determines the dataset's spatial resolution (Chang 2015). In contrast to vector data, raster cells only contain one variable, indicated as the *cell values* of a raster dataset. For this research, vector data were available as data frames stored in CSV files and shapefiles, whereas raster datasets were provided as GeoTIFF format. A *Comma Separated Value* (CSV) file is a very basic textual data format in which data is often structured as a data frame (DF). A DF is a matrix with rows indicating observations, and columns representing variables that are allowed to have different data types such as text, numbers and dates (Shafranovich 2005). Secondly, a shapefile is a vector data format to store spatial features, linked to their geometric location by means of spatial coordinates (Bivand et al. 2008). One type of shapefile is a spatial points data frame (SPDF), which is a DF with two or three columns representing spatial coordinates for each point, and other columns containing values of one or more variables associated with those coordinates (Bivand et al. 2008). Lastly, the *Geometric Tagged Image File Format* (GeoTIFF) is a file format for storing raster images such as aerial photographs that can be embedded with metadata like a coordinate reference system or geometric projection (GDAL/OGR contributors 2018).

For all soil and crop datasets, a general pre-processing step was transforming them (if necessary) into another coordinate reference system. On the one hand, a common coordinate system is the *World Geodetic System* (WGS), used worldwide for mapping and navigation (Eurocontrol and IfEN 1998). A well-known example that makes use of WGS is the *Global Positioning System* (GPS). The currently applied WGS version is WGS84, and the latitudes and longitudes are expressed in degrees. On the other hand, the standard geodetic XY coordinate system for the Netherlands is *Rijksdriehoeksstelsel* (RD). The central point of this system is the steeple of the *Onze Lieve Vrouwetoren* in Amersfoort (de Bruijne et al. 2005). The presently used version of this reference system is RDNew, and longitudes (x) and latitudes (y) are expressed in meters. Because of its high accuracy for geospatial data about the Netherlands, this reference system was very suitable to use as a basis for the data of this research. In addition, some soil datasets were already available in RDNew coordinates. Therefore, the rest of the datasets were transformed into RDNew coordinates as well.

3.2.2 Field and fertilization boundaries

The Dutch government increasingly encourages a more open geo-information infrastructure, freely accessible for stakeholders that are interested in this information. For that purpose, many geo-data portals are available on the web, such as the *Nationaal Georegister* (NGR) (Kadaster 2018a) and its linked website *Publieke dienstverlening op de Kaart* (PDOK) (Kadaster 2018b). These portals provide a broad range of geo-information, such as the *Basisregistratie Gewaspercelen* (BRP). The BRP is a database containing boundaries and attributes (in vector data format) about the location of all registered agricultural parcels across the Netherlands and Flanders, Belgium. These attributes are linked to their corresponding farm and include each farmer's stocktaking of crop yield per season (De Vos et al. 2010). For this research, the main field boundary was extracted as a shapefile from the BRP database from 2015.

Since a few years, experiments have been conducted on a number of parcels at the farm of Van Den Borne, including the parcel that was studied in this research. For an experiment on comparing nitrogen (N) input with crop yield in 2011, the parcel was subdivided into four initial fertilization zones distributed across the parcel from north to south (Areda 2013). In 2015, a similar experiment was conducted based on the same fertilization zones, but with different N inputs. Moreover, additional sensor-based N fertilizer was applied on the whole field, except for a narrow strip of land perpendicular to the initial fertilization zones (Figure 4.1). (Van den Brande 2015). The shapefiles containing the fertilization zones were provided by Jacob Van Den Borne (Van den Borne 2018a). For this research, these datasets were used to compare them with the delineated management zones, and to investigate whether differences occurred in crop yield between specific fertilization zones. The methods for these analysis steps are explained in section 3.5.

3.2.3 Soil and crop images from UAVs

Aerial images of the parcel at Van Den Borne were recorded in 2015 as part of the research from Van Den Brande (Van den Brande 2015), and were also provided by Lammert Kooistra for the current research. The images were taken by means of a multispectral camera on board of an eBee (a UAV used by the company Aurea Imaging) at a resolution of 5 cm x 5 cm, and stored as GeoTIFFs that contained multiple spectral bands. The coordinate system of both GeoTIFFs was WGS84 UTM zone 31. One dataset was recorded on 14 April 2015, showing the field's bare soil, and other datasets were taken during the growing season. The last recording of that season was done on 15 August 2015, which was also used in this research in addition to the bare soil image from 14 April. The bare soil image was recorded in three spectral bands: *Red*, *Green*, *Blue* (RGB), containing 8-bit grayscale values ranging from 0 to 255, and the crop images were taken in four bands: *Green*, *Red*, *Red-Edge*, and *NIR*, having reflectance values ranging from 0 to 1. The value of 0 in both images indicated no reflection of electromagnetic radiation, whereas 255 and 1 represented 100% reflection, respectively.

In order to reduce the amount of pixels, and thus the memory size of the images, the first pre-processing step was masking the images ('cutting out' a piece) according to a rectangular polygon with its extents adopted from the field boundary. The next steps were aggregating the images to a resolution of 2m x 2m, and transforming them to RDNew coordinates. And finally, the images were masked a second time according to the field boundary itself. One study proposed calculating spectral indices of aerial images (Bartholomeus and Kooistra 2012), in order to highlight variations in reflectance values and to standardize differences in radiometry. For the current research, two spectral reflectance indices of bare soil were calculated and stored as two new raster images: the SUMVIS (sum of the three spectral bands) and the NDRG (Normalized Difference Red – Green). The equations to calculate both indices based on the three spectral bands of the bare soil image are as follows.

$$SUMVIS = \text{Red+Green+Blue} \qquad (3.1)$$

$$NDRG = \frac{\text{Red} - \text{Green}}{\text{Red} + \text{Green}} \qquad (3.2)$$

Based on the different spectral bands from the vegetation image that was recorded in August 2015, two vegetation indices were calculated and stored as two new images: the NDVI and WDVI (see also section 2.2.1). In fact, Aurea Image had already calculated these indices, which were included as two additional GeoTIFFs to the rest of the dataset. The equations to calculate both indices are as follows.

$$NDVI = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}} \qquad (3.3)$$

$$WDVI = \mathrm{NIR_{tot}} - \frac{\mathrm{NIR_{soil}}}{\mathrm{Red_{soil}}} \times \mathrm{Red_{tot}} \qquad (3.4)$$

In addition to the images containing spectral indices or vegetation indices, two other images were created from both GeoTIFFs. The bare soil image from April was used to create an image showing the field in *True Colour*, meaning that the field was depicted in its actual colours covering the RGB colour space of the image (Lillesand et al. 2008). On the other hand, the image of August was used to create an image showing the field in *False Colour*, meaning that the image was depicted in colours that differed from an actual True colour image, for instance to highlight a specific feature that was present in the image (Lillesand et al. 2008). In fact, the green band was depicted in blue, the red band in green and the NIR band in red. These images were only used to describe spatial patterns within the field (described in section 4.2.2), so they were not included in the procedure to delineate potential MZs.

3.2.4 Soil and crop measurements from manual sampling and ground sensors

In contrast to the raster-based aerial images discussed in the previous section, other datasets were available in the form of CSV files containing one pair of coordinates and one or more attributes representing different soil or crop yield variables. One of the variables was electric conductivity (EC) (in mS/m), measured on 20 March 2015 (before the start of cultivation and fertilization) with an EM38-MK2 sensor at soil depths of 0.5 m, 1.0 m, 1.5 m, and 3.0 m. Another variable was potato crop yield (in ton/ha), measured during harvest with a *Yield Master Pro* system on 4 October 2015. Datasets of both variables were provided for this research by Jacob Van Den Borne (Van den Borne 2018a). Three other variables were elevation, soil depth of the A-horizon, and soil organic matter, all measured during fieldwork at the farm of Van Den Borne on 20 March 2015 as well. Elevation was expressed in NAP, which is an abbreviation of *Normaal Amsterdams Peil*, used to represent the Dutch vertical coordinate reference system expressed in m above sea level (de Bruijne et al. 2005). The A-horizon is the top soil layer, which typically contains large amounts of humus and other organic matter, and is therefore very suitable for cultivating different kinds of crops such as potatoes and maize (Van den Brande 2015; Stoorvogel et al. 2015). Soil depth of the A-horizon was measured in m below ground level, and soil organic matter in percentages. Just as the aerial images discussed in the previous section, measurements of these three variables were acquired as part of the research of Van Den Brande (Van den Brande 2015), and were also provided by Lammert Kooistra for this research.

Except for elevation and soil depth that contained RDNew coordinates, all datasets contained WGS84 coordinates. Therefore, a first pre-processing step was to transform them into RDNew coordinates as well, as described in section 3.2.1. A second step was to check the point datasets for influential points, categorized as outliers or inliers. An outlier is an observation that is far outside the general pattern of values in a dataset, that potentially has influence on statistical analyses (Osborne and Overbay 2004). On the other hand, an inlier is an observation that significantly differs from its neighbouring values, but still falls within the general pattern of other values in a (spatial) dataset (Córdoba et al. 2016). One way to detect outliers was to calculate the mean of a variable and adding or subtracting three times the standard deviation. All values below or above these thresholds were considered as potential outliers. However, an important condition to be able to use this method is that data should approximately come from a normal distribution (Miller 1991; Osborne and Overbay 2004). Therefore, normality of all soil and crop yield variables was checked by creating boxplots. If a variable was approximately normally distributed, the method described above was used. However, if the boxplots showed a highly skewed distribution, an alternative method was used. The purpose of this method was to detect and remove inliers, based on Moran's local index I_i checking for spatial autocorrelation in a geographic dataset (Anselin 1995). This index was calculated for each observation and determined the degree of similarity compared to its neighbouring points. A positive value indicated similarity, whereas a negative I_i indicated unusually low or high values compared to its neighbouring points. To calculate the degree of autocorrelation based on the index, a spatial weighting matrix had to be calculated. Spatial weighs indicate the amount of interaction between observations in a spatial dataset (Dray, Legendre, and Peres-Neto 2006). For calculating the index, the pairs of coordinates were extracted from each point dataset, after which the neighbouring points per observation were identified based on a pre-defined Euclidian distance, and spatial weighs were calculated for each neighbouring point per observation. Potential inliers were visualized by creating so-called Moran scatterplots, with an x-axis containing observations, and the y-axis indicating the spatial lag of the variable (Anselin 1996). For this research, observations having a negative local Moran index, or containing statistically significant indices ($p < 0.05$) were removed from the datasets. Results of this procedure are described in section 4.1.2.

After detection and removal of influential points, the spatial point datasets were interpolated by means of ordinary kriging. Kriging is a geostatistical regression-based method to predict unknown values at output locations based on known values at input locations (Cressie 1990). The method uses (semi-)variograms as input information, which are functions describing the amount of spatial dependence of a random variable, and are expressed as relations between the distance and the variance of the difference between paired observations of that variable (Calder and Cressie 2009). The first pre-processing step with respect to spatial interpolation of the point datasets was to test and fit a number of empirical variograms for each variable in an automated way by the *AutofitVariogram* function in R (Hiemstra et al. 2009). This led to estimated variograms with parameters *model*, *nugget*, *sill*, and *range*. Only spherical and exponential models were tested, since those two types of models were suggested to be the most suitable for spatial interpolation of datasets used in plant and soil science (Gili 2013). The nugget represents all unexplained variance including measurement error, the sill indicates the maximum semi-variance between pairs of observations, and the range is the spatial distance at which the value of the model variogram reaches the maximum value (or sometimes 95%) of the sill (Calder and Cressie 2009). The *AutofitVariogram* algorithm iterates over all listed variogram models and selects the model having the smallest residual sum of squares compared to the sample variogram (Hiemstra et al. 2009), which was used in the end as input for the kriging function. Secondly, a reference raster dataset (one of the pre-processed aerial images) was transformed into a spatial points data frame with its coordinates serving as new prediction locations. Third, a kriging function was executed with these locations as input, together with the experimental variogram and the existing point locations of the dependent variable. The electric conductivity and crop yield datasets had a huge amount of observations, so in order to save processing time and memory, the amount of neighbours to include in the kriging predictions was fixed to a maximum value. Lastly, all interpolated points were transformed into raster datasets, and masked and projected to match the reference system, extents, and spatial resolution of the pre-processed aerial images discussed in section 3.2.3. Potato crop yield was only interpolated for assessing descriptive statistics, correlations, and spatial patterns.

One of the goals in the research of (Van den Brande 2015) was to establish relationships between soil properties and nitrogen (N) content in the top soil. Besides soil depth of the A-horizon and soil organic matter (SOM), another soil variable was Total SOM, which was also used as a variable in this research.

Total SOM was calculated by means of equation 3.5. In this equation, *Total SOM* is expressed in ton/ha, *SOM* in percentage, *soil depth* (of the A-horizon) in meters, and *soil bulk density* in g/cm^3 (or ton/m^3). The bulk density of the sandy soil at the farm of Van Den Borne was 1.13 g/cm^3 (Bakker 2014). The last number in the equation refers to the fact that one hectare is equal to 10,000 m^2.

$$Total\ SOM = \frac{SOM}{100\%} \times soil\ depth \times soil\ bulk\ density \times 10000 \qquad (3.5)$$

3.3 Descriptive statistics and mapping of datasets

Similar to data pre-processing, data analysis was performed in Rstudio as well. The data analysis steps applied on all pre-processed soil and potato crop variables are summarized in Figure 3.5, and will be further explained in this and the coming sections. First of all, descriptive statistics, scatterplots, and raster-based maps were created in order to get a better overview on the value ranges, spatial patterns, and interrelations between all covariates.

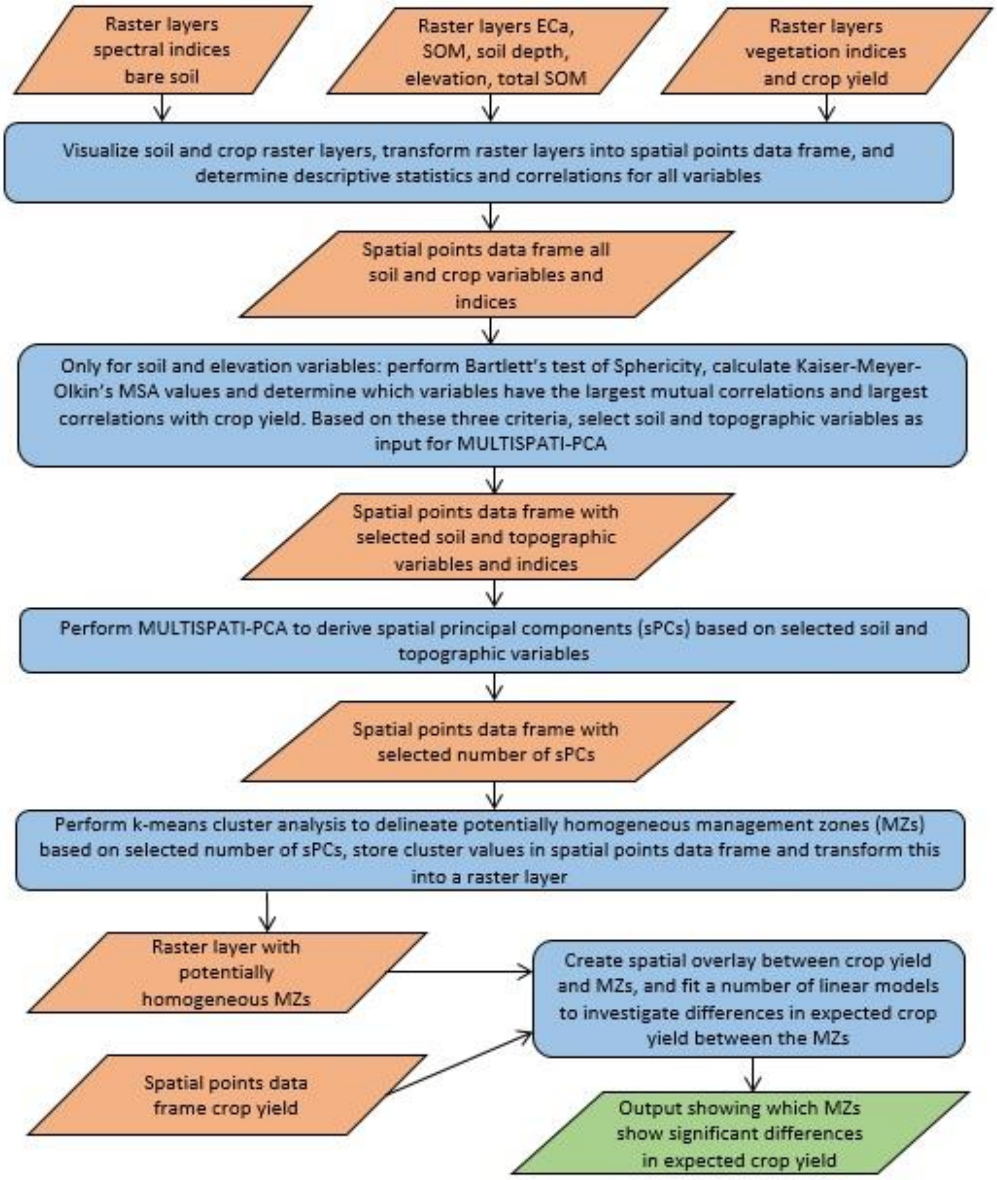

Raster layers spectral indices bare soil
Raster layers ECa, SOM, soil depth, elevation, total SOM
Raster layers vegetation indices and crop yield
Visualize soil and crop raster layers, transform raster layers into spatial points data frame, and determine descriptive statistics and correlations for all variables
Spatial points data frame all soil and crop variables and indices
Only for soil and elevation variables: perform Bartlett's test of Sphericity, calculate Kaiser-Meyer-Olkin's MSA values and determine which variables have the largest mutual correlations and largest correlations with crop yield. Based on these three criteria, select soil and topographic variables as input for MULTISPATI-PCA
Spatial points data frame with selected soil and topographic variables and indices
Perform MULTISPATI-PCA to derive spatial principal components (sPCs) based on selected soil and topographic variables
Spatial points data frame with selected number of sPCs
Perform k-means cluster analysis to delineate potentially homogeneous management zones (MZs) based on selected number of sPCs, store cluster values in spatial points data frame and transform this into a raster layer
Raster layer with potentially homogeneous MZs
Create spatial overlay between crop yield and MZs, and fit a number of linear models to investigate differences in expected crop yield between the MZs
Spatial points data frame crop yield
Output showing which MZs show significant differences in expected crop yield

Figure 3.5 - Flow chart data analysis, describing the steps to create descriptive statistics, maps, and spatial principal components, and in the end to delineate and validate potentially homogeneous MZs

Before the start of the actual data analysis, all pre-processed raster datasets were transformed into one large spatial point data frame (SPDF) with two columns indicating the pairs of RDNew coordinates, and the other columns representing the soil and crop variables associated with those coordinates. This SPDF was used as a basis to generate descriptive statistics, scatterplots and correlation matrices from those variables. For each variable, the sample size, minimum, maximum, mean, standard deviation, and coefficient of determination were calculated. For bivariate analysis, a correlation matrix and density scatterplots for each pair of variables were created. Two ways of describing correlation is by Pearson's correlation coefficient (*r*), and by including regression lines in scatterplots. Pearson's *r* ranges between -1 to 1 and explains how strong two variables correlate (Ott and Longnecker 2015). The closer the values approach those extremes, the larger the respective negative or positive association between the variables. A regression line depicts the goodness of fit between two variables. The closer the points in a scatterplot approach this line, the stronger the correlation between the two variables on both axes of the scatterplot (Ott and Longnecker 2015).

Visualization of all soil and crop datasets, but also from the derived PCs and MZs (explained in section 3.4) were performed in QGIS Desktop (version 2.18) (QGIS 2018). Among other (commercial) software such as ArcGIS (Environmental Systems Research Institute 2018), this open-source geographic information system (GIS) is widely used to compile, process, analyse, and visualize geographical data and information, to manage those data and information in geographical databases, and share it with many applications on the web (Burrough 1986; Chang 2015). For mapping all datasets, an aerial image from the study area was downloaded from Google maps (Google 2018) to use as a background layer in the maps. However, since the image was downloaded as a plain JPEG image, it lost its metadata concerning spatial resolution, coordinate system, and spatial extent, so georeferencing the image was necessary. Georeferencing is a procedure to (visually) relate and match the spatial coordinate system of one spatial dataset to another dataset with a different reference system, or to a new image without a coordinate system (Hackeloeer et al. 2014). This method was performed by deriving the coordinates of the location (in WGS84 format) from Google maps (Google 2018), converting them to RDNew coordinates, and projecting them on the JPEG image in the form of digital Ground Control Points (GCPs) in QGIS. Also the topographic map in Figure 3.3 was obtained and processed in that way.

3.4 Delineation of potential management zones

Delineation of potentially homogeneous management zones for the study field at Van Den Borne farm was performed in two steps. First of all by a special kind of principal component analysis called MULTISPATI-PCA, and secondly by a k-means cluster analysis based on the principal components that were derived from that analysis.

3.4.1 MULTISPATI Principal component analysis

Before conducting an actual principal component analysis, input variables were selected from the SPDF based on a number of assumptions. First, the correlation matrix (included in Appendix 5) was examined to check which soil and topographic variables had a noteworthy correlation with crop yield ($r \geq 0.3$), since crop yield was going to be used for validation of the derived management zones (explained in section 3.5). Second, Bartlett's test of Sphericity was executed to check if the soil variables were mutually related or not. This was done by testing the hypothesis that the correlation matrix was an identity matrix (a matrix with values of 1 on the diagonal, and other values being zero) or not, indicating that the variables were unrelated or related (Snedecor and Cochran 1989). A significant test ($p < 0.05$) would mean some degree of association between the variables, making them useful as input for PCA.

Third, a value called the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO-MSA) statistic was calculated, both for the complete dataset of soil and topographic variables, and for each of these variables separately. This statistic has often been applied in factor analysis, but may be used in PCA as well. It is a measure to investigate if the proportion of variance among all variables is caused by underlying factors or not (Hutcheson and Sofroniou 1999). A commonly used rule of thumb is that if KMO values ≥ 0.5, then variables are suitable as input for factor analysis or PCA, which was used as one of the assumptions for variable selection.

Principal component analysis (PCA) is an unsupervised method to transform possibly correlated variables from a dataset into a new set of uncorrelated variables, called principal components (PCs). One purpose is dimensionality reduction, achieved by only retaining the PCs that account for the largest part of variance from the complete dataset (Jolliffe 2014). The selected soil and topographic variables were used as input for MULTISPATI principal component analysis, which is a special kind of PCA that takes spatial information into account by calculating Moran's local index to detect spatial autocorrelation between observations, and calculates spatial principal components (sPCs) by maximizing the product of variance and spatial autocorrelation (Dray et al. 2008). The procedure to obtain sPCs from the soil and topographic variables was as follows. Ordinary PCA was performed on all selected soil and topographic variables to create a scree plot, which is a function showing the explained variance for each PC in descending order for the purpose of deciding how many components to retain in subsequent analysis (Jolliffe 2014). Next, ordinary PCA was performed on the selected variables a second time with the chosen number of retained PCs. The spatial coordinates were not used in the PCA itself, rather to calculate spatial autocorrelation based on Moran's index and to store those values in a weighing matrix, in the same way as was applied for the method to detect potential inliers described in section 3.2.4. Together with the duality diagram obtained from ordinary PCA on selected soil and crop variables, the spatial weighs were used as input for the MULTISPATI-PCA algorithm to generate spatial principal components (sPCs). Lastly, the sPC scores were linked to the pair of RDNew coordinates, after which each sPC was transformed into a raster file.

3.4.2 K-means clustering and smoothing of classification results

Management zones (MZs) were created by means of k-means clustering, which is an unsupervised classification method that allocates the values of given input variables into one of the *k* number of clusters, in such a way that the sum of squares within each cluster (WSS) is as low as possible (Hartigan and Wong 1979). The sPCs created from the MULTISPATI-PCA algorithm served as input variables. Similar to the procedure for PCA, the clustering algorithm was first performed to create a scree plot to decide how many clusters to retain. However, instead of the explained variance, this scree plot showed the total within sum of squares (WSS) for each given number of clusters in descending order. Next, the cluster algorithm was performed again based on the retained number of clusters, after which the cluster values were attached to the pair of RDNew coordinates and transformed into a raster file.

In addition, a paper suggested smoothing of classification results by means of non-linear spatial filters that respond according to the order of pixels in a particular area of an image (Arce 2005), in order to remove isolated pixels, small cluster patches, and sharp edges on boundaries between clusters (or MZs). A large range of filters with different matrix sizes ranging from 3 x 3 pixels to 35 x 35 pixels and different functions (such as *min*, *max*, *mean*, *median*, and *mode*) were explored, and the result that appeared most practical for management purposes was selected for further analysis.

3.5 Validation of potential management zones

Section 3.2.2 mentioned that fertilization experiments had taken place on the field, for which the parcel was subdivided in four fertilizer zones, and that additional fertilizer (N) had been applied on the whole field except for a vertical strip of land perpendicular to these fertilizer zones (Van den Brande 2015). Based on these fertilizer zones, three areas containing two or more delineated MZs were selected for validation of those MZs. The first area covered the whole field excluding the strip without additional N input, the second was located in the upper north-west corner of the field, and the third in the mid-west of the field (Figure 4.20). For validating the MZs, the spatial points data frame of crop yield was used instead of the interpolated raster, and the cluster values were attached to the points by means of a spatial overlay.

Validation of MZs was conducted by fitting a number of linear models representing the relationship between MZs and crop yield, in order to investigate whether significant differences occurred in crop yield between the delineated MZs. A book chapter by (Corwin and Lesch 2010) suggested that geostatistical mixed linear models (MLM) could be an effective way to map soil and plant properties. Moreover, (Schabenberger and Pierce 2001) proposed using MLMs for making statistical inferences about data in plant and soil sciences, since these types of models allow incorporating dependence and non-constant variance among observations, which is not the case for ordinary linear models (Ott and Longnecker 2015). The paper of (Córdoba et al. 2016) recommended to draw stratified random samples from the response variable to use in MLMs, with MZs serving as strata. Therefore, from each MZ a random sample of 100 crop yield point observations was drawn for each of the three validation areas to use in the statistical models. Next, four mixed linear models in addition to a one-way ANOVA were fitted on the random samples from each of the three described areas, with clusters (MZs) as explanatory variable and crop yield as response variable. Similar to the procedure in the paper by (Córdoba et al. 2016), the four MLMs were modelled with spatially correlated error terms based on spherical and exponential functions, both with and without estimated nugget effect, while the one-way ANOVA was modelled with independent error terms. Only MLMs with spherical and exponential correlations were evaluated, since these were considered most suitable for statistical modelling in plant and soil sciences (Gili 2013). For each of the three areas, the total set of five models was evaluated by means of Akaike's information criterion (AIC), which is a method to compare the relative qualities for a number of fitted statistical models, in which quality is defined by the inclusiveness of information in a given model (Akaike 1981). For each of the three validation areas, the statistical model having the lowest AIC was considered to be of best quality, and was selected to make further statistical inferences about. For each of the three models, two hypotheses were tested. The first hypothesis tested the overall performance of the model to evaluate whether any difference occurred in expected crop yield between the delineated MZs, by means of an F-test at a significance level of $\alpha = 0.05$:

H0: Expected crop yield is the same among all MZs;
Ha: At least two of the MZs show significant differences in expected crop yield.

For pairwise comparisons, the second hypothesis tested whether a significant difference occurred in expected crop yield between each pair of MZs, by means of t-tests based on Tukey's HSD at a significance level of $\alpha = 0.05$:

H0: No significant difference occurs in expected crop yield between two MZs;
Ha: The difference in expected crop yield between two MZs is significant.

4. Results

This chapter discusses the results of this research, starting with the data pre-processing results, followed by an interpretation of the descriptive statistics, correlations and output maps of all used data, the results of the management zone delineation based on PCA and cluster analysis, and the validation of the potential management zones.

4.1 Pre-processed datasets

This section starts with a description of the designated fertilizer boundaries. Next, the results regarding the detection and removal of inliers for the spatial point datasets and spatial interpolation by means of kriging are explained.

4.1.1 Fertilizer boundaries

For the fertilization experiment that took place in 2015, four zones with different fertilizer (N) levels were assigned on the field. As Figure 4.1 indicates, these N inputs were 90 kg/ha, 0 kg/ha, 162 kg/ha, and 252 kg/ha ranging from north to south (Van den Brande 2015). These N levels were applied before the start of the growing season. Besides this, the field was treated with additional fertilizer during the growing season, except for one vertical strip of land perpendicular to the initial fertilizer zones (in Figure 4.1 indicated in blue). For this research, these fertilizer zones were used as a basis to select areas as input for linear models to validate delineated management zones (explained in section 3.5).

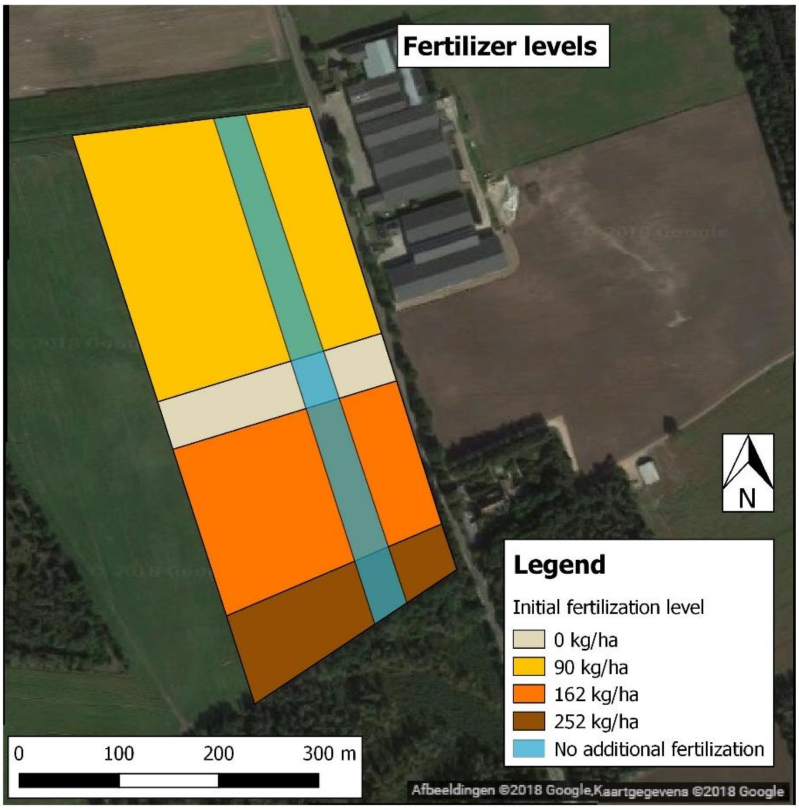

Fertilizer levels
N
Legend
Initial fertilization level
0 kg/ha
90 kg/ha
162 kg/ha
252 kg/ha
No additional fertilization
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

Figure 4.1 – Fertilization map potato field, with the four initial N inputs 90 kg/ha, 0 kg/ha, 162 kg/ha, and 252 kg/ha, and a vertical strip without additional fertilizer, adopted from (Van den Brande 2015)

4.1.2 Detection and removal of influential points

As explained in section 3.2.4, some of the data were available as spatial point datasets. After transforming these datasets to RDNew coordinates and removing missing values (if present), boxplots were examined to check normality of the data and potential influential points (outliers or inliers). Elevation, soil depth A-horizon, and soil organic matter (SOM) all appeared to be normally distributed, except for one or two potential outliers (Figure 4.2 or Appendix 2), so for those data the method to detect and remove outliers was applied based on values outside the range of 3 standard deviations away from the mean. Appendix 1 includes the initial number of spatial point observations of these datasets. For elevation 1 out of 133 observations was removed, for soil depth 2 out of 133, and for SOM none of the 25 observations were removed.

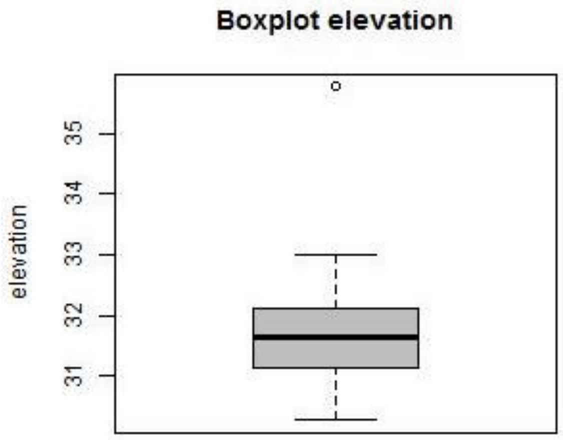


Figure 4.2 – Boxplot elevation. Data looks normally distributed, apart from one potential outlier

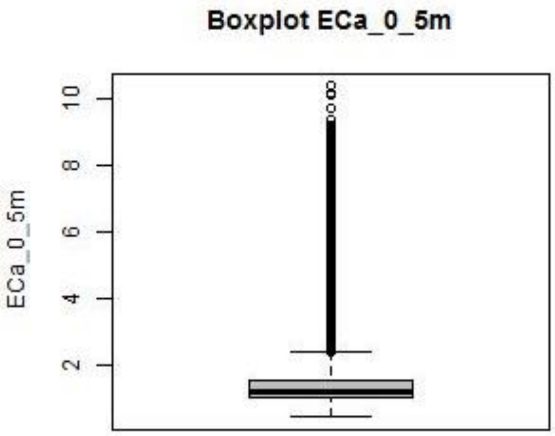


Figure 4.3 – Boxplot ECa at 0-0.5 m soil depth. Data does not appear to be normally distributed and potential outliers are clearly visible

Electric conductivity at all four soil depths and crop yield did not show normal distributions (Figure 4.3 or Appendix 2), so for those datasets the procedure to detect and remove inliers was conducted. To detect potential inliers based on Moran's local index, neighbouring points were identified for each observation at a Euclidian distance of 2.5 meters for EC and 2 meters for crop yield, and given a spatial weight. Potential inliers were visualized by Moran scatterplots, represented as black dots as shown in Figure 4.4, Figure 4.5, and Appendix 3. Apart from Moran's local index, these points were also based on a number of other diagnostic statistics, that were not considered for removal of inliers in this research. The slope of the line in the scatterplot is equal to Moran's overall *I* index, which is similar to Moran's local index, yet based on the sum of all weighted observations (Anselin 1996). The upper-left corner of the plot relative to the dashed lines indicates low values surrounded by high values, whereas the lower-right corner indicates high values surrounded by low values. Moreover, the more scattered the observations are, the greater the potential influence. For instance, Figure 4.4 and Figure 4.5 indicate that crop yield contained stronger influential points than EC. In the end, about 0.7% of observations from the four EC datasets were removed (with almost 30,000 observations remaining), and 12% of observations from crop yield (with more than 26,000 observations remaining).

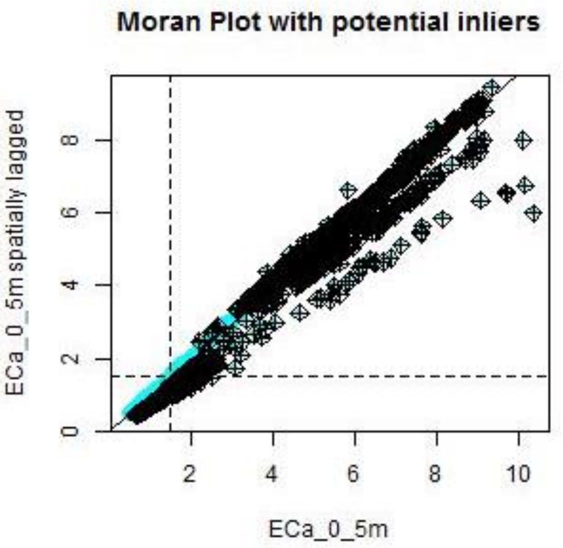


Figure 4.4 – Moran plot ECa at 0-0.5 m soil depth. Points follow straight line, but many potential inliers are visible (represented as black points)

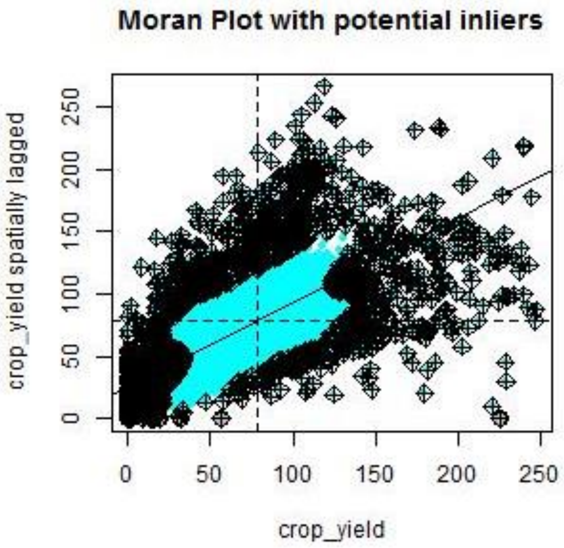


Figure 4.5 – Moran plot crop yield. Points are scattered, and many potential inliers are visible (represented as black points)

4.1.3 Spatial interpolation by means of kriging

As explained in section 3.2.4, an experimental variogram was fitted for each spatial point variable by means of the *AutofitVariogram* function in R (Hiemstra et al. 2009). Output was generated in the form of semi-variograms, in addition to a table containing the parameters *model*, *nugget*, *sill*, and *range* for each variable (Appendix 4). Most of the variograms were fitted as exponential models, except for elevation and crop yield that were fitted as spherical models. The minimum nugget was zero (all ECa variables), whereas the maximum nugget was 254.3 (crop yield). Regarding the sill, the minimum value was 0.006 (soil depth), and the maximum sill was 555.6 (crop yield). Lastly, the minimum range was 14.8 (soil depth), while the maximum range was 1654.1 (elevation).

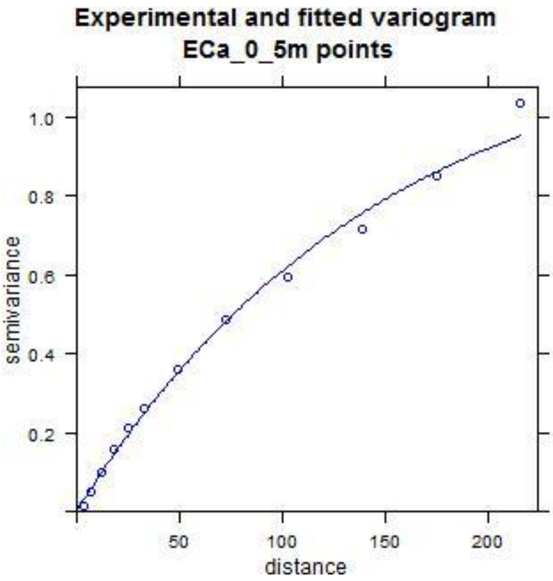


Figure 4.6 – Experimental (dots) and fitted variogram (line) ECa at 0-0.5 m soil depth

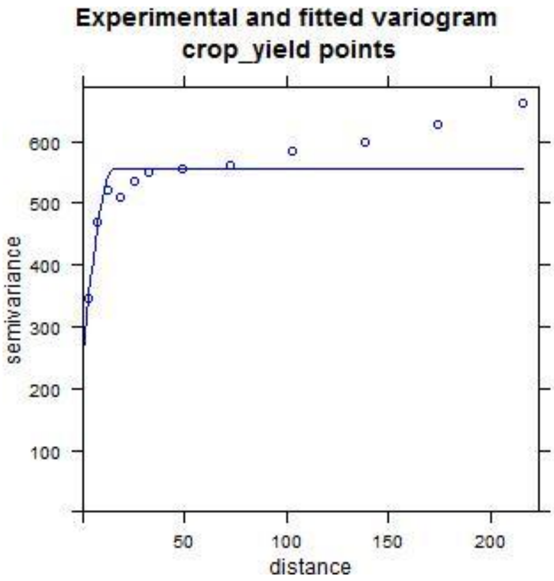


Figure 4.7 – Experimental (dots) and fitted variogram (line) crop yield

Two semi-variogram plots are shown in Figure 4.6 (EC at 0-0.5 meter soil depth) and Figure 4.7 (crop yield). The range of the EC variogram is relatively large compared to the range of the crop yield variogram, which corresponds to the fact that the EC variogram gradually increases, while the crop yield variogram has a steep increase in the beginning. In addition, the plots also display the respective nuggets of 0 and 254.3 and sills of 1.24 and 555.6 for EC and for crop yield.

4.2 Descriptive statistics and visual assessment of datasets

This section includes interpretations on the descriptive statistics, correlations and visualizations of all soil and crop variables. Descriptive statistics of these variables are summarized in Table 4.1. The number of observations (n) represents the amount of pixels in each raster dataset that were transformed into a SPDF. The table shows the same amount for all variables, which is one indication that the extents and resolutions of all datasets were properly matched with each other.

Table 4.1 – Descriptive statistics of all soil and crop variables, including number of observations in SPDF (n), minimum, mean, maximum, standard deviation and coefficient of variation

Variable	n	min	mean	max	sd	cv
Aerial image bare soil						
Bare soil Red (-)	31790	83.937	142.274	179.077	6.963	0.049
Bare soil Green (-)	31790	81.899	128.084	169.233	7.938	0.062
Bare soil Blue (-)	31790	67.430	119.297	156.759	6.623	0.056
Bare soil NDRG (-)	31790	-0.058	0.053	0.077	0.015	0.276
Bare soil SUMVIS (-)	31790	244.477	389.655	499.822	19.847	0.051
Soil parameters and elevation						
Eca 0-0.5 m (mS/m)	31790	0.437	1.592	10.503	1.272	0.799
Eca 0-1.0 m (mS/m)	31790	1.679	3.369	15.630	1.660	0.493
Eca 0-1.5 m (mS/m)	31790	3.343	5.191	16.584	1.626	0.313
Eca 0-3.0 m (mS/m)	31790	4.283	6.861	16.140	1.510	0.220
Elevation (m + NAP)	31790	30.296	31.584	32.983	0.622	0.020
Soil depth (m - gr. lev.)	31790	0.274	0.425	0.566	0.042	0.100
Soil OM (%)	31790	4.095	4.356	4.758	0.170	0.039
Total soil OM (ton/ha)	31790	129.519	209.611	300.564	24.112	0.115
Aerial image potato crops						
Crop yield (ton/ha)	31790	0.036	75.858	151.792	18.702	0.247
Crop NDVI (-)	31790	0.289	0.838	0.919	0.053	0.063
Crop WDVI (-)	31790	0.013	0.258	0.513	0.067	0.260

4.2.1 Descriptive statistics and correlations

The aerial image of the bare field taken in April 2015 is represented by the three individual RGB bands and spectral indices NDRG and SUMVIS. Table 4.1 indicates that the RGB values extend between 67.430 and 179.077, which is right in the middle of the expected range between 0 and 255 (belonging to an 8-bit colour space) that each band can take. Equation 3.2 specifies that the SUMVIS index is the sum of the three RGB bands, from which observed values range between 244.477 and 499.822, which is also right in the middle of its expected value range (between 0 and 765). The mean of the Red band is a little higher than the means of the two other colour bands. Since the coefficient of variation (CV) is a ratio between the standard deviation and the mean, this leads to a lower CV. In addition, because the NDRG is a ratio between the Red and Green bands (calculated with equation 3.2), expected values could range from -1 to 1. However, the range of NDRG values is very close to zero (with numbers between -0.058 and 0.077), which is caused by the fact that the values of the Red and Green bands are so near to each other.

The correlation matrix in Appendix 5 shows (very) strong mutual correlations ranging from 0.7 to 0.95 between the three RGB bands and the SUMVIS, so these bands and index appear to be closely related to each other. However, the correlations with other indices and variables appear to be (much) lower (Appendix 5). For instance, the correlation between the SUMVIS and NDRG index is only -0.355 (Figure 4.8), while the correlations with soil variables are sometimes even lower, ranging from -0.01 to 0.33. Also the correlations with potato crop yield are not that high; the correlation between for instance the SUMVIS and crop yield is only -0.16. In general, the NDRG index appears to perform better regarding its correlations with other soil and crop variables. For instance, the correlation between NDRG and total soil organic matter is -0.341 (Figure 4.9), and the correlation between NDRG and EC at 0.5 m depth is 0.3. Also the correlations with crop variables are higher than the ones related to the SUMVIS. For example, the correlation between NDRG and crop yield is 0.31.

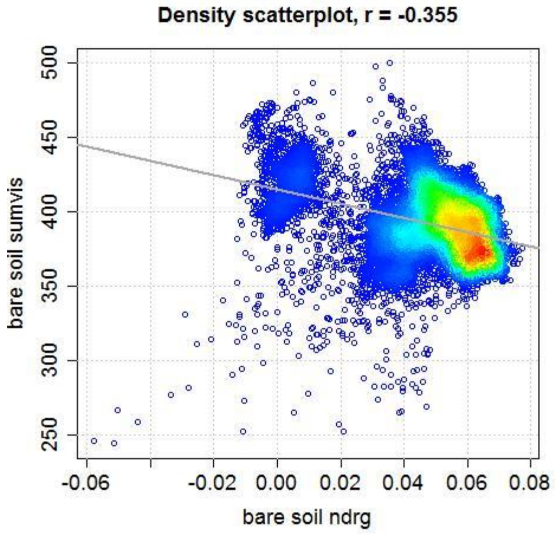


Figure 4.8 – Density scatterplot bare soil NDRG vs. bare soil SUMVIS with a weak negative correlation of r = -0.355

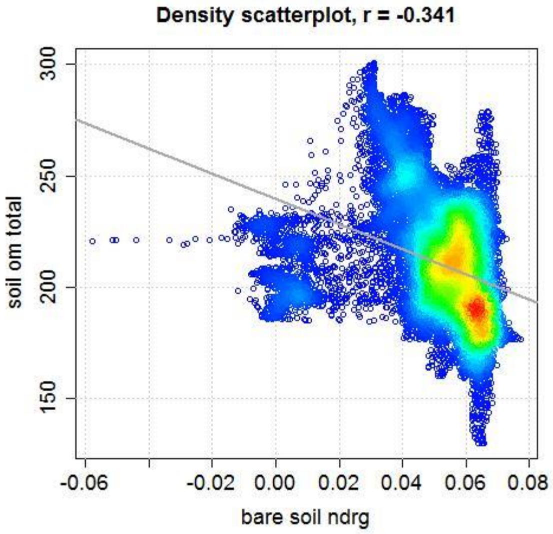


Figure 4.9 – Density scatterplot bare soil NDRG vs. total soil organic matter with a weak negative correlation of r = -0.341

In addition to a correlation matrix, density scatterplots such as shown in Figure 4.8 and Figure 4.9 give an even more complete overview of the relations between pairs of variables. This is because the patterns of the individual observations are visible, and regression lines are included in the plots as well. Figure 4.8 shows a decreasing pattern for SUMVIS values for increasing NDRG values. Additionally, the point density gets higher for larger values of NDRG, indicating that the NDRG contains many pixel values ranging between approximately 0.04 and 0.08 compared to the SUMVIS. A similar pattern is visible in Figure 4.9, although this plot contains less isolated observations than the plot in Figure 4.8. Moreover, observations of soil organic matter are vertically more spread out than observations of the SUMVIS index.

Apparent electric conductivity (ECa) was measured at soil depths of 0.5m, 1.0m, 1.5m and 3.0m. Table 4.1 shows ECa values ranging between 0.437 mS/m and 16.584 mS/m across all depths, but appear to increase when soil depth increases. For instance, the mean ECa at 0.5m depth is 1.592 mS/m, while the mean at 3.0m depth is 6.861 mS/m. These values correspond with expected EC values ranging between 0-10 mS/m for sandy soil, and EC larger than 10 mS/m for more loamy soil types (Geonics Ltd. 1980). Elevation ranges smoothly from 30.296 m to 32.983 m, while the soil depth of the A-horizon has values between 0.274 m and 0.566 m. Interpolated values of SOM extend between 4.095% and

4.758%. Total soil organic matter was calculated with help of equation 3.5, in which soil depth was multiplied by the fraction of SOM and bulk density of the soil. As a result, values of total soil organic matter range between 129.519 ton/ha and 300.564 ton/ha.

A shown in Appendix 5, mutual correlations between the ECa variables at the different soil depths are very strong, and extend between 0.91 and 0.96. Also correlations between EC and other soil variables are moderate to strong. For instance, the correlation between ECa at 0.5 meters depth and elevation equals -0.524 (Figure 4.10), and the correlation with total soil organic matter is 0.555 as shown in Figure 4.11. The correlations between ECa and potato crop variables are also moderate, as demonstrated for example by a correlation of -0.4 between ECa at 3.0m depth and crop yield. The correlations between elevation and other variables (except the WDVI) are also moderate. For instance, the correlations with soil depth, soil OM and crop yield are -0.31, 0.3 and 0.33, respectively. Also soil depth and total SOM show noteworthy, and sometimes even strong correlations between each other and with other variables; the correlation between for example soil depth and total SOM is 0.94, and the correlation between soil depth and crop yield is -0.24.

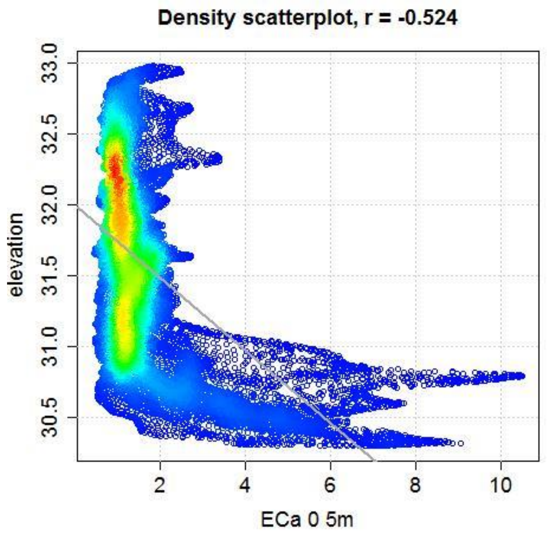


Figure 4.10 – Density scatterplot ECa at 0-0.5 m soil depth vs. elevation with a moderate negative correlation of r = -0.524

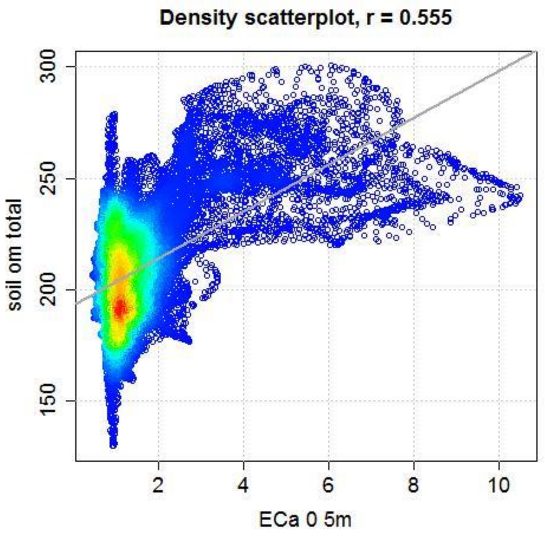


Figure 4.11 – Density scatterplot ECa at 0-0.5 m soil depth vs. total soil organic matter with a moderate positive correlation of r = 0.555

The scatterplots in Figure 4.10 and Figure 4.11 depict the relations between ECa at a soil depth of 0.5 meters, and elevation and total SOM. The figure on the left shows a negative association, with generally speaking lower elevation values for higher ECa values. However, a large point density is visible on the left part of the plot, indicating that a relatively small range of ECa values is present at the field with a large range of elevation values. A similar pattern is visible in the scatterplot on the right, with the difference that it is vertically flipped compared to the other plot, depicting a positive relation between ECa and total SOM.

Three potato crop variables were included in this research: crop yield that was measured during harvest in October 2015, and NDVI and WDVI that were two vegetation indices derived from the aerial image of the field recorded in August 2015. Table 4.1 indicates that interpolated crop yield values range from 0.036 to 151.792 ton/ha, with a mean of 75.858 and standard deviation of 18.702. The NDVI and WDVI were calculated with equations 3.3 and 3.4 respectively, and since they are ratios between the NIR and Red bands, their expected values range between -1 and 1. In fact, the NDVI ranges from 0.289 to 0.919, with a mean of 0.838 and standard deviation of 0.053, while the WDVI extends between 0.013 and 0.513, with a mean of 0.258 and standard deviation of 0.067. However, the WDVI has a lower coefficient of variation, because its mean is much lower than the mean of the NDVI.

As the correlation matrix in Appendix 5 indicates, mutual correlations between crop yield, NDVI and WDVI are moderate till good. For example, the correlation between crop yield and WDVI is 0.429 (Figure 4.12), and the correlation between NDVI and WDVI is 0.66. This plot shows a positive relation, but with a very scattered pattern of observations. Density increases towards the center of the plot, so many intermediate crop yield values correspond to many intermediate WDVI values. Figure 4.13 shows a decreasing relation between ECa at 0.5 meters soil depth and crop yield, with a high point density cloud on the left part of the plot, indicating many ECa pixels that are lower than 2 mS/m are present for a broader range of crop yield values.

Finally, as described in the previous sections, correlations between crop yield and other spectral indices and soil variables show a varying range of magnitudes. Some correlations are moderate, such as the one between ECa at 0.5 m soil depth and crop yield (r = -0.391 as shown in Figure 4.13), while others are very weak, such as the correlation between the Red band of the bare soil image and crop yield, which equals -0.07. The strength of correlations between potato crop yield and other variables is one of the considerations for deciding which variables to include and which ones to exclude in the principal component analysis described in section 4.3.1.

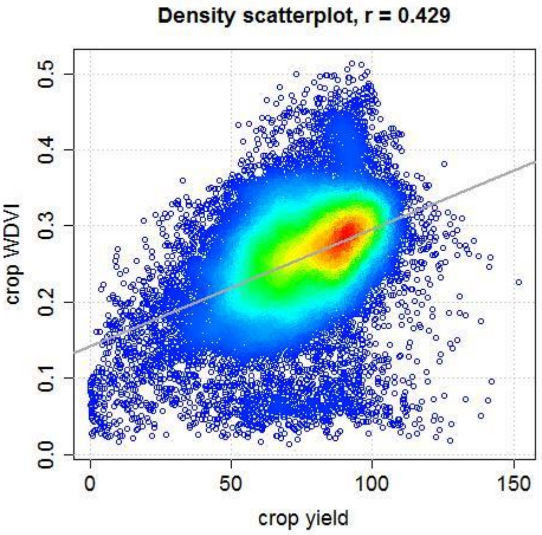


Figure 4.12 – Density scatterplot crop yield vs. crop WDVI with a moderate positive correlation of r = 0.429

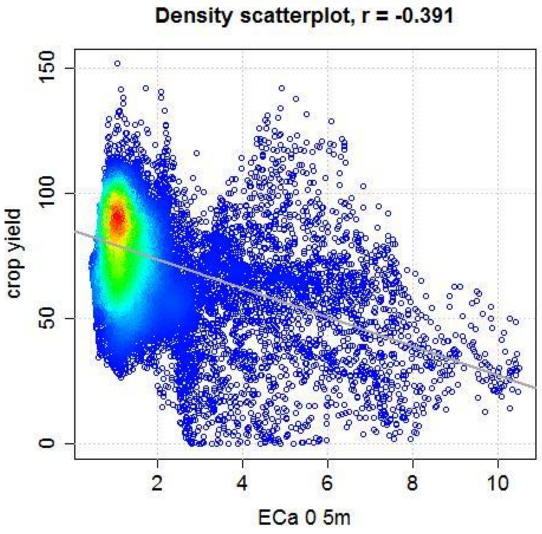


Figure 4.13 – Density scatterplot ECa at 0-0.5 m soil depth vs. crop yield with a weak negative correlation of r = -0.391

4.2.2 Visual assessment of soil and potato crop maps

The True Colour image of the field depicting bare soil (map A in Appendix 10) shows a gradual darkening pattern ranging from south to north, with some dark patches of land spread across the field. Another notable feature is the green strip of land bordering with the field in the west. Maps of the three spectral RGB bands, as well as the map depicting the SUMVIS index (map C in Appendix 10) show very low pixel values at the north part of the field, which is also the case for a large patch of land ranging from the west to the center of the field. Values of surrounding pixels gradually increase, and the south and east parts of the field show relatively high values compared to the other areas. On the contrary, the NDRG map (map D in Appendix 10) shows high values in a large part of the field except for the areas on the edges of the field. Especially the headland in the north of the field has very low NDRG pixel values. In addition, on all of these maps except the map of the Blue band, the same deviating vertical strip in the west of the field is visible that was also observed on the True colour image.

The four maps depicting ECa at different soil depths, such as the one in Figure 4.14 show all a very similar pattern, although the values increase with increasing soil depth. The patches of land in the south show low EC values, while values increase more to the north of the field. Especially the headland in the upper north shows very high EC values. For ECa at 3.0 meters soil depth, the contrast between pixels in the south and north parts of the field is even larger. In addition, on all ECa maps, a similar seemingly contradicting vertical strip of land is visible in the west of the field, which was also the case for the maps of the spectral bands and indices from the aerial image of bare soil. The soil organic matter map (map F in Appendix 10) shows high SOM values at the edges of the field, with a gradual decrease towards the upper-mid east part of the field containing a patch of land with low SOM values. The map depicting the soil depth of the A-horizon shows a very scattered pattern. Patches of land with high or low values alternate between each other, although the headland in the north of the field contains relatively large soil depths compared to the rest of the field. The calculated total SOM map (map G in Appendix 10) shows a combination of patterns visible on both the SOM and soil depth map, with slightly lower values on the areas where SOM is also lower. Lastly, the elevation map (map H in Appendix 10) shows a smooth pattern with gradually decreasing values towards the headland in the north of the field.

The potato crop yield map (Figure 4.15) shows high yield values in the south, but in the center and north parts of the field, some areas with (very) low crop yields are visible, partly due to the fertilization experiment that was performed in 2015. Especially the headland in the north suffers from very low yield values. This pattern is also visible on the NDVI and WDVI maps (maps I and J in Appendix 10) However, the WDVI shows relatively high pixel values in the center of the field where crop yield shows lower pixel values. Another notable feature on the WDVI and NDVI maps is the vertical strip of land in the mid-east part of the field with relatively low values compared to other parts of the land, which is to a lesser extent also visible on the crop yield map (Figure 4.15). The False Colour image (map B in Appendix 10) shows a very similar pattern to the NDVI and WDVI maps. Moreover, since vegetation reflects a large portion of NIR radiation, and because the NIR band is visualized in Red, pixel values on the False Colour map show a dominant red colour. Lastly, the tramlines (or tractor paths) are clearly visible in that image and other potato crop maps as well, indicated by contours with low pixel values located between vertical strips of land containing high pixel values.

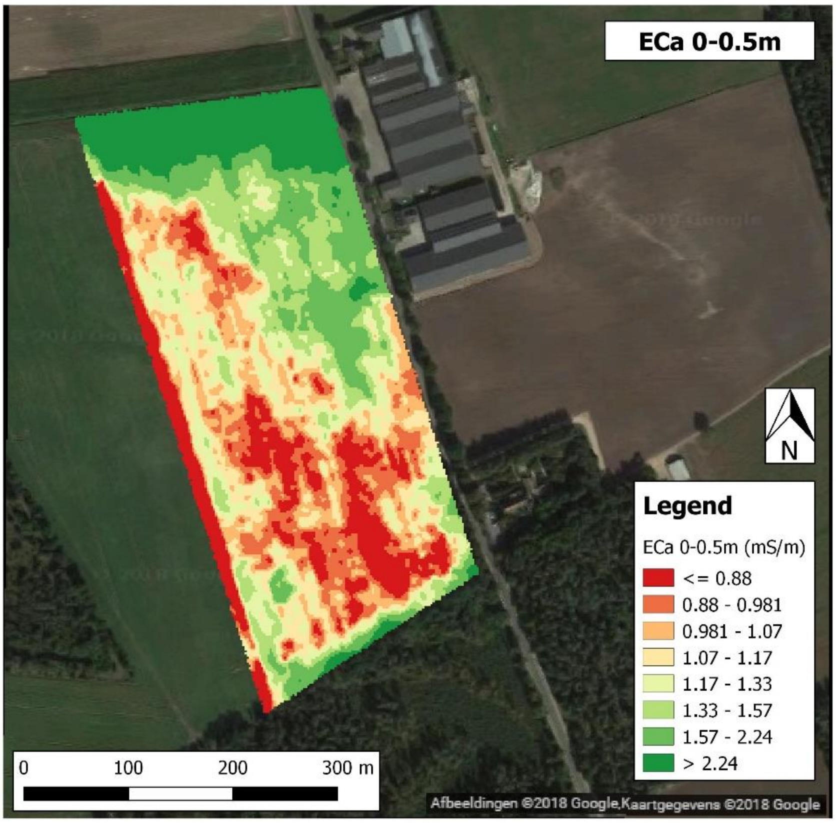


Figure 4.14 – Map of electric conductivity (mS/m) at 0-0.5 m soil depth

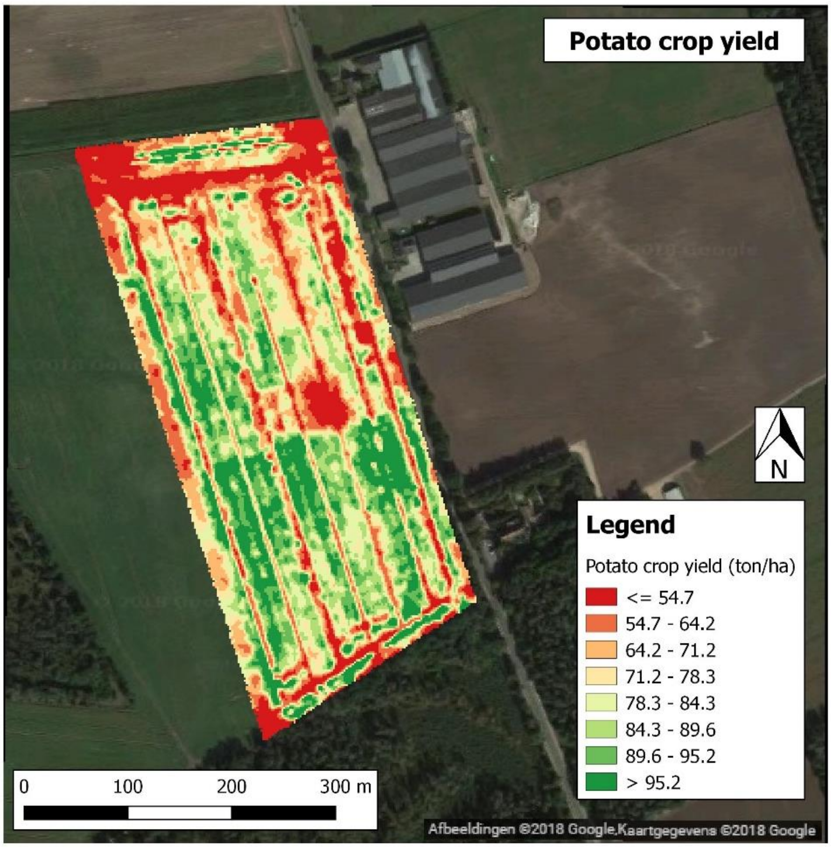


Figure 4.15 – Map of potato crop yield (ton/ha)

4.3 Delineation of potential management zones

This section describes the results regarding the delineation of potentially homogeneous management zones. First the MULTISPATI-PCA output is interpreted, after which the clustering results are explained.

4.3.1 Principal components obtained from MULTISPATI–PCA

Soil and topographic variables were selected based on three assumptions. First of all, mutual correlations, and correlations with crop yield were evaluated (explained in section 4.2.1). Based on this information, bare soil NDRG, the ECa variables at all soil depths, and elevation were initially selected as input for PCA, since their absolute correlations with crop yield were higher than 0.3 (which was considered moderate or better). Second, Bartlett's test of Sphericity was performed to check if the variables had some degree of association or not. As the output in Appendix 6 indicates, the chi-square value of this tests equals 1.875 x 10^6 and its p-value equals 0.000, which is statistically significant meaning that the soil and topographic variables are suitable as input for PCA. Third, Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO-MSA) statistics were calculated, both overall and for each variable separately. Initially, the overall KMO was 0.607, but the KMO values of five variables including the RGB colour bands of the bare soil image, the SUMVIS, and soil organic matter (SOM) were below the critical value of 0.5 (Appendix 6), making them potentially unsuitable for PCA. After eliminating these five variables and calculating the KMO values again, the overall KMO increased to 0.766, and all individual KMO values above the threshold of 0.5 (Appendix 6). This information led to the same variables that were selected from the correlation analysis, in addition to soil depth A-horizon and total SOM. In the end, these 8 variables (Table 4.2) were selected as input for the MULTISPATI-PCA algorithm.

As explained in section 3.4.1, the MULTISPATI-PCA algorithm is a special kind of PCA that incorporates spatial information in its calculations (Dray et al. 2008), and was performed with the selected soil and topographic variables as input. Spatial principal components (sPCs) were determined by maximizing the product of explained variance and spatial autocorrelation that was present across the selected variables. Based on the PCA scree plot (Appendix 7), four sPCs were chosen to retain for further analysis, explaining 94.08% of the total variance in the dataset with soil and elevation variables. Table 4.2 shows the communalities and loadings of these four sPCs. The PCA loadings suggest that the first sPC mostly represents the ECa variables at all four soil depths, the second sPC is dominated by soil depth A-horizon and total SOM, the third sPC by bare soil NDRG, and the fourth sPC is mostly represented by elevation. The communalities range from 0.213 to 0.993. The ECa variables contain such low communalities due to very low loadings in sPC2 till sPC4, but relatively high loadings in sPC1, while bare soil NDRG and elevation contain high communalities because of very high loadings in sPC3 and sPC4, respectively.

Table 4.2 – Communalities and PCA loadings of the 4 retained sPCs. The numbers in blue/bold represent variables with the highest communalities or variables with the highest loadings for each sPC

Variable	Communalities of first 4 sPCs	sPC1 (64.46%)	sPC2 (14.76%)	sPC3 (8.24%)	sPC4 (6.62%)
Bare soil NDRG	**0.993**	0.170	-0.203	**0.950**	0.139
Eca 0-0.5 m	0.315	**-0.408**	-0.135	-0.009	0.361
Eca 0-1.0 m	0.269	**-0.419**	-0.157	0.017	0.262
Eca 0-1.5 m	0.242	**-0.423**	-0.168	0.010	0.188
Eca 0-3.0 m	0.213	**-0.418**	-0.183	0.025	0.061
Elevation	**0.929**	0.286	0.435	-0.062	**0.809**
Soil depth	0.553	-0.311	**0.544**	0.274	-0.292
Total soil OM	0.486	-0.311	**0.608**	0.129	-0.053

PCA biplots as shown in Figure 4.16 and Figure 4.17 give an even better overview about the loadings and mutual relationships between variables. Figure 4.16 indicates that all four ECa variables are strongly correlated, which is also the case for soil depth and total SOM. It also shows that ECa is negatively correlated with elevation, and that soil depth and total SOM are both negatively correlated with bare soil NDRG. This was also observed from the correlation matrix in Appendix 5. Moreover, ECa is the most important variable to explain spatial variability in the first sPC (negatively correlated) and most of the spatial variability in sPC3 is explained by bare soil NDRG (positively correlated), while soil depth and total soil OM are the second most important for these two axes, but most important in explaining spatial variability in sPC2 (all positively correlated). Lastly, elevation explains most of the spatial variability in sPC4. All these patterns were also observed in Table 4.2. A complete overview of biplots depicting loadings and correlations between the pairs of all 4 sPCs is included in Appendix 8.

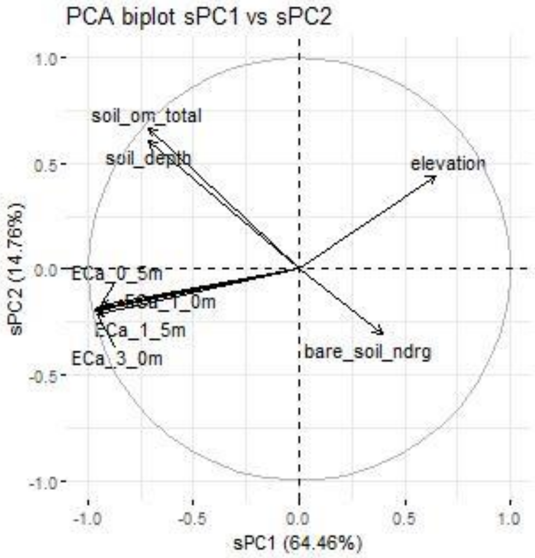


Figure 4.16 – PCA biplot sPC1 (64.46%) and sPC2 (14.76%)

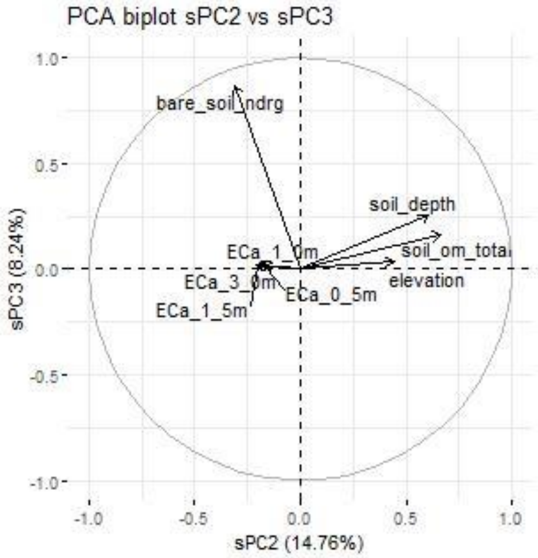


Figure 4.17 – PCA biplot sPC2 (14.76%) and sPC3 (8.24%)

The four sPCs were visualized and depicted as ranges of PC scores distributed across the potato field. For instance, the map of sPC1 shown in Figure 4.18 shows very low PC scores in the headland in the north of the field in addition to a patch of land with relatively low scores in the east of the field and high PC scores in the center and south of the field. This corresponds quite well with the ECa patterns observed in Figure 4.14, especially since ECa variables are strongly and negatively correlated with sPC1. The spatial distribution of sPC2 is visualized in Figure 4.19, which has patches of land with low scores in the north and center of the field, and high sPC scores in the south of the field. This shows very strong similarities with the patterns observed for soil depth and Total SOM such as shown in maps E and G of Appendix 10. The map of sPC3 (map A in Appendix 11) shows relatively high scores in the center of the field, while the edges contain very low sPC scores. This corresponds quite well with the NDRG patterns observed in map D of Appendix 10, although soil depth also seems to have a small influence on sPC3. Moreover, the seemingly deviating strip of land in the west of the field is also visible in the map of sPC3, similar to what was observed in the NDRG map. Lastly, the map of sPC4 (map B in Appendix 11) shows very low sPC scores in the north-west of the field, and higher scores towards the south-east of the field. This shows a lot of similarities with the distribution of elevation across the field (especially because elevation is strongly correlated with sPC4), although also some influence of ECa on sPC4 seems to be present in the northern part of the field.

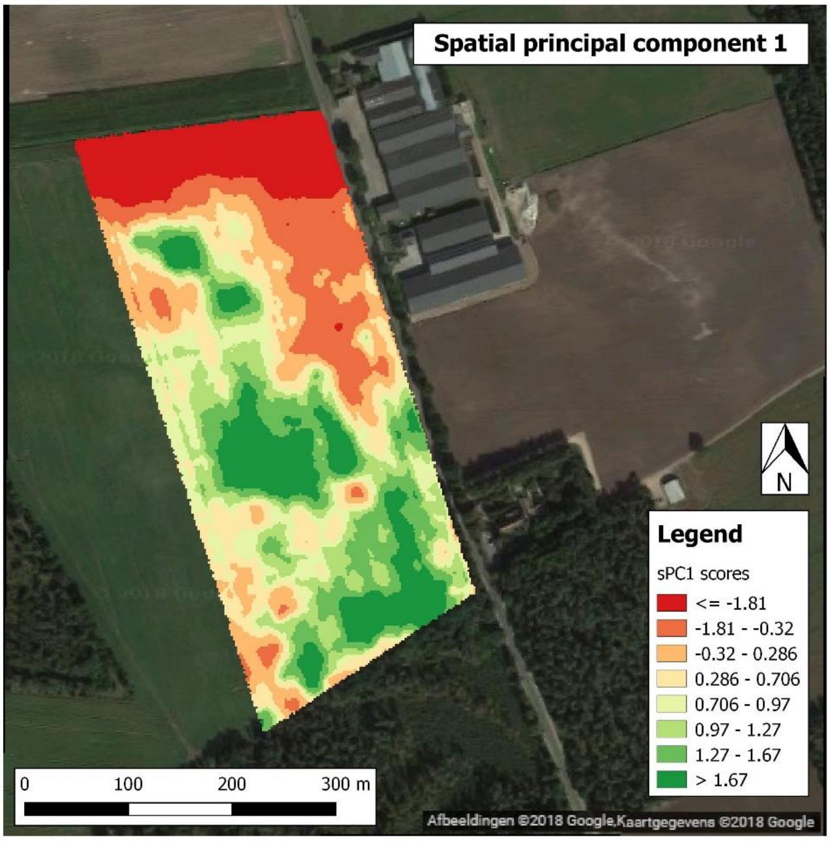


Figure 4.18 – Map of spatial principal component 1 including sPC scores

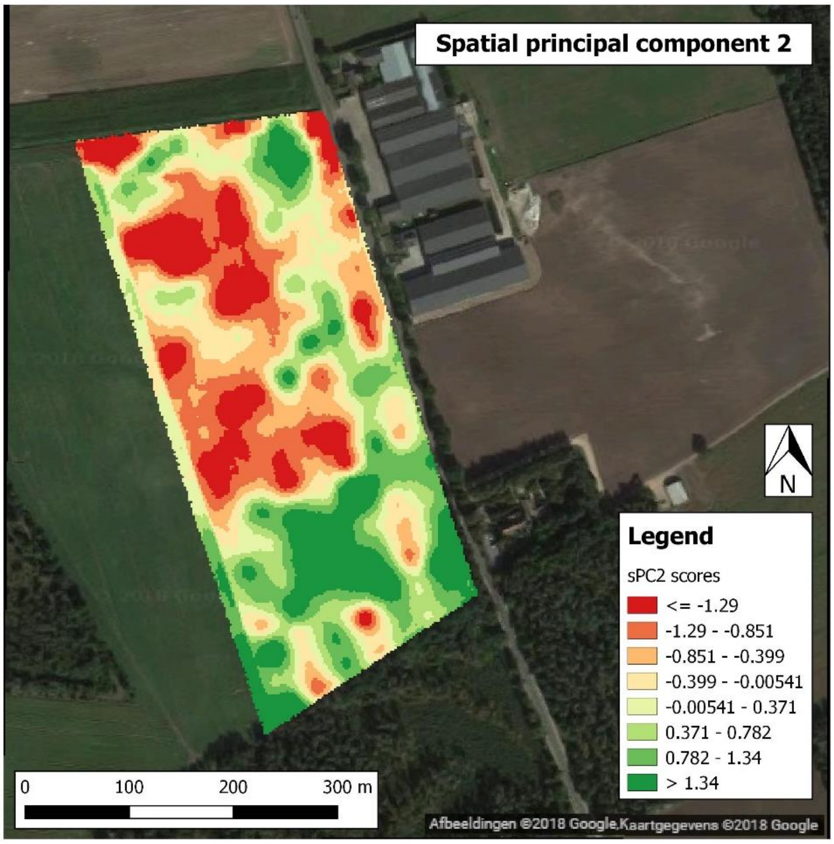


Figure 4.19 – Map of spatial principal component 2 including sPC scores

4.3.2 Management zones obtained from cluster analysis

In order to actually delineate the potentially homogeneous management zones (MZs), the spatial principal components (sPCs) derived from the MULTISPAT-PCA algorithm were used as input for k-means clustering. As explained before, this algorithm is an unsupervised classification method that allocates the values of given input variables (potato crop yield points) into one of the predefined *k* number of clusters (MZs), in such a way that the sum of squares within each cluster (WSS) is as low as possible (Hartigan and Wong 1979). The clustering scree plot in Appendix 7 shows a steep decline of WSS between one and two clusters, but the distinction between larger numbers of clusters is not clear. Therefore, the k-means algorithm was performed with two, three, and four clusters and visually compared. As suggested by (Arce 2005), smoothing of clustering results was proposed to remove isolated pixels, small cluster patches and sharp boundary edges between MZs, which was performed by running a number of spatial filters with different matrix sizes and different functions. In the end, the map with four MZs smoothed by a modal filter with a size of 35 x 35 pixels was selected for further analysis (Figure 4.20), because it showed the most functional classification results to use for management practices in precision agriculture. Moreover, the clusters in that map gave a clear distinction between the north and the south of the field, which was also observed to a certain degree in the patterns of potato crop yield (Figure 4.15).

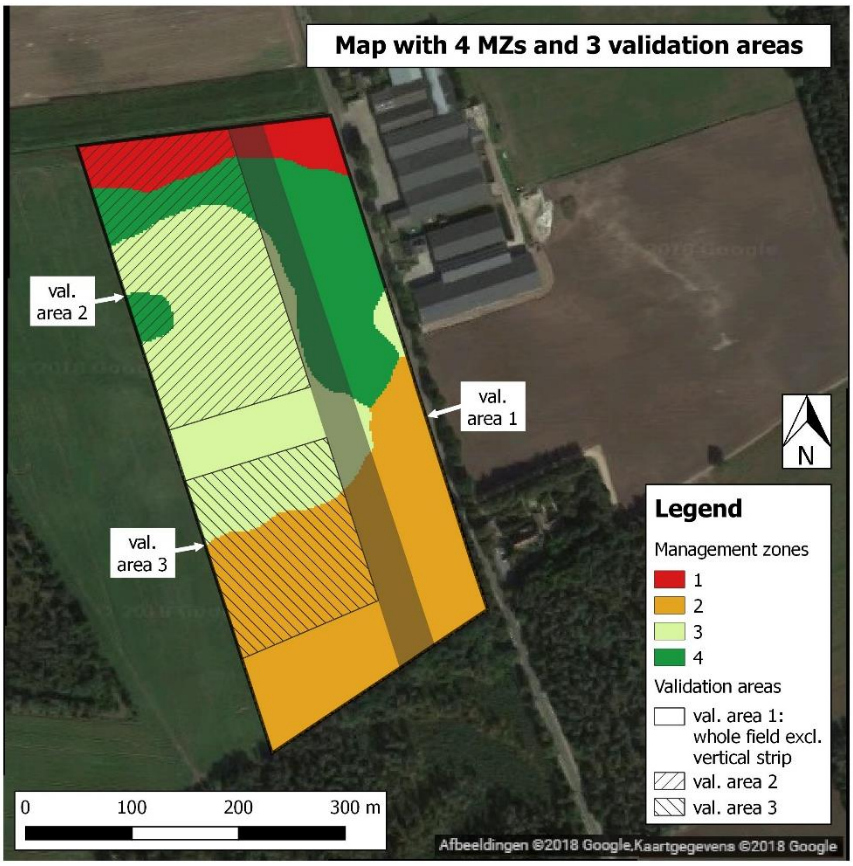

Map with 4 MZs and 3 validation areas
val. area 2
val. area 1
val. area 3
N
Legend
Management zones
1
2
3
4
Validation areas
val. area 1: whole field excl. vertical strip
val. area 2
val. area 3
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

Figure 4.20 – Cluster map with four management zones, and the indication of three areas used for validation by means of linear models (section 4.4)

4.4 Validation of potential management zones

Three areas for validation of MZs were assigned to the potato field, since those areas contained two or more MZs (Figure 4.20). Validation area 1 covers the whole field excluding the strip without additional N input, validation area 2 is located in the north-west of the field in the zone with 90 kg/ha N input, and validation area 3 in the mid-west of the field in the zone with 162 kg/ha N input. For each of these three areas, one out of five statistical models was chosen to compare expected crop yield between the delineated MZs by means of the two different hypotheses described in section 3.5.

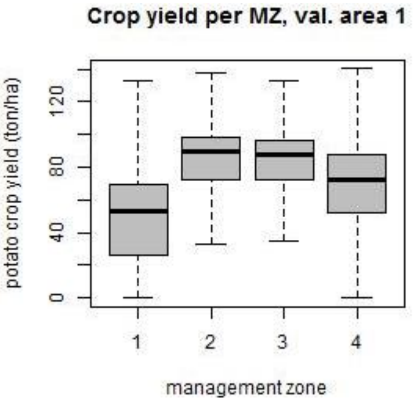


Figure 4.21 – Side-by-side boxplots crop yield within the 4 management zones in val. area 1. Significant differences occur between MZ 1-2, MZ 1-3, MZ 2-4

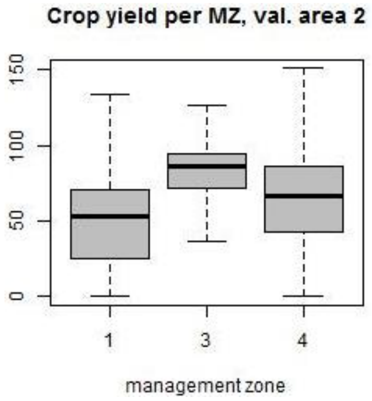


Figure 4.22 – Side-by-side boxplots crop yield within the 3 management zones in val. area 2. Significant difference occurs between MZ 1-3

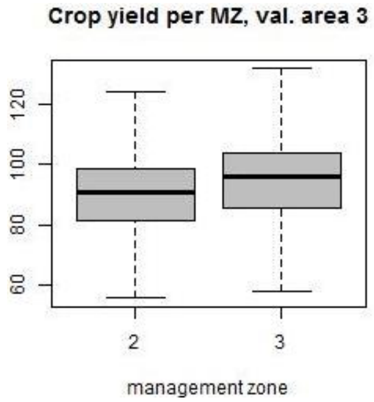
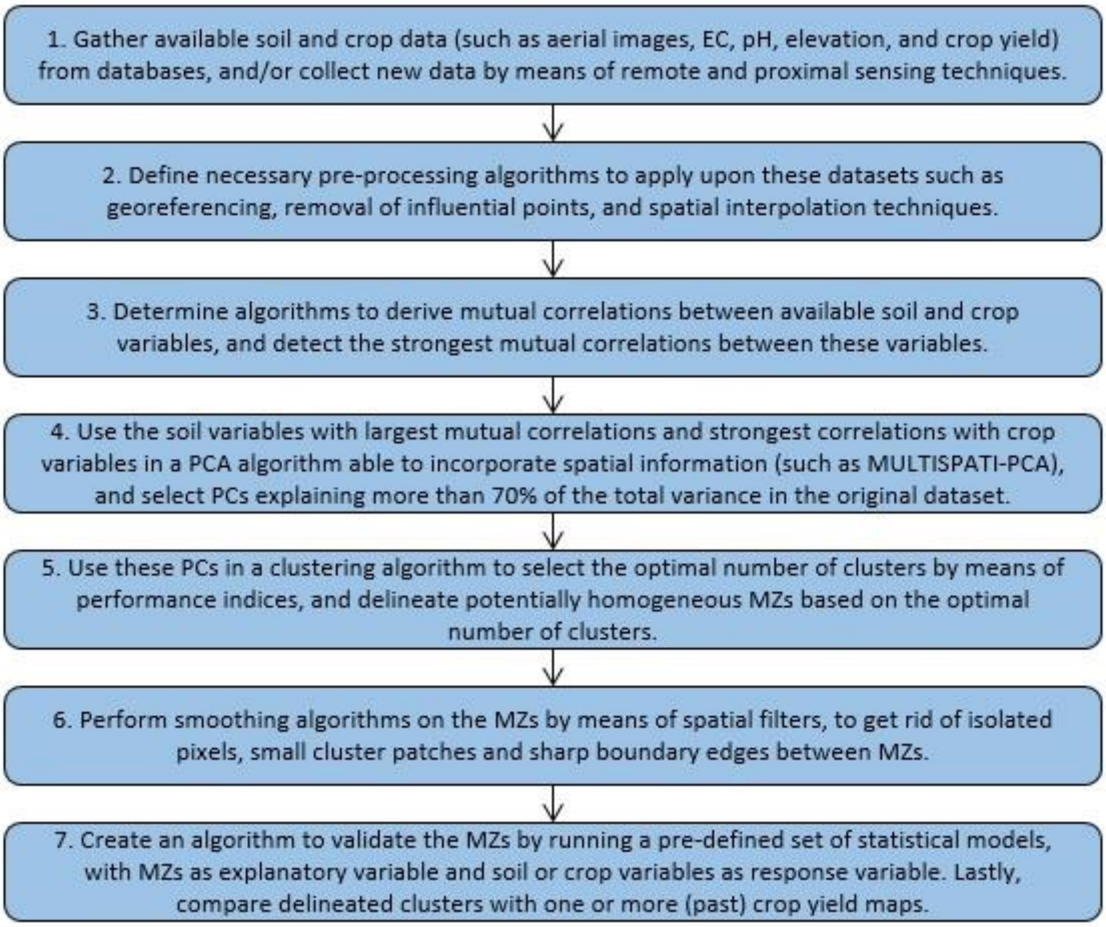


Figure 4.23 – Side-by-side boxplots crop yield within the 2 management zones in val. area 3. The difference between the two zones is not significant

First of all, for validation area 1, the model with the lowest AIC is a mixed model with error terms having a spherical correlation with nugget effect (Appendix 9). The ANOVA table shows an F-value of 6.043 (for *clusters*), with a significant p-value of 0.000, so the null hypothesis is rejected and expected crop yield between at least two MZs is significantly different. The boxplots in Figure 4.21 and the least square (LS) means (Appendix 9) indicate noteworthy mean differences in crop yield between MZ1-2, MZ1-3, and MZ1-4, as well as between MZ2-4 and MZ3-4. Pairwise comparisons show that for three of these pairs significant differences in expected crop yield occur: between MZ1-2 (25.451), MZ1-3 (23.195), and MZ2-4 (15.633). However, the mean differences between MZ1-4 (9.818) and MZ3-4 (13.377) are also remarkable despite the non-significant p-values. Both the boxplots and the pairwise comparisons clearly show no difference between MZ2-3. Secondly, for validation area 2 a mixed model with exponential correlation without nugget effect is chosen. The ANOVA table shows an F-value of 7.373 and a significant p-value of 0.001, so the null hypothesis is rejected and expected crop yield between at least two MZs is significantly different. The boxplots in Figure 4.22 and the LS means show notable mean differences in crop yield between all three MZs, although the pairwise comparisons in Appendix 9 reveal that expected crop yield is statistically significant only between MZ1-3. However, the mean differences for the other two contrasts (13.544 and 13.603 respectively) are noteworthy as well, regardless of the non-significant p-values. Lastly, for validation area 3, a mixed model with spherical correlation without nugget effect is selected. Since this area contains only two clusters, just one pairwise comparison is made. Therefore, the two hypotheses test the same assertion, namely if the difference in crop yield between MZ2-3 is significant. This is not the case, because the p-value for both the F-test and the t-test is 0.087 (Appendix 9). Also the boxplot in Figure 4.23 and the LS mean (5.329) do not show a large difference in crop yield between the two MZs.

5. Discussion

The main goal of this research was to define and investigate a method to delineate potentially homogeneous management zones (MZs), based on soil and crop data obtained by means of remote and proximal sensing techniques. As suggested by a number of authors, such a method could be incorporated in a decision support system (DSS), which is an information system to support a business in decision making, for instance for management practices in precision agriculture (Heijting et al. 2011; McBratney et al. 2005). Data become increasingly more abundant and of higher quality, so more and more farmers have the potential to use such a DSS in their daily business. Several DSSs have been proposed in the past to support management practices in precision agriculture. For instance, one paper discussed the applications of a software program called Management Zone Analyst (MZA) (Fridgen et al. 2004). This software is able to perform a fuzzy c-means (FCM) clustering algorithm based on descriptive statistics such as a variance-covariance matrix, and to evaluate the delineated clusters by means of a number of performance indices as a guidance on how many clusters to choose as input for management practices. In addition to these kinds of software environments, similar DSSs could also be developed to incorporate in freely available farm management apps, for instance to use on smartphones for decision making purposes in arable farming. For this research, a potential DSS was developed and investigated as well. Guided by the consecutive steps of this system outlined in Figure 5.1, the coming subsections provide a discussion on the results that were described in chapter 4.



1. Gather available soil and crop data (such as aerial images, EC, pH, elevation, and crop yield) from databases, and/or collect new data by means of remote and proximal sensing techniques.

2. Define necessary pre-processing algorithms to apply upon these datasets such as georeferencing, removal of influential points, and spatial interpolation techniques.

3. Determine algorithms to derive mutual correlations between available soil and crop variables, and detect the strongest mutual correlations between these variables.

4. Use the soil variables with largest mutual correlations and strongest correlations with crop variables in a PCA algorithm able to incorporate spatial information (such as MULTISPATI-PCA), and select PCs explaining more than 70% of the total variance in the original dataset.

5. Use these PCs in a clustering algorithm to select the optimal number of clusters by means of performance indices, and delineate potentially homogeneous MZs based on the optimal number of clusters.

6. Perform smoothing algorithms on the MZs by means of spatial filters, to get rid of isolated pixels, small cluster patches and sharp boundary edges between MZs.

7. Create an algorithm to validate the MZs by running a pre-defined set of statistical models, with MZs as explanatory variable and soil or crop variables as response variable. Lastly, compare delineated clusters with one or more (past) crop yield maps.

Figure 5.1 – Possible decision support system (DSS) to delineate potentially homogenous zones for an arable field as a basis for management practices in precision agriculture

5.1 Available datasets and data pre-processing

A whole range of soil and crop variables are potentially available to use in a DSS to support management practices in precision agriculture. Research including a number of soil and landscape variables showed that at least three key variables could be identified to form a stable basis for delineation of management zones: EC, elevation and pH (Van Meirvenne et al. 2012; Vitharana et al. 2008). However, pH was not included in this research, because this variable had not been measured yet for the potato field of Van den Borne farm at the time of this research. In addition, two papers suggested that crop yield data of at least five years should be used for the purpose of delineating stable management zones (Boydell and McBratney 2002; Stoorvogel et al. 2015). In addition, a paper by (Mulla 2012) recommended the use of historical remote sensing images in combination with real-time remote sensing data in high spectral and spatial resolutions for improved delineation of management zones. For this thesis research, potato crop data in the form of crop yield and vegetation indices were available from only one year (2015), so evaluating the performance and stability of delineated MZs could not be achieved by comparing crop data from multiple years.

Pre-processing for all datasets was performed by a number of steps explained in section 3.2. One such pre-processing step was the detection and removal of influential points in spatial point datasets (ECa, soil OM, soil depth, elevation, and crop yield). Influential points possibly consisted of outliers that are observations falling far outside the general range of values in a dataset (Osborne and Overbay 2004), while inliers are observations that do fall within the general range of values, but are significantly different from their neighbouring points (Córdoba et al. 2016). Both inlier and outlier detection were tried on all point datasets. On the one hand, removal of inliers worked well for ECa and crop yield due to a high point density with many near neighbouring points and because of possible sensor-based measurement errors, but this did not work well for the low point density datasets obtained by manual sampling including SOM, soil depth A-horizon and elevation, because neighbouring points were much further apart, and the number of measurement errors was very low. On the other hand, removal of outliers worked well for the manually sampled point datasets, since they met the condition required for outlier detection to be normally distributed, which was not the case for ECa and crop yield (indicated by the boxplots in Appendix 2). Hence, in the end the decision was made to perform inlier detection and removal only on the ECa datasets and crop yield, and outlier detection and removal only on SOM, soil depth A-horizon and elevation, from which the results were explained in section 4.1.2. Moreover, one approach to detecting inliers was to evaluate Moran scatterplots (Appendix 3). Apart from Moran’s local index, the points in a Moran scatterplot were also based on a number of other diagnostic statistics, that were not taken into account to remove inliers, because this would lead to a huge loss of observations. Therefore, inliers were solely removed based on Local Moran’s index and statistically significant observations ($p < 0.05$). These approaches were in contrast with the methods by (Córdoba et al. 2016) and (Gili et al. 2017) that performed both outlier and inlier removal on all spatial point datasets, and conducted inlier removal by means of both Moran’s local index and all other derived diagnostic statistics.

The next pre-processing step was transforming the spatial point datasets into raster datasets by means of ordinary kriging. Variograms and variogram parameters included in Appendix 4 that were used as input for kriging functions show a large variety in results between the different datasets. For instance, the minimum nugget was zero, which was the case for all EC datasets. One reason could be that (almost) no measurement errors had taken place, or that no unexplained variance was left in the model. On the other hand, the maximum nugget was 254.3 for potato crop yield. Possible reasons for this large nugget are measurement errors, or unexplained variation at short distances between points, which could partly be caused by the tramlines (or tractor paths) within the field that are for instance visible in Figure 4.15. Also large differences in both the sills and ranges between the different datasets was observed (Appendix 4). the minimum sill was 0.006 (soil depth), and the maximum sill was 555.6 (crop yield), whereas the minimum range was 14.8 (soil depth) and the maximum range 1654.1 (elevation). The larger ranges could indicate a large spatial dependence over distances that fall far beyond the borders of the potato field, causing the variogram function to increase very slowly. On the other hand, variograms of crop yield and soil depth have very small ranges that cause steep increases in the beginning of the variograms, indicating spatial dependence at very small distances between points. As explained before, experimental variograms were fitted in an automated way by means of the *AutofitVariogram* function in R (Hiemstra et al. 2009), since eight point variables had to be interpolated by means of ordinary kriging based on those variograms. However, this led to a lack of control on the variogram parameters, which could be one reason for the large variety in variogram curves and parameters and large differences in nuggets and ranges. As suggested by (Calder and Cressie 2009), manually determining the different variogram functions and parameters could have optimized the models, which could have given improved variograms and improved interpolation results.

Two other pre-processing steps were explored. First of all, shadows caused by trees on the edges south of the potato field are visible on the False colour image, NDVI, and WDVI (maps B, I, and J in Appendix 10). Therefore, methods for shadow removal suggested by (Murali and Govindan 2013) and (Qin et al. 2013) were investigated to normalize pixels. However, because of the relatively small patches of shadow present on the maps opposed to the complexity of those methods, it was decided to leave these steps out of scope for the current research. In addition, tramlines (or tractor paths) between the crop cultivation areas were visible on those maps as well, which were assumed to be of great influence on the MZ validation results. For that reason, Otsu's binary thresholding was tried on these crop images (Otsu 1979), in order to remove these patterns. However, this led to a huge loss of data in the form of pixels, partly because of a low-yield area in the right-center of the field. In addition, decreasing Otsu's threshold value caused less data to be removed, but also made the tramlines disappear again. Hence, in the end it was decided not to perform this pre-processing on the different crop images.

In the end, raster datasets of all soil and crop variables were resampled to a spatial resolution of 2m x 2m (Appendix 10). This was a trade-off between spatial resolution and processing efficiency. At this spatial resolution, many spatial patterns were still visible in the field, but for coarser resolutions, more and more information started to get lost in the maps of all variables. On the other hand, when the resolution became too fine (1m x 1m or smaller), the number of pixels would exceed 50,000, which started to influence processing time and memory. The number of pixels within the extents of the field boundary for a spatial resolution of 2m x 2m was 31790 pixels (in Table 4.1 represented as spatial points), which is a good compromise between the detail of spatial patterns visible in the field, and processing time and memory.

In conclusion, for the purpose of developing a proposed DSS for management practices in precision agriculture, it is important to evaluate which kinds of soil and crop data are available. Besides this, it is advised to investigate soil and landscape features, as well as variations in weather conditions for the location where management practices are intended. Based on these kinds of information, decisions can be made about which additional data are needed, and which pre-processing steps such as detection and removal of influential points, spatial interpolation, and georeferencing are required to fulfil the first 2 steps of the DSS presented in Figure 5.1.

5.2 Statistical description and visual assessment of datasets

Section 4.2 described descriptive statistics, correlations between pairs of variables, and visual assessment of all variables. This last step was performed in QGIS. QGIS, as well as other GIS software, has the possibility to use so-called *Open Layers* linked with the web containing aerial images or terrain maps of a given area, which could be included as background layers in maps of spatial datasets (QGIS 2018). However, all available layers were available in WGS84 coordinates only, whereas the soil and crop datasets for this research were stored or transformed into RDNew coordinates. Normally speaking, GIS software has the ability to project datasets with different coordinate systems *On The Fly*, which means that all layers are automatically stacked in accordance to one and the same coordinate system. However, this did not work out properly, so as explained in section 3.3, the alternative was to download an aerial image of the potato field from Google maps (Google 2018), pre-process this image by means of georeferencing, and use it as a background layer in the maps of all spatial datasets.

One category of soil variables were spectral reflectance indices derived from the aerial image of bare soil. Calculation of spectral reflectance indices from aerial photographs of bare soil were suggested by (Bartholomeus and Kooistra 2012; Bartholomeus et al. 2014), in order to quantify soil organic matter (SOM) and soil organic carbon (SOC). Moderate to strong correlations between these soil parameters and spectral reflectance indices were observed, suggesting that higher spectral index values led to lower SOC or SOM values. Two spectral indices were used in this research: the sum of the three visible RGB colour bands (SUMVIS) and the Normalized Difference Red Green (NDRG), respectively calculated by means of equations 3.1 and 3.2 in section 3.2.3. The correlation between SOM and NDRG was indeed noteworthy (-0.34), but there was no correlation between SOM and SUMVIS ($r = -0.01$). Additionally, a number of papers found significant correlations between soil moisture and soil spectral reflectance (Hadjimitsis et al. 2013; Weidong et al. 2002). In general, lower reflectance values indicated larger soil moisture levels, whereas higher values represented dryer soils (Figure 5.2). This information suggested that the SUMVIS could also be an effective measure to quantify soil moisture, because as equation 3.1 showed, higher reflectance values would lead to higher SUMVIS values. The SUMVIS of the potato parcel (map C in Appendix 10) showed smaller pixel values (hence wetter soil) in the west of the field than in the north and east parts of the field, which matched very well with the respective dark and light patterns observed on the aerial image of bare soil (map A in Appendix 10). Other soil variables showed interesting features and significant relations as well. As described before, ECa (Figure 4.14) had low pixel values in the south of the field, and high values in the north, especially in the headland of the field. One reason for the large values in the north is that this is the entrance of the field where the soil is disrupted by agricultural vehicles that manoeuvre to enter or leave the field. Another reason for those high values could be the descending elevation pattern (map H in Appendix 10), causing water run-off towards the ditch in the north of the field (also visible in Figure 3.3), disposing dissolved minerals and SOM in the soil at that location. In fact, ECa showed very high mutual correlations, but also moderate to strong correlations with other soil variables. Papers such as (Fraisse, Sudduth, and Kitchen 2001) and (Córdoba et al. 2016) observed significant correlations between ECa and soil depth as well. The correlations between EC and SOM were in general lower, which was also the case in the paper by (Moral et al. 2010). One possible reason could be that SOM had a very smooth pattern with a patch of land in the mid-west of the field containing very low pixel values, which did not seem to match very well with the described EC patterns. Lastly, correlations between ECa and crop yield were moderate (ranging between -0.39 and -0.42), while in papers such as (Johnson et al. 2003), very high correlations between ECa and crop yield were observed (extending between 0.79 and 0.99).

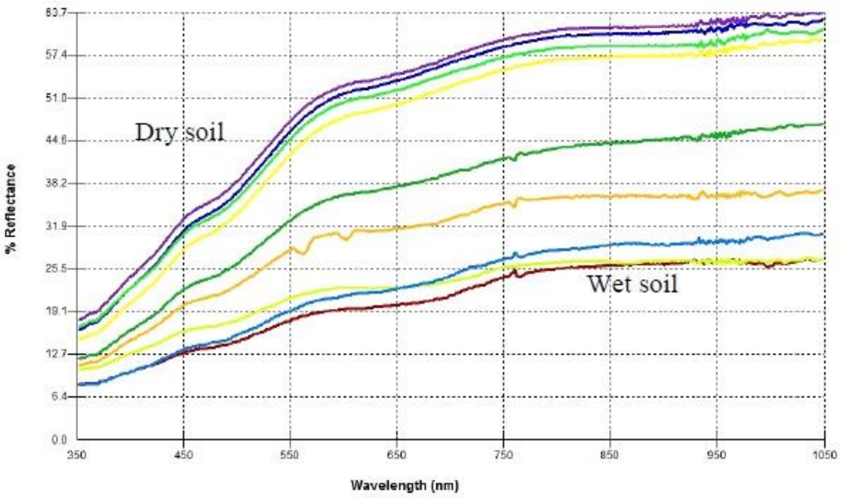


Figure 5.2 – Spectral reflectance curves ranging from wet soil to dry soil (Hadjimitsis et al. 2013)

Spatial patterns observed in other soil and crop images corresponded well with maps created in past studies, such as the researches by (Areda 2013) and (Van den Brande 2015). Nevertheless, a number of seemingly contradictory patterns were visible between the maps of crop yield (Figure 4.15) and the False colour image, NDVI, and WDVI (maps B, I, and J in Appendix 10), although the correlations between these three variables were moderate to strong (ranging from 0.43 to 0.66, as Appendix 5 shows). First of all, on almost all aerial images and maps with indices (for instance in map D of Appendix 10), a vertical strip of land on the border of the field in the west was observed with atypical pixel values compared to neighbouring pixel values. Information from Jacob Van Den Borne revealed that the field boundary at that location had not been taken into account very strictly. In other words, some of the crops cultivated on the neighbouring field in the west were accidently planted on the potato field as well, which is the cause of the deviating strip of land in the west of the field. Second, as expected, the vertical strip of land without additional N input (Figure 4.1) showed low values in WDVI and NDVI, and a dark shade in the False Colour image. However, these patterns did not neatly follow the boundaries of that strip, most likely due to drift (scattering of particles in the air by wind) that occurred when spraying fertilizer, causing some of the crops in that strip to receive additional N-input after all. Third, as expected, crop yield in the fertilization zone with 0 kg/ha N input in the middle of the field was low (Figure 4.15), but the WDVI map (map J in Appendix 10) showed high pixel values on that location. Information from Jacob Van Den Borne revealed that late in the growing season of 2015, some additional fertilizer had been applied on that location, which positively affected the leaves of the plants (visible on the WDVI map), whereas potato crop yield in that area was already low (visible on the crop yield map) due to the lack of initial N input before the start of the growing season. Lastly, potato crop yield in the headland in the north of the field (Figure 4.15) appeared to have very low values. One possible explanation is that this is the entrance of the field where the soil and crops had been disrupted by agricultural vehicles regularly manoeuvring to enter or leave the field (which was also observed at the EC map). A second reason could be that during the growing season of 2015, a leakage from a water hose had taken place that negatively influenced the production of potatoes at that location.

In conclusion, interpreting descriptive statistics and correlations, and comparing spatial patterns between all soil and crop variables are important first steps to get a good overview of the interrelations and coherence between variables (partly indicated by step 3 in Figure 5.1) to serve as a basis for further analysis such as PCA and cluster analysis, which are discussed in the following sections.

5.3 Delineation of potential management zones

As explained earlier, the first step for delineation of management zones (MZs) was to conduct a special type of PCA, called MULTISPATI-PCA that took spatial information into account to generate spatial principal components (sPCs) based on selected input variables. Selection of soil and elevation variables was based on three criteria. The first one was the correlation matrix (Appendix 5), where initially good correlations were found between NDRG, ECa and elevation on the one hand, and crop yield on the other hand. Secondly, Bartlett's test of Sphericity was performed, which was significant, meaning that the set of soil and crop variables were suitable to use as input for PCA (Appendix 6). However, a disadvantage of such statistical tests is that they always tend to become significant for large sample sizes (Snedecor and Cochran 1989), so for 31790 observations (as indicated by Table 4.1), the test was definitely going to be significant. The third criterion to detect which variables were suitable, was the KMO Measure of Sampling Adequacy statistic (Appendix 6). The procedure to calculate KMO values was based on a partial correlation matrix explaining correlations between one variable and all the remaining variables together, rather than describing pairwise correlations as shown in the correlation matrix in Appendix 5. Based on this information, two other variables (Soil depth A-horizon and Total SOM) were selected as input for PCA, in addition to the other soil variables mentioned before. In comparison, the paper by (Moral et al. 2010) conducted a correlation analysis to select variables and found EC at two soil depths, and soil textures represented by clay, coarse sand, and fine sand to be the most representative soil characteristics to be included in PCA. An alternative for variable selection was described by (Schenatto et al. 2017), which determined mutual spatial autocorrelation between soil variables, and autocorrelation between these variables and crop yield. Variables that did not show significant autocorrelation were eliminated, leading to elevation and soil penetration resistance (SPR) at a number of soil depths to be the remaining variables to use as input for PCA.

Another commonly used approach was to first investigate correlations between soil variables visually, and then use PCA itself as a method for variable selection as a basis for MZ delineation. As described earlier, researches by (Vitharana et al. 2008) and (Van Meirvenne et al. 2012) found ECa, Elevation and pH as output of PCA to the most important features for explaining variability in soils from a number of study sites in Belgium. Papers by (Fraisse et al. 2001) and (Molin and Castro 2008) also detected ECa and Elevation, and ECa and pH respectively, but the first study additionally found the soil's slope as one of the most important variables for explaining variability in the study fields. These findings were merely based on ordinary PCA algorithms, but as described in section 2.3.4, also approaches based on MULTISPATI-PCA have been investigated. For example, the procedure described in a paper by (Córdoba et al. 2016) found ECa at two soil depths, elevation, and soil depth to be of most importance for explaining spatial variability, whereas a study by (Gili et al. 2017) found good correlations between the derived sPCs on the one hand, and SOM, clay + silt, pH, EC, and phosphorous on the other hand. The results in the current research (Table 4.2) correspond well with the findings from these and other past researches, since ECa, Soil depth, Total SOM, NDRG of bare soil, and Elevation were found to explain the largest part of spatial variability in the potato field, expressed in four sPCs accounting for 94.08% of the total variance in the set of selected soil and topographic variables.

Either the variables accounting for the largest part of variance, or the (spatial) principal components themselves could have been used as input for cluster analysis algorithms to delineate potential MZs. For instance, (Fraisse et al. 2001) used original variables such as elevation, slope, curvature, and EC as input to derive both three and six management zones in an unsupervised classification algorithm in a GIS software environment. In addition, (Scudiero et al. 2013) used NDVI and normalized soil variables such as ECa as input for an FCM cluster algorithm, and came up with a five-cluster solution. Besides these approaches, researches by (Molin and Castro 2008) and (Moral et al. 2010) used two principal components derived from various soil variables as input to generate a map with three MZs for a study site in Brazil, and a map with two MZs for a study site in Spain, respectively. Lastly, a paper by (Gili et al. 2017) compared a number of methods to delineate potential MZs based on original soil variables and based on sPCs derived from the MULTISPATI-PCA algorithm. The study found solutions with two, three and four clusters, and concluded that the choices for the method and for the number of clusters depended on the objectives of crop management, main yield-limiting factors, and the agro-ecological conditions of an arable field.

Also a number of approaches to determine the optimal number of clusters have been proposed. For instance, that number could have been chosen by observing a scree plot, or by calculating a number of cluster performance indices. A cluster scree plot as shown in Appendix 7 indicates the total within sum of squares (WSS) for each given number of clusters in descending order on a bar plot, and can be used to decide how many clusters to retain. However, cluster performance indices are a more exact method to decide upon the optimal number of clusters. For instance, the study by (Schenatto et al. 2017) used the fuzziness performance index (FPI), which is a measure of the degree to which different classes share a cluster membership, and the modified partition entropy (MPE), which estimates the amount of disorder caused by a specific chosen number of clusters (Boydell and McBratney 2002), to come up with an optimal number of four cluster for each of two study fields. In addition to selecting a suitable number of clusters, (Arce 2005) proposed statistical models in the form of non-linear spatial filters for smoothing classification results. For instance, the procedure by (Córdoba et al. 2016) derived a two cluster-solution based on a median filter with a size of 9 x 9 pixels, in addition to FPI and MPE performance indices and two spatial principal components as input for the MULTISPATI-PCA algorithm.

For this research, two, three and four cluster-solutions were evaluated, based on the four sPCs as input variables. The number of clusters were merely chosen based on the cluster scree plot on the right side in Appendix 7. Cluster performance indices were not calculated, because these methods were found in papers only at the very end of my thesis. In addition, a large range of filter sizes and functions were tried for smoothing of classification results. In the end, the four cluster-solution was chosen (Figure 4.20) smoothed with a modal filter of 35 x 35 pixels. This is in contrast with the type and size of filter used by (Córdoba et al. 2016), although it showed the most functional classification results to use as a basis for management practices in precision agriculture. A map with two MZs (map A in Appendix 12) did not seem to cover all the spatial variation within the field, and a map with three MZs (map B in Appendix 12) kept showing small cluster patches scattered across the field, regardless of the chosen type and size of filter. Moreover, the map with four clusters gave a clear distinction between the north and the south of the field, which was also observed to some degree from the potato yield pattern shown in Figure 4.15. The four delineated MZs (Figure 4.20) appeared to be influenced by a number of sPCs, and hence by a number of original soil variables. Especially MZ1, but also MZ2 seemed to be dominated by sPC1 and ECa, although patterns of Total SOM were visible in MZ2 as well. Also some influence of ECa was visible in MZ3 and MZ4, although these zones mostly appeared to be dominated by Total SOM and bare soil NDRG. It was a bit harder to detect elevation patterns in the clusters though, since elevation was mostly correlated with sPC4, which only accounted for a small part of the total variance in the dataset (6.62% as indicated by Table 4.2).

The results discussed in this section can be used as input for step 3 to step 6 of the DSS described in Figure 5.1. These steps can be summarized as follows. Variable selection as input for PCA could be achieved by a number of ways, such as correlation analysis and the calculation of KMO variables, or by variable selection based on a PCA algorithm itself. It is up to the researcher to choose the most appropriate PCA algorithm (such as ordinary PCA or MULTISPATI-PCA). Moreover, a decision should be made to use either the original variables, or the (spatial) PCs explaining the largest proportion of variance as input for a cluster analysis such as k-means of FCM. Finally, a suitable number of clusters should be selected based on either a cluster scree plot, or on performance indices such as FPI or MPE, and the classification results could optionally be smoothed for instance by non-linear spatial filters.

5.4 Validation of potential management zones

The final stage of this research was to validate the delineated MZs, which was conducted for the three validation areas shown in Figure 4.20. For validation area 1, the vertical strip of land was excluded since it was expected to negatively influence the validation results. Validation area 2 and 3 were chosen based on the fact that they contained two or more MZs. As explained in section 4.4, validation was performed by a number of statistical models that were fitted on a stratified random sample of 100 observations per MZ, as suggested by (Gili 2013) and (Córdoba et al. 2016), for each of the three areas. Model selection was performed by Akaike's Information Criterion (AIC). To potentially improve performance of models, other sampling and model selection methods could have been used as well, such as bootstrapping and jackknife resampling. Bootstrapping is a resampling method to draw random samples of observations with replacement, which could be useful for estimating sampling distributions or hypothesis testing. Jackknife resampling is an improved method that systematically leaves out one observation of a dataset, calculates a parameter estimate or fits a model on the remaining observations, and finally determines the average of those calculations or model fits (James et al. 2013). For this research, the model outputs showing expected crop yield means and pairwise comparisons between MZs (included in Appendix 9) appeared to match with the distributions of the boxplots in Figure 4.21 to Figure 4.23 (which were based on the total number of observations in each MZ). Hence, the methods for sampling and model selection seemed to have performed appropriately.

Many studies used statistical models, such as one-way ANOVA or mixed linear models, and most of them found significant differences in soil parameters or crop yield between the different MZs (Córdoba et al. 2016; Molin and Castro 2008; Scudiero et al. 2013). The model outputs (Appendix 9) and boxplots (Figure 4.21 to Figure 4.23) in this research showed significant, or at least noteworthy pairwise differences in expected crop yield between all MZs, expect for the difference between MZ2 and MZ3, which was neither statistically significant, nor practically significant. Evaluated by models fitted on samples of validation area 1 and area 3, only very small difference in expected crop yield was observed between these two zones, indicating that they possibly could have been merged into one MZ. Considering the spatial patterns, one reason for this could be that MZ2 and MZ3 were particularly delineated based on sPC2 (Figure 4.17), which in turn was dominated by soil depth and Total SOM. However, the correlations between these variables and crop yield were relatively low (-0.23 and -0.24, respectively), suggesting that these variables probably should have been left out of the PCA.

Validation of MZs is an important last stage (indicated by step 7 in Figure 5.1) to investigate whether MZ delineation had worked out well or not. It is up to the researchers and farmers to choose which kinds of models and sampling methods for this purpose, since their performance depend on the available data, on the PCA and cluster analysis results, and more broadly speaking on the agro-ecological conditions of agricultural locations and the varying weather patterns per growing season (Gili et al. 2017).

6. Conclusions and recommendations

This chapter describes the conclusions and recommendations of this thesis research. Conclusions are based on the research questions in section 1.3, while the recommendations are guided by Figure 5.1.

6.1 Conclusions

The main objective of this research was to define and investigate a method to determine how soil variation related to crop yield to delineate potentially homogeneous management zones for precision agriculture, based on data obtained from remote and proximal sensing technology. This objective was addressed by four sub-research questions.

The first research question was intended to investigate which datasets and methods were available for delineating management zones. From the literature review, a broad variety of remote and proximal sensing data such as aerial images, ECa, (SOM), and crop yield could be collected for this purpose. In addition, a large range of methods were available to delineate MZs, but the most widely used approaches were correlation analysis to detect the most strongly associated variables, principal component analysis for variable reduction, and cluster analysis to actually delineate the MZs, either based on original soil and crop variables, or on (spatial) principal components as input variables.

The second research question anticipated on detecting spatial patterns and relations among soil and crop variables. Electric conductivity showed low values in the south of the field, whereas (very) high values were observed in the north and east parts of the field. Elevation as well as SOM showed very smooth (descending) patters, whereas other soil variables showed more scattered patterns in the field. Mutual correlations between EC were very strong (0.91 and higher), whereas correlations between other soil variables were weak to very strong (ranging from 0.20 to 0.94). Bare soil indices, except for NDRG, showed very weak or even meaningless correlations with other soil and crop variables. Crop yield, NDVI and WDVI showed distinctions between the north and south parts of the field, and clearly deviation patterns in the zones with no initial an addition N input. Correlations between these crop variables were moderate to strong, ranging from 0.43 to 0.66.

The third research question was intended on developing and investigating a method to delineate potentially homogeneous MZs. This was performed by first pre-processing available soil and crop datasets. Second, descriptive and visual assessments of those data were made, and third, variable selection was executed based on a correlation analysis, Bartlett’s test of Sphericity, and KMO-MSA statistics. Fourth, the selected variables were used as input for MULTISPATI-PCA, from which four sPCs explaining 94.08% of the total variance were used as input for k-means cluster analysis, which led to four potential MZs. Lastly, classification results were smoothed by means of non-linear spatial filters.

The last research question aimed at validating the delineated MZs. Validation was performed for three areas in the field that were based on designated fertilizer zones. For each of the three areas, a stratified random sample of crop yield was drawn as input for a number of statistical models. Significant differences ($p < 0.05$), or at least practical differences (9 ton/ha or more) in expected crop yield were found between all MZs, except between MZ2 and MZ3.

The consecutive data pre-processing and analysis stages investigated in this research could be incorporated in a decision support system (DSS) to support management practices in precision agriculture. Data become increasingly more abundant and of higher quality, so more and more farmers have the potential to use such a DSS in their daily businesses.

6.2 Recommendations

- Collect and include soil pH measurements to serve as an additional variable in future research.
- Investigate additional pre-processing steps for aerial images and crop yield maps, such as Otsu's binary thresholding to eliminate pixels representing soil, and algorithms to normalize pixels that are influenced by shadows of objects such as trees alongside a field.
- In addition to the *AutofitVariogram* function, investigate alternative methods for variogram estimation in order to optimize variogram parameters and variogram fitting.
- Explore other kriging methods such as cokriging and regression kriging, which allow additional (soil and crop) variables in their models that could improve spatial predictions.
- Determine performance indices such as FPI and MPE to determine an optimal number of clusters.
- Explore additional statistical models and sampling methods to validate the delineated MZs.

References

Adamczyk, Joanna and Antonia Osberger. 2015. ‘Red-Edge Vegetation Indices for Detecting and Assessing Disturbances in Norway Spruce Dominated Mountain Forests’. *International Journal of Applied Earth Observations and Geoinformation* 37(Complete):90–99.

Akaike, Hirotugu. 1981. ‘Likelihood of a Model and Information Criteria’. *Journal of Econometrics* 16(1):3–14. Retrieved (http://www.sciencedirect.com/science/article/pii/0304407681900713).

Anselin, Luc. 1995. ‘Local Indicators of Spatial Association—LISA’. *Geographical Analysis* 27(2):93–115.

Anselin, Luc. 1996. ‘The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association’. Pp. 111–25 in *Spatial Analytical Perspectives on GIS*. London: Taylor & Francis.

Arce, Gonzalo R. 2005. *Nonlinear Signal Processing: A Statistical Approach*. John Wiley & Sons.

Areda, A. T. 2013. ‘A Multi-Sensor Analysis of Vegetation Indices for Leaf Area Index Retrieval in Precision Agriculture’. Wageningen University and Research.

Bakker, N. J. 2014. ‘Exploring the Impact of Soil Compaction on Relative Transpiration by Potatoes : A Case Study Performed at “Van Den Borne Aardappelen”’. Wageningen University and Research.

Barnes, Edward M. et al. 2003. ‘Remote- and Ground-Based Sensor Techniques to Map Soil Properties’. *Photogrammetric Engineering & Remote Sensing* 69(6):619–30. Retrieved (https://www.ingentaconnect.com/content/asprs/pers/2003/00000069/00000006/art00002).

Bartholomeus, H. M. and L. Kooistra. 2012. ‘Use of Aerial Photographs for Assessment of Soil Organic Carbon and Delineation of Agricultural Management Zones’. P. 212 in *Geophysical Union (EGU) Conference, Vienna, Austria, 22 - 27 April, 2012*, vol. 14.

Bartholomeus, H., J. M. Suomalainen, and L. Kooistra. 2014. ‘Estimation of within Field Variation of SOM Using UAV Based RGB and Elevation Data’. 79, , .

Bartholomeus, Harm et al. 2011. ‘Soil Organic Carbon Mapping of Partially Vegetated Agricultural Fields with Imaging Spectroscopy’. *International Journal of Applied Earth Observation and Geoinformation* 13(1):81–88. Retrieved (http://www.sciencedirect.com/science/article/pii/S0303243410000796).

Bauer, Marvin E. and Jan E. Cipra. 1973. ‘Identification of Agricultural Crops by Computer Processing of ERTS MSS Data’. *LARS Technical Reports* 20.

Ben-Dor, E. et al. 2009. ‘Using Imaging Spectroscopy to Study Soil Properties’. *Remote Sensing of Environment* 113:S38–55. Retrieved (http://www.sciencedirect.com/science/article/pii/S0034425709000753).

Ben-Hur, Asa and Isabelle Guyon. 2003. ‘Detecting Stable Clusters Using Principal Component Analysis’. *Methods in Molecular Biology (Clifton, N.J.)* 224:159—182. Retrieved (https://doi.org/10.1385/1-59259-364-X:159).

De Benedetto, Daniela et al. 2013. ‘An Approach for Delineating Homogeneous Zones by Using Multi-Sensor Data’. *Geoderma* 199(0):117–27. Retrieved (http://www.sciencedirect.com/science/article/pii/S0016706112003230).

Bezdek, James C., Robert Ehrlich, and William Full. 1984. ‘FCM: The Fuzzy c-Means Clustering Algorithm’. *Computers & Geosciences* 10(2–3):191–203. Retrieved (http://www.sciencedirect.com/science/article/pii/0098300484900207).

Bhatti, A. U., D. J. Mulla, and B. E. Frazier. 1991. ‘Estimation of Soil Properties and Wheat Yields on Complex Eroded Hills Using Geostatistics and Thematic Mapper Images’. *Remote Sensing of Environment* 37(3):181–91. Retrieved (http://www.sciencedirect.com/science/article/pii/003442579190080P).

Bivand, Roger S., Edzer J. Pebesma, Virgilio Gómez-Rubio, and Edzer Jan Pebesma. 2008. *Applied Spatial Data Analysis with R*. Springer.

Van den Borne, Jacob. 2018a. 'Van Den Borne Aardappelen'. Retrieved 31 July 2018 (https://www.vandenborneaardappelen.com/).

Van den Borne, Jacob. 2018b. 'Van Den Borne Loonwerk GPS'. Retrieved 31 July 2018 (http://www.loonwerkgps.nl/).

Boydell, B. and A. B. McBratney. 2002. 'Identifying Potential Within-Field Management Zones from Cotton-Yield Estimates'. *Precision Agriculture* 3(1):9–23. Retrieved (http://dx.doi.org/10.1023/A%3A1013318002609).

Van den Brande, Marnix. 2015. 'Sensing the Nitrogen Balance in Potatoes'. Wageningen University and Research.

de Bruijne, A., J. Van Buren, A. Koesters, and H. Van Der Marel. 2005. *De geodetische referentiestelsels van Nederland : definitie en vastlegging van ETRS89, RD en NAP en hun onderlinge relaties = Geodetic reference frames in the Netherlands : definition and specification of ETRS89, RD and NAP, and their mutual relationships*. Delft: Nederlandse Commissie voor Geodesie. Retrieved (http://lib.ugent.be/catalog/rug01:000937054).

Burrough, P. A. 1986. 'Principles of Geographical Information Systems for Land Resources Assessment'. *Geocarto International* 1(3):54. Retrieved (https://doi.org/10.1080/10106048609354060).

Calder, C. A. and N. Cressie. 2009. 'Kriging and Variogram Models A2 - Thrift, Rob KitchinNigel'. Pp. 49–55 in *International Encyclopedia of Human Geography*. Oxford: Elsevier. Retrieved (http://www.sciencedirect.com/science/article/pii/B9780080449104004612).

Castrignanò, A., M. T. F. Wong, M. Stelluti, D. De Benedetto, and D. Sollitto. 2012. 'Use of EMI, Gamma-Ray Emission and GPS Height as Multi-Sensor Data for Soil Characterisation'. *Geoderma* 175–176(0):78–89. Retrieved (http://www.sciencedirect.com/science/article/pii/S0016706112000377).

Chang, Kang-Tsung. 2015. *Introduction to Geographic Information Systems*. McGraw-Hill Science/Engineering/Math.

Christy, C. D. 2008. 'Real-Time Measurement of Soil Attributes Using on-the-Go near Infrared Reflectance Spectroscopy'. *Computers and Electronics in Agriculture* 61(1):10–19. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169907001639).

Clevers, G. Jan, Lammert Kooistra, and M. Marnix van den Brande. 2017. 'Using Sentinel-2 Data for Retrieving LAI and Leaf and Canopy Chlorophyll Content of a Potato Crop'. *Remote Sensing* 9(5).

Clevers, J. G. P. W. 1991. 'Application of the WDVI in Estimating LAI at the Generative Stage of Barley'. *ISPRS Journal of Photogrammetry and Remote Sensing* 46(1):37–47. Retrieved (http://www.sciencedirect.com/science/article/pii/092427169190005G).

Clevers, J. G. P. W. et al. 2001. 'MERIS and the Red-Edge Position'. *ITC Journal* 3(4):313–20.

Clevers, J. G. P. W. and L. Kooistra. 2012. 'Using Hyperspectral Remote Sensing Data for Retrieving Canopy Chlorophyll and Nitrogen Content'. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5(2):574–83.

Colomina, I. and P. Molina. 2014. 'Unmanned Aerial Systems for Photogrammetry and Remote Sensing: A Review'. *ISPRS Journal of Photogrammetry and Remote Sensing* 92:79–97. Retrieved (http://www.sciencedirect.com/science/article/pii/S0924271614000501).

Cordoba, M., C. Bruno, J. Costa, and M. Balzarini. 2013. 'Subfield Management Class Delineation Using Cluster Analysis from Spatial Principal Components of Soil Variables'. *Comput. Electron. Agric.*

97:6–14.

Córdoba, Mariano A., Cecilia I. Bruno, José L. Costa, Nahuel R. Peralta, and Mónica G. Balzarini. 2016. ‘Protocol for Multivariate Homogeneous Zone Delineation in Precision Agriculture’. *Biosystems Engineering* 143:95–107. Retrieved (http://www.sciencedirect.com/science/article/pii/S153751101530204X).

Corwin, D. L. and S. M. Lesch. 2005. ‘Apparent Soil Electrical Conductivity Measurements in Agriculture’. *Computers and Electronics in Agriculture* 46(1–3):11–43. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169904001243).

Corwin, D. L. and S. M. Lesch. 2003. ‘Application of Soil Electrical Conductivity to Precision Agriculture’. *Agronomy Journal* 95(3):455–71.

Corwin, D. L. and S. M. Lesch. 2010. ‘Delineating Site-Specific Management Units with Proximal Sensors’. Pp. 139–65 in *Geostatistical Applications for Precision Agriculture*, edited by M. A. Oliver. Springer Netherlands. Retrieved (http://dx.doi.org/10.1007/978-90-481-9133-8_6).

Costa, José. 2012. *Principal Component Analysis with Georeferenced Data an Application in Precision Agriculture*.

Cressie, Noel. 1990. ‘The Origins of Kriging’. *Mathematical Geology* 22(3):239–52. Retrieved (http://dx.doi.org/10.1007/BF00889887).

Davatgar, N., M. R. Neishabouri, and A. R. Sepaskhah. 2012. ‘Delineation of Site Specific Nutrient Management Zones for a Paddy Cultivated Area Based on Soil Fertility Using Fuzzy Clustering’. *Geoderma* 173–174(0):111–18. Retrieved (http://www.sciencedirect.com/science/article/pii/S0016706111003533).

Dekkers, J. M. J. 1981. *De Bodemgesteldheid van Het Landbouwbedrijf van de Gebr. Laarakker Te Reusel*. Wageningen: Stiboka.

Delegido, J. et al. 2013. ‘A Red-Edge Spectral Index for Remote Sensing Estimation of Green LAI over Agroecosystems’. *European Journal of Agronomy* 46:42–52. Retrieved (http://www.sciencedirect.com/science/article/pii/S1161030112001542).

Ding, Chris and Xiaofeng He. 2004. ‘K-Means Clustering via Principal Component Analysis’. P. 29-- in *Proceedings of the Twenty-first International Conference on Machine Learning*, *ICML ’04*. New York, NY, USA: ACM. Retrieved (http://doi.acm.org/10.1145/1015330.1015408).

Dray, Stéphane, Pierre Legendre, and Pedro R. Peres-Neto. 2006. ‘Spatial Modelling: A Comprehensive Framework for Principal Coordinate Analysis of Neighbour Matrices (PCNM)’. *Ecological Modelling* 196(3):483–93. Retrieved (http://www.sciencedirect.com/science/article/pii/S0304380006000925).

Dray, Stéphane, Sonia Saïd, and Françis Débias. 2008. ‘Spatial Ordination of Vegetation Data Using a Generalization of Wartenberg’s Multivariate Spatial Correlation’. *Journal of Vegetation Science* 19(1):45–56. Retrieved (https://doi.org/10.3170/2007-8-18312).

Environmental Systems Research Institute. 2018. ‘Esri ArcGIS’. *Esri ArcGIS*. Retrieved 1 July 2018 (http://desktop.arcgis.com/en/).

Eurocontrol and IfEN. 1998. *WGS 84 Implementation Manual*. Retrieved (https://www.icao.int/safety/pbn/Documentation/EUROCONTROL/Eurocontrol WGS 84 Implementation Manual.pdf).

Feng, Wei et al. 2015. *Remote Estimation of above Ground Nitrogen Uptake during Vegetative Growth in Winter Wheat Using Hyperspectral Red-Edge Ratio Data*.

Fleming, K. L., D. F. Heermann, and D. G. Westfall. 2004. ‘Evaluating Soil Color with Farmer Input and

Apparent Soil Electrical Conductivity for Management Zone Delineation'. *Agron. J.* 96(6):1581–87. Retrieved (https://www.agronomy.org/publications/aj/abstracts/96/6/1581).

Fleming, K. L., D. G. Westfall, D. W. Wiens, and M. C. Brodahl. 2000. 'Evaluating Farmer Defined Management Zone Maps for Variable Rate Fertilizer Application'. *Precision Agriculture* 2(2):201–15. Retrieved (http://dx.doi.org/10.1023/A%3A1011481832064).

Fraisse, C. W., K. A. Sudduth, and N. R. Kitchen. 2001. 'Delineation of Site-Specific Management Zones by Unsupervised Classification of Topographic Attributes and Soil Electrical Conductivity'. *Transactions of the ASAE* 44(1):155.

Fridgen, Jon J. et al. 2004. 'Management Zone Analyst (MZA)'. *Agron. J.* 96(1):100–108. Retrieved (https://www.agronomy.org/publications/aj/abstracts/96/1/100).

Fritzmeier-Umwelttechnik. 2016. 'Brochure Your System for Intelligent Crop Management' edited by F. U. GmbH. Retrieved (http://www.fritzmeier-umwelttechnik.com).

GDAL/OGR contributors. 2018. 'GDAL/OGR Geospatial Data Abstraction Software Library'. *Open Source Geospatial Foundation*. Retrieved 1 July 2018 (http://gdal.org).

Geladi, Paul, Hans Isaksson, Lennart Lindqvist, Svante Wold, and Kim Esbensen. 1989. 'Principal Component Analysis of Multivariate Images'. *Chemometrics and Intelligent Laboratory Systems* 5(3):209–20. Retrieved (http://www.sciencedirect.com/science/article/pii/0169743989800498).

Geonics Ltd. 1980. 'Electric Conductivity of Soil and Rocks' edited by Geonics Ltd. 20.

Geonics Ltd. 2008. 'EM38-MK2 Ground Conductivity Meter Operating Manual' edited by Geonics Ltd. 40.

Gili, Adriana, Cristian Álvarez, Ramiro Bagnato, and Elke Noellemeyer. 2017. *Comparison of Three Methods for Delineating Management Zones for Site-Specific Crop Management*.

Gili, Adriana Anahí. 2013. 'Modelación de La Variación Espacial de Variables Edáficas y Su Aplicación En El Diseño de Planes de Muestreo de Suelos'.

Google. 2018. 'Google Maps'. Retrieved 4 July 2018 (https://www.google.nl/maps/).

Graesser, Jordan and Navin Ramankutty. 2017. 'Detection of Cropland Field Parcels from Landsat Imagery'. *Remote Sensing of Environment* 201:165–80. Retrieved (http://www.sciencedirect.com/science/article/pii/S0034425717303930).

Grisso, Robert et al. 2009. *Precision Farming Tools: Soil Electrical Conductivity*. Virginia Polytechnic Institute and State University.

Hackeloeer, Andreas, Klaas Klasing, Jukka M. Krisp, and Liqiu Meng. 2014. 'Georeferencing: A Review of Methods and Applications'. *Annals of GIS* 20(1):61–69. Retrieved (https://doi.org/10.1080/19475683.2013.868826).

Hadjimitsis, Diofantos G. et al. 2013. *Detection of Water Pipes and Leakages in Rural Water Supply Networks Using Remote Sensing Techniques*. Retrieved (http://www.intechopen.com/books/export/citation/EndNote/remote-sensing-of-environment-integrated-approaches/detection-of-water-pipes-and-leakages-in-rural-water-supply-networks-using-remote-sensing-techniques).

Haghverdi, Amir, Brian G. Leib, Robert A. Washington-Allen, Paul D. Ayers, and Michael J. Buschermohle. 2015. 'Perspectives on Delineating Management Zones for Variable Rate Irrigation'. *Computers and Electronics in Agriculture* 117:154–67. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169915002264).

Hartigan, John A. and Manchek A. Wong. 1979. 'Algorithm AS 136: A k-Means Clustering Algorithm'. *Applied Statistics* 100–108.

Heijting, S., S. De Bruin, and A. K. Bregt. 2011. 'The Arable Farmer as the Assessor of Within-Field Soil Variation'. *Precision Agriculture* 12(4):488–507. Retrieved (http://dx.doi.org/10.1007/s11119-010-9197-y).

Hiemstra, Paul H., Edzer J. Pebesma, Chris J. W. Twenhöfel, and Gerard B. M. Heuvelink. 2009. 'Real-Time Automatic Interpolation of Ambient Gamma Dose Rates from the Dutch Radioactivity Monitoring Network'. *Computers & Geosciences* 35(8):1711–21. Retrieved (http://www.sciencedirect.com/science/article/pii/S0098300409000867).

Hively, W. Dean et al. 2011. 'Use of Airborne Hyperspectral Imagery to Map Soil Properties in Tilled Agricultural Fields'. *Applied and Environmental Soil Science* 2011:13. Retrieved (http://dx.doi.org/10.1155/2011/358193).

Holzapfel, C. B. et al. 2009. 'Estimating Canola (Brassica Napus L.) Yield Potential Using an Active Optical Sensor'. *Canadian Journal of Plant Science. Revue Canadienne de Phytotechnie* 89(6):1149–60. Retrieved (http://europepmc.org/abstract/AGR/IND44297825).

Hutcheson, G. D. and N. Sofroniou. 1999. *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*. SAGE Publications. Retrieved (https://books.google.nl/books?id=NiF4--8lvf0C).

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer.

Johnson, Cinthia K., David A. Mortensen, Brian J. Wienhold, John F. Shanahan, and John W. Doran. 2003. 'Site-Specific Management Zones Based on Soil Electrical Conductivity in a Semiarid Cropping System'. *Agron. J.* 95(2):303–15. Retrieved (https://www.agronomy.org/publications/aj/abstracts/95/2/303).

Jolliffe, Ian. 2014. 'Principal Component Analysis'. in *Wiley StatsRef: Statistics Reference Online*. American Cancer Society. Retrieved (https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06472).

Kadaster. 2018a. 'Nationaal Georegister'. Retrieved 1 July 2018 (http://nationaalgeoregister.nl/).

Kadaster. 2018b. 'Publieke Dienstverlening Op de Kaart'. Retrieved 1 July 2018 (https://www.pdok.nl/).

Kadaster. 2018c. 'Tijdreis over 200 Jaar Geografie'. Retrieved 3 July 2018 (http://www.topotijdreis.nl/).

Keller, Thomas, Janine A. Sutter, Knud Nissen, and Tomas Rydberg. 2012. 'Using Field Measurement of Saturated Soil Hydraulic Conductivity to Detect Low-Yielding Zones in Three Swedish Fields'. *Soil and Tillage Research* 124(0):68–77. Retrieved (http://www.sciencedirect.com/science/article/pii/S0167198712000980).

Khanal, Sami, John Fulton, and Scott Shearer. 2017. 'An Overview of Current and Potential Applications of Thermal Remote Sensing in Precision Agriculture'. *Computers and Electronics in Agriculture* 139:22–32. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169916310225).

Kitchen, N. R. et al. 2005. *Development of a Conservation-Oriented Precision Agriculture System: Crop Production Assessment and Plan Implementation*.

Kitchen, N. R., D. F. Hughes, K. A. Sudduth, and S. J. Birrell. 1996. 'Comparison of Variable Rate to Single Rate Nitrogen Fertiliser Application: Corn Production and Residual Soil NO3-N'. *Site-Specific Management for Agricultural Systems* 427–42.

Kooistra, L., E. A. Beza, J. Verbesselt, J. van den Borne, and W. van der Velde. 2012. 'Integrating Remote-, Close Range- and in-Situ Sensing for High-Frequency Observation of Crop Status to Support Precision Agriculture'. Pp. 15–20 in *Proceedings Sensing a Changing World, Wageningen, The Netherlands, 9 - 11 May, 2012*. Wageningen: Wageningen University. Retrieved (http://edepot.wur.nl/242658).

Kooistra, Lammert. 2011. *Verificatie Remote versus near Sensing Voor Toepassingen in Precisie Landbouw*. Wageningen: Wageningen University. Retrieved (http://edepot.wur.nl/191040).

Kooistra, Lammert, Harm Bartholomeus, Peter Lerink, and Erik van Valkengoed. 2011. *Plaatsspecifiek Perceelmanagement van de Kaart*. Wageningen: Wageningen University and Research Centre.

Kuang, B. et al. 2012. 'Sensing Soil Properties in the Laboratory, in Situ, and on-Line: A Review'. *Advances in Agronomy* 114:155–223. Retrieved (http://edepot.wur.nl/185800).

Li, Fei et al. 2014. 'Improving Estimation of Summer Maize Nitrogen Status with Red Edge-Based Spectral Vegetation Indices'. *Field Crops Research* 157:111–23. Retrieved (http://www.sciencedirect.com/science/article/pii/S0378429013004322).

Lillesand, T. M., R. W. Kiefer, and J. W. Chipman. 2008. *Remote Sensing and Image Interpretation*. 6th ed. John Wiley & Sons.

López-Lozano, R., M. A. Casterad, and J. Herrero. 2010. 'Site-Specific Management Units in a Commercial Maize Plot Delineated Using Very High Resolution Remote Sensing and Soil Properties Mapping'. *Computers and Electronics in Agriculture* 73(2):219–29. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169910000876).

Lund, E. D., C. D. Christy, and P. E. Drummond. 1999. 'Practical Applications of Soil Electrical Conductivity Mapping'. 9.

Ma, Ruijun, Alex Mcbratney, B. Whelan, Budiman Minasny, and Michael Short. 2011. *Comparing Temperature Correction Models for Soil Electrical Conductivity Measurement*.

Maestrini, Bernardo and Bruno Basso. 2018. 'Predicting Spatial Patterns of Within-Field Crop Yield Variability'. *Field Crops Research* 219:106–12. Retrieved (http://www.sciencedirect.com/science/article/pii/S0378429017315435).

Masini, Nicola and Rosa Lasaponara. 2007. 'Investigating the Spectral Capability of QuickBird Data to Detect Archaeological Remains Buried under Vegetated and Not Vegetated Areas'. *Journal of Cultural Heritage* 8(1):53–60. Retrieved (http://www.sciencedirect.com/science/article/pii/S1296207406001087).

McBratney, Alex, Brett Whelan, Tihomir Ancev, and Johan Bouma. 2005. 'Future Directions of Precision Agriculture'. *Precision Agriculture* 6(1):7–23.

McDermid, G. J. et al. 2009. 'Remote Sensing and Forest Inventory for Wildlife Habitat Assessment'. *Forest Ecology and Management* 257(11):2262–69. Retrieved (http://www.sciencedirect.com/science/article/pii/S0378112709001595).

McNeill, J. D. 1980. *Electromagnetic Terrain Conductivity Measurement at Low Induction Numbers, Technical Note TN-6*. Geonics Ltd.

Van Meirvenne, Marc et al. 2012. 'Key Variables for the Identification of Soil Management Classes in the Aeolian Landscapes of North–west Europe'. *Geoderma* 199(0):99–105. Retrieved (http://www.sciencedirect.com/science/article/pii/S001670611200287X).

Miller, Jeff. 1991. 'Short Report: Reaction Time Analysis with Outlier Exclusion: Bias Varies with Sample Size'. *The Quarterly Journal of Experimental Psychology Section A* 43(4):907–12. Retrieved (https://doi.org/10.1080/14640749108400962).

Milne, A. E., R. Webster, D. Ginsburg, and D. Kindred. 2012. 'Spatial Multivariate Classification of an Arable Field into Compact Management Zones Based on Past Crop Yields'. *Computers and Electronics in Agriculture* 80(0):17–30. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169911002353).

Mohanty, Hrushikesha, Prachet Bhuyan, and Deepak Chenthati. 2015. *Big Data: A Primer*. Springer.

Molin, José Paulo and Cesar Nunes de Castro. 2008. ‘Establishing Management Zones Using Soil Electrical Conductivity and Other Soil Properties by the Fuzzy Clustering Technique’. *Scientia Agricola* 65(6):567–73.

Moral, F. J., J. M. Terrón, and J. R. Marques da Silva. 2010. ‘Delineation of Management Zones Using Mobile Measurements of Soil Apparent Electrical Conductivity and Multivariate Geostatistical Techniques’. *Soil and Tillage Research* 106(2):335–43. Retrieved (http://www.sciencedirect.com/science/article/pii/S0167198709002268).

Moran, M. S., Y. Inoue, and E. M. Barnes. 1997. ‘Opportunities and Limitations for Image-Based Remote Sensing in Precision Crop Management’. *Remote Sensing of Environment* 61(3):319–46. Retrieved (http://www.sciencedirect.com/science/article/pii/S003442579700045X).

Mulla, David J. 2012. ‘Twenty Five Years of Remote Sensing in Precision Agriculture: Key Advances and Remaining Knowledge Gaps’. *Biosystems Engineering* 114(4):358–71. Retrieved (http://www.sciencedirect.com/science/article/pii/S1537511012001419).

Murali, Saritha and V. K. Govindan. 2013. ‘Shadow Detection and Removal from a Single Image Using LAB Color Space’. *Cybernetics and Information Technologies* 13(1):95–103.

Nagy, Attila, János Fehér, and János Tamás. 2018. ‘Wheat and Maize Yield Forecasting for the Tisza River Catchment Using MODIS NDVI Time Series and Reported Crop Statistics’. *Computers and Electronics in Agriculture* 151:41–49. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169918303922).

El Nahry, A. H., R. R. Ali, and A. A. El Baroudy. 2011. ‘An Approach for Precision Farming under Pivot Irrigation System Using Remote Sensing and GIS Techniques’. *Agricultural Water Management* 98(4):517–31. Retrieved (http://www.sciencedirect.com/science/article/pii/S0378377410003185).

Naturalis. 2018. ‘Geologie van Nederland’. *Leiden*. Retrieved 12 June 2018 (http://www.geologievannederland.nl/).

Nawar, Said, Ronald Corstanje, Graham Halcro, David Mulla, and Abdul M. Mouazen. 2017. ‘Delineation of Soil Management Zones for Variable-Rate Fertilization: A Review’. Pp. 175–245 in *Advances in Agronomy*, vol. 143. Elsevier.

Oliver, Y. M., M. J. Robertson, and M. T. F. Wong. 2010. ‘Integrating Farmer Knowledge, Precision Agriculture Tools, and Crop Simulation Modelling to Evaluate Management Options for Poor-Performing Patches in Cropping Fields’. *European Journal of Agronomy* 32(1):40–50. Retrieved (http://www.sciencedirect.com/science/article/pii/S1161030109000410).

Osborne, Jason W. and Amy Overbay. 2004. ‘The Power of Outliers (and Why Researchers Should Always Check for Them)’. *Practical Assessment, Research & Evaluation* 9(6):1–12.

Otsu, N. 1979. ‘A Threshold Selection Method from Gray-Level Histograms’. *IEEE Transactions on Systems, Man, and Cybernetics* 9(1):62–66.

Ott, R. L. and M. T. Longnecker. 2015. *An Introduction to Statistical Methods and Data Analysis*. Cengage Learning. Retrieved (https://books.google.nl/books?id=VAuyBQAAQBAJ).

Pádua, Luís et al. 2017. ‘Very High Resolution Aerial Data to Support Multi-Temporal Precision Agriculture Information Management’. *Procedia Computer Science* 121:407–14. Retrieved (http://www.sciencedirect.com/science/article/pii/S1877050917322482).

Papadopoulos, N. and A. Sarris. 2006. ‘Integrated Geophysical Survey to Characterize the Subsurface Properties below and around the Area of Saint Andreas Church (Loutraki, Greece)’. Retrieved (http://www.chnt.at/proceedings-chnt-14/).

Parcak, Sarah H. 2009. *Satellite Remote Sensing for Archaeology*. Routledge.

Paul Obade, Vincent de and Rattan Lal. 2013. ‘Assessing Land Cover and Soil Quality by Remote Sensing and Geographical Information Systems (GIS)’. *CATENA* 104(0):77–92. Retrieved (http://www.sciencedirect.com/science/article/pii/S0341816212002202).

Pierce, Francis J. and Peter Nowak. 1999. ‘Aspects of Precision Agriculture’. Pp. 1–85 in *Advances in Agronomy*, vol. Volume 67, edited by L. S. Donald. Academic Press. Retrieved (http://www.sciencedirect.com/science/article/pii/S0065211308605131).

QGIS. 2018. ‘QGIS - A Free and Open Source Geographic Information System’. Retrieved 1 July 2018 (https://qgis.org/en/site/).

Qin, Yin-Shi, Shui-Fa Sun, Xian-Bing Ma, Song Hu, and Bang-Jun Lei. 2013. *A Shadow Removal Algorithm for ViBe in HSV Color Space*.

Raizman, E. A., H. Barner Rasmussen, L. E. King, F. W. Ihwagi, and I. Douglas-Hamilton. 2013. ‘Feasibility Study on the Spatial and Temporal Movement of Samburu’s Cattle and Wildlife in Kenya Using GPS Radio-Tracking, Remote Sensing and GIS’. *Preventive Veterinary Medicine* 111(1):76–80. Retrieved (http://www.sciencedirect.com/science/article/pii/S0167587713001451).

Roberts, Roland K. et al. 2004. ‘Adoption of Site-Specific Information and Variable-Rate Technologies in Cotton Precision Farming’. *Journal of Agricultural and Applied Economics* 36(01). Retrieved (http://ideas.repec.org/a/ags/joaaec/42943.html).

Rokhmana, Catur Aries. 2015. ‘The Potential of UAV-Based Remote Sensing for Supporting Precision Agriculture in Indonesia’. *Procedia Environmental Sciences* 24:245–53. Retrieved (http://www.sciencedirect.com/science/article/pii/S1878029615001000).

Roy, D. P. and L. Yan. 2018. ‘Robust Landsat-Based Crop Time Series Modelling’. *Remote Sensing of Environment*. Retrieved (http://www.sciencedirect.com/science/article/pii/S003442571830316X).

RStudio. 2018. ‘RStudio’. Retrieved 1 July 2018 (https://www.rstudio.com/).

Saey, Timothy et al. 2013. ‘Identifying Soil Patterns at Different Spatial Scales with a Multi-Receiver EMI Sensor’. *Soil Science Society of America Journal* 77:382–90. Retrieved (http://dx.doi.org/10.2136/sssaj2012.0276).

Sanches, Guilherme M., Paulo S. G. Magalhães, Armando Z. Remacre, and Henrique C. J. Franco. 2018. ‘Potential of Apparent Soil Electrical Conductivity to Describe the Soil PH and Improve Lime Application in a Clayey Soil’. *Soil and Tillage Research* 175:217–25. Retrieved (http://www.sciencedirect.com/science/article/pii/S0167198717301770).

Schabenberger, Oliver and Francis J. Pierce. 2001. *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC press.

Schaepman, Michael E. et al. 2009. ‘Earth System Science Related Imaging Spectroscopy—An Assessment’. *Remote Sensing of Environment* 113:S123–37.

Schans, D. A. van der and W. van den Berg. 2013. *Testen, Validatie En Toepassing van Het Veris-Sensorplatform Veldanalyse van Twee Percelen Op Veenkoloniale Grond*. 526, , : PPO AGV. Retrieved (http://edepot.wur.nl/259239).

Schenatto, Kelyn et al. 2017. ‘Normalization of Data for Delineating Management Zones’. *Computers and Electronics in Agriculture* 143:238–48. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169917302715).

Schirrmann, Michael, Robin Gebbers, Eckart Kramer, and Jan Seidel. 2011. ‘Soil PH Mapping with an On-The-Go Sensor’. *Sensors* 11(1).

Scudiero, Elia et al. 2013. ‘Delineation of Site-Specific Management Units in a Saline Region at the

Venice Lagoon Margin, Italy, Using Soil Reflectance and Apparent Electrical Conductivity'. *Computers and Electronics in Agriculture* 99(0):54–64. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169913002019).

SenseFly. 2018. 'Https://Www.Sensefly.Com/Drones/Ebee.Html'. *Lausanne, Switzerland*. Retrieved (https://www.sensefly.com/drones/ebee.html).

Shafranovich, Y. 2005. *Common Format and MIME Type for Comma-Separated Values (CSV) Files, RFC 4180*. Retrieved (https://www.rfc-editor.org/info/rfc4180).

Snedecor, G. W. and W. G. Cochran. 1989. *Statistical Methods*. Iowa State University Press. Retrieved (https://books.google.nl/books?id=LGikkgEACAAJ).

Stoorvogel, J. J., L. Kooistra, and J. Bouma. 2015. 'Managing Soil Variability at Different Spatial Scales as a Basis for Precision Agriculture'. Pp. 37–72 in *Soil-Specific Farming: Precision Agriculture*, *Advances in Soil Science*, edited by R. Lal and B. A. Stewart. 506, , : CRC Press. Retrieved (http://edepot.wur.nl/393927).

Suomalainen, Juha et al. 2014. 'A Lightweight Hyperspectral Mapping System and Photogrammetric Processing Chain for Unmanned Aerial Vehicles'. *Remote Sensing* 6(11).

Thessler, Sirpa, Lammert Kooistra, Frederick Teye, Hanna Huitu, and Arnold Bregt. 2011. 'Geosensors to Support Crop Production: Current Applications and User Requirements'. *Sensors* 11(7):6656–84. Retrieved (http://www.mdpi.com/1424-8220/11/7/6656).

USGS. 2018. 'Landsat'. Retrieved 6 August 2018 (https://landsat.usgs.gov/).

Vantage Agrometius. 2018. 'VERIS MSP Series Bodemscanner'. Retrieved (https://www.vantage-agrometius.nl/product/veris-msp3-bodemscanner).

Viscarra Rossel, R. A. 2011. 'Proximal Soil Spectroscopy (Keynote Presentation)' edited by R. A. Viscarra Rossel. *The Second Global Workshop on Proximal Soil Sensing* 4.

Viscarra Rossel, R. A. and A. B. McBratney. 1998. 'Laboratory Evaluation of a Proximal Sensing Technique for Simultaneous Measurement of Soil Clay and Water Content'. *Geoderma* 85(1):19–39. Retrieved (http://www.sciencedirect.com/science/article/pii/S0016706198000238).

Vitharana, Udayakantha W. A., Marc Van Meirvenne, Liesbet Cockx, and Jean Bourgeois. 2006. 'Identifying Potential Management Zones in a Layered Soil Using Several Sources of Ancillary Information'. *Soil Use and Management* 22(4):405–13.

Vitharana, Udayakantha W. A., Marc Van Meirvenne, David Simpson, Liesbet Cockx, and Josse De Baerdemaeker. 2008. 'Key Soil and Topographic Properties to Delineate Potential Management Classes for Precision Agriculture in the European Loess Area'. *Geoderma* 143(1–2):206–15. Retrieved (http://www.sciencedirect.com/science/article/pii/S0016706107003217).

Di Vittorio, Courtney A. and Aris P. Georgakakos. 2018. 'Land Cover Classification and Wetland Inundation Mapping Using MODIS'. *Remote Sensing of Environment* 204:1–17. Retrieved (http://www.sciencedirect.com/science/article/pii/S0034425717305114).

Van der Voort, D. 2016. 'Exploring the Usability of Unmanned Aerial Vehicles for Non-Destructive Phenotyping of Small-Scale Maize Breeding Trials'. Wageningen University and Research.

De Vos, J. A., P. J. T. Van Bakel, I. E. Hoving, and R. A. Smidt. 2010. 'Raising Surface Water Levels in Peat Areas with Dairy Farming: Upscaling Hydrological, Agronomical and Economic Effects from Farm-Scale to Local Scale'. *Agricultural Water Management* 97(11):1887–97. Retrieved (http://www.sciencedirect.com/science/article/pii/S0378377410002222).

Vrindts, E. et al. 2005. 'Management Zones Based on Correlation between Soil Compaction, Yield and Crop Data'. *Biosystems Engineering* 92(4):419–28. Retrieved (http://www.sciencedirect.com/science/article/pii/S1537511005001959).

Wang, Jianghao, Yong Ge, Gerard B. M. Heuvelink, Chenghu Zhou, and Dick Brus. 2012. 'Effect of the Sampling Design of Ground Control Points on the Geometric Correction of Remotely Sensed Imagery'. *International Journal of Applied Earth Observation and Geoinformation* 18:91–100. Retrieved (http://www.sciencedirect.com/science/article/pii/S0303243412000025).

Weidong, Liu et al. 2002. 'Relating Soil Surface Moisture to Reflectance'. *Remote Sensing of Environment* 81(2):238–46. Retrieved (http://www.sciencedirect.com/science/article/pii/S0034425701003479).

Weiss, Michael D. 1996. 'Precision Farming and Spatial Economic Analysis: Research Challenges and Opportunities'. *American Journal of Agricultural Economics* 78(5):1275–80. Retrieved (http://www.jstor.org/stable/1243506).

Whelan, B. M. and A. B. McBratney. 2003. 'Definition and Interpretation of Potential Management Zones in Australia'. in *Proceedings of the 11th Australian Agronomy Conference, Geelong, Victoria*.

Wold, Svante, Kim Esbensen, and Paul Geladi. 1987. 'Principal Component Analysis'. *Chemometrics and Intelligent Laboratory Systems* 2(1):37–52. Retrieved (http://www.sciencedirect.com/science/article/pii/0169743987800849).

Wolfert, Sjaak, Lan Ge, Cor Verdouw, and Marc-Jeroen Bogaardt. 2017. 'Big Data in Smart Farming – A Review'. *Agricultural Systems* 153:69–80. Retrieved (http://www.sciencedirect.com/science/article/pii/S0308521X16303754).

Yang, Chenghai, James H. Everitt, and Dale Murden. 2011. 'Evaluating High Resolution SPOT 5 Satellite Imagery for Crop Identification'. *Computers and Electronics in Agriculture* 75(2):347–54. Retrieved (http://www.sciencedirect.com/science/article/pii/S0168169910002632).

Zhang, Chunhua and JohnM Kovacs. 2012. 'The Application of Small Unmanned Aerial Systems for Precision Agriculture: A Review'. *Precision Agriculture* 13(6):693–712. Retrieved (http://dx.doi.org/10.1007/s11119-012-9274-5).

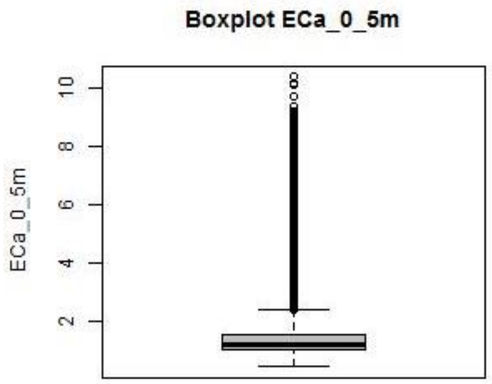
Zhang, Xiaodong, Lijian Shi, Xinhua Jia, George Seielstad, and Craig Helgason. 2010. 'Zone Mapping Application for Precision-Farming: A Decision Support Tool for Variable Rate Application'. 11(2):103–14. Retrieved (http://dx.doi.org/10.1007/s11119-009-9130-4).

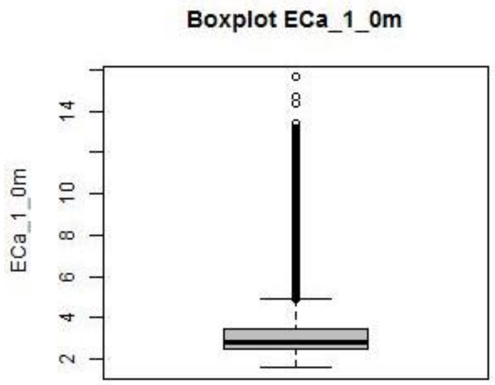
Appendices

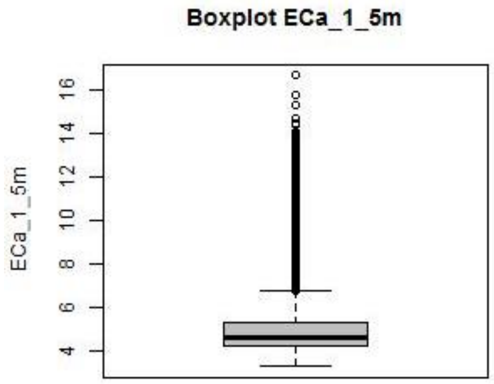
Appendix 1. Overview initially available datasets

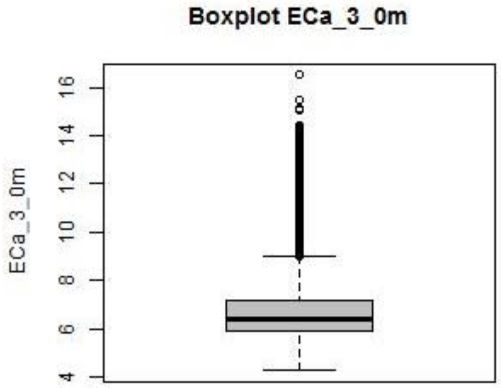
Dataset	Format	Ref. system	Contents	Meas. unit	Acquis. date	Provided by
Aerial image bare soil	GeoTIFF with spatial resolution of 0.035m	WGS84-UTM31	RGB colour bands	-	13-4-2015	Aurea imaging
Aerial image potato crops	GeoTIFF with spatial resolution of 0.132m	WGS84-UTM31	Spectral bands NIR, Red Edge, Red, Green; vegetation indices WDVI, NDVI, NDRE	-	21-8-2015	Aurea imaging
Apparent electric conductivity	CSV with 54312 spatial point observations (for 2 fields)	WGS84 standard	Eca at soil depths of 0.5m, 1.0m, 1,5m and 3.0m	mS/m	20-3-2015	Jacob Van den Borne
Crop yield	CSV with 29987 spatial point observations	WGS84 standard	Potato crop yield	ton/ha	4-10-2015	Jacob Van den Borne
Elevation	CSV with 133 spatial point observations	RDNew	Elevation	m + NAP	20-3-2015	Marnix Van den Brande via Lammert Kooistra
Soil depth A-horizon	CSV with 133 spatial point observations	RDNew	Soil depth	m - ground level	20-3-2015	Marnix Van den Brande via Lammert Kooistra
Soil organic matter	CSV with 25 spatial point observations	RDNew	SOM	%	20-3-2015	Marnix Van den Brande via Lammert Kooistra

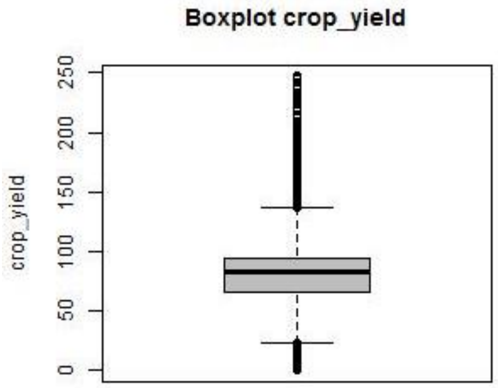
Appendix 2. Boxplots spatial point datasets

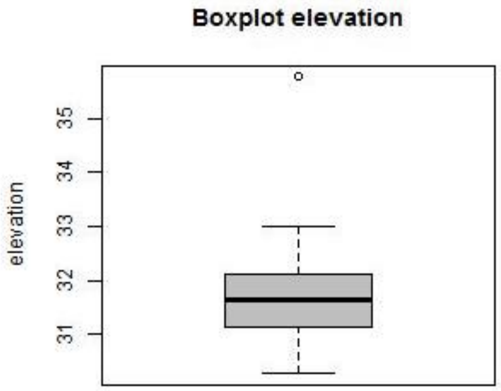

Boxplot ECa_0_5m

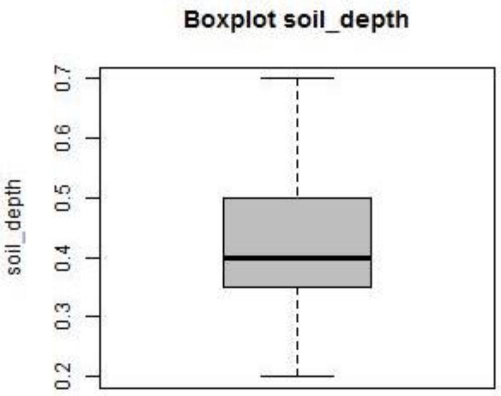

Boxplot ECa_1_0m

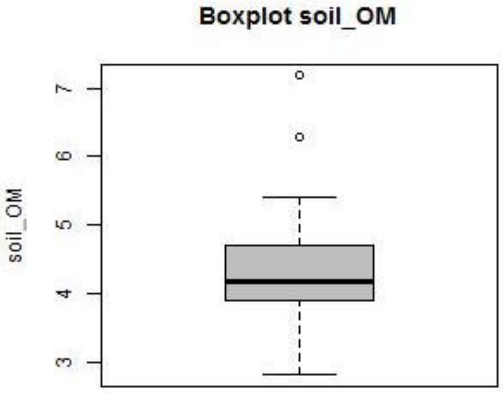

Boxplot ECa_1_5m


Boxplot ECa_3_0m


Boxplot crop_yield

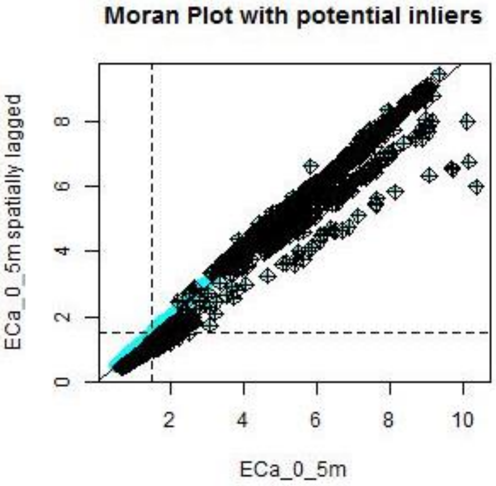

Boxplot elevation

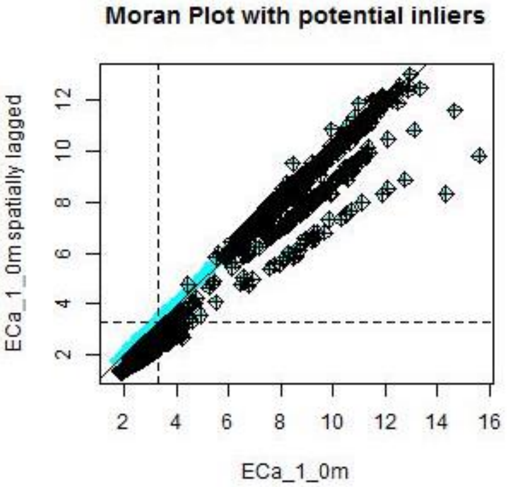

Boxplot soil_depth

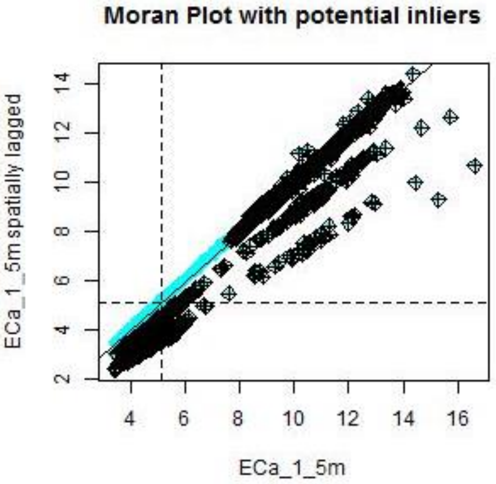

Boxplot soil_OM

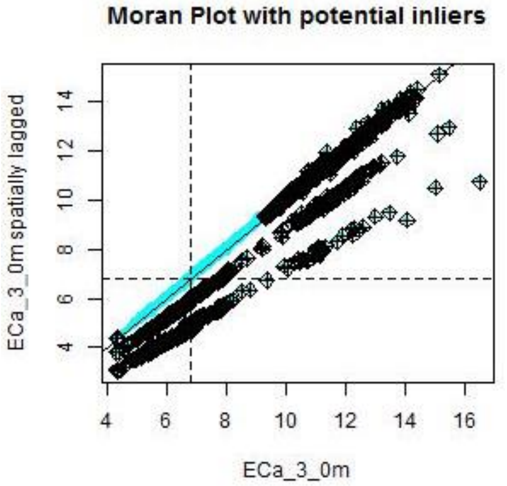
Appendix 3. Moran plots EC and crop yield

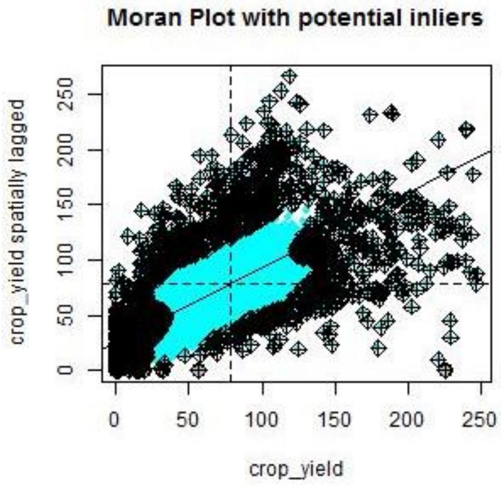
The x-axis represents a spatial point variable and the y-axis indicates the spatial lag (or distance) between point pairs for that variable. Black dots represent potential inliers, while blue dots indicate observations that do not significantly deviate from their neighbours.



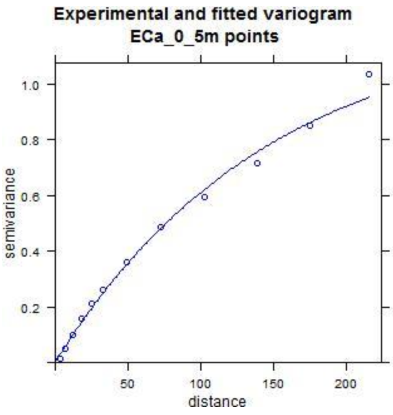


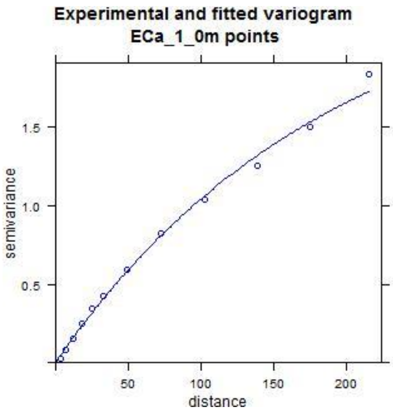


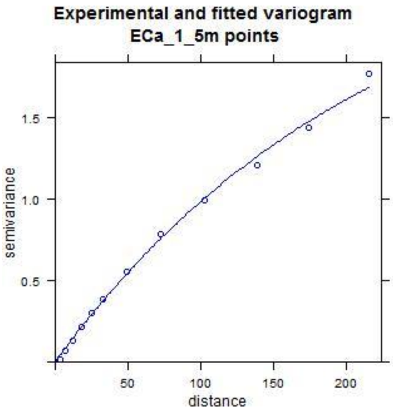


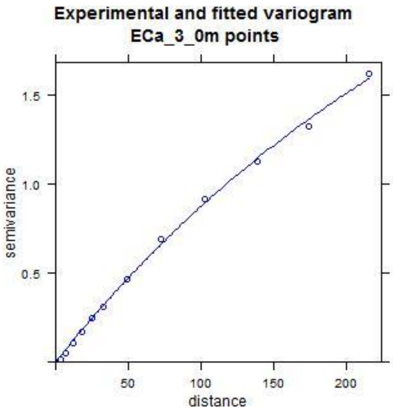


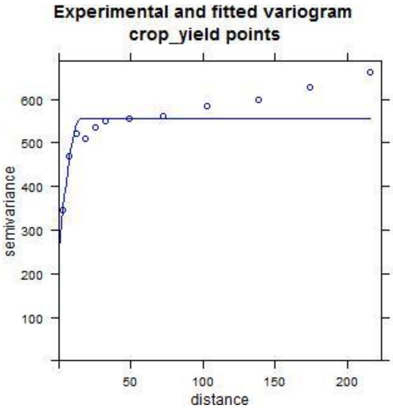
Appendix 4. Variograms and variogram parameters spatial point datasets

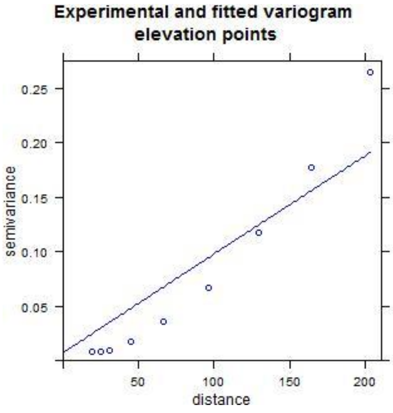


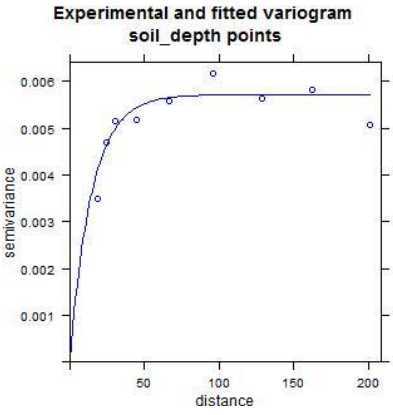


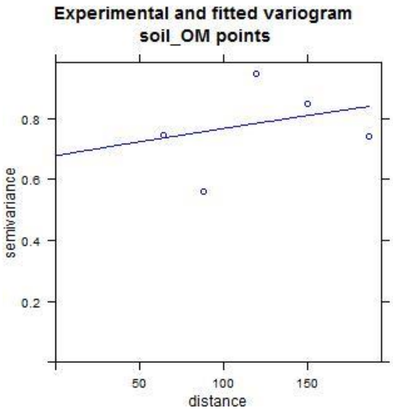








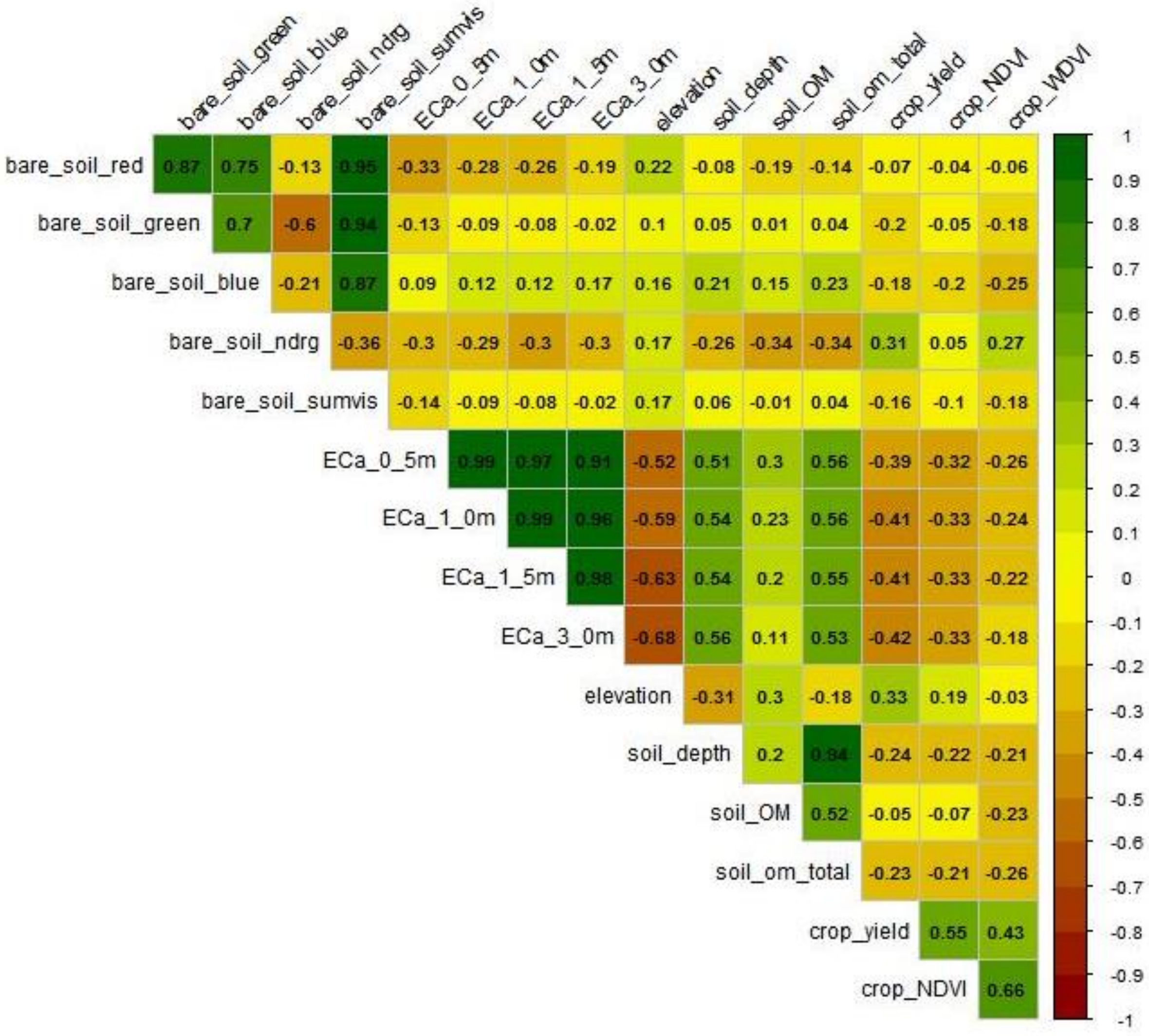




Variogram parameters

Variable	model	nugget	sill	range
crop_yield points	Sph	254.335	555.574	15.294
ECa_0_5m points	Exp	0.000	1.241	147.751
ECa_1_0m points	Exp	0.000	2.547	190.446
ECa_1_5m points	Exp	0.000	2.706	222.027
ECa_3_0m points	Exp	0.000	3.203	315.022
elevation points	Sph	0.007	1.009	1654.113
soil_depth points	Exp	0.000	0.006	14.825
soil_OM points	Exp	0.678	1.930	1347.722

Appendix 5. Correlation matrix soil and crop variables

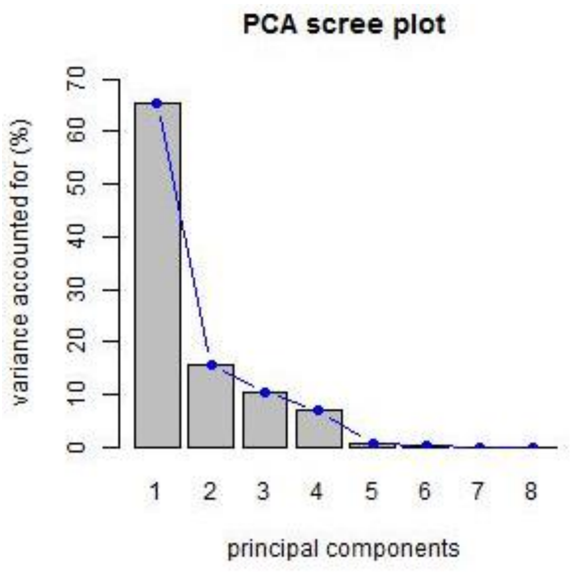


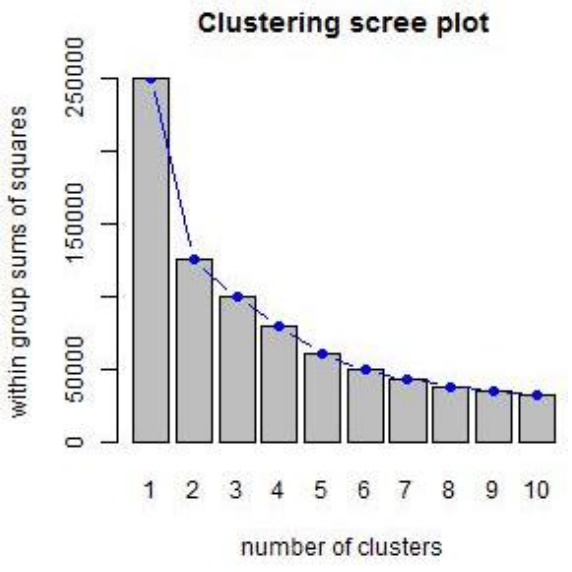
	bare_soil_green	bare_soil_blue	bare_soil_ndrg	bare_soil_sumvis	ECa_0_5m	ECa_1_0m	ECa_1_5m	ECa_3_0m	elevation	soil_depth	soil_OM	soil_om_total	crop_yield	crop_NDVI	crop_WDVI
bare_soil_red	0.87	0.75	-0.13	0.95	-0.33	-0.28	-0.26	-0.19	0.22	-0.08	-0.19	-0.14	-0.07	-0.04	-0.06
bare_soil_green		0.7	-0.6	0.94	-0.13	-0.09	-0.08	-0.02	0.1	0.05	0.01	0.04	-0.2	-0.05	-0.18
bare_soil_blue			-0.21	0.87	0.09	0.12	0.12	0.17	0.16	0.21	0.15	0.23	-0.18	-0.2	-0.25
bare_soil_ndrg				-0.36	-0.3	-0.29	-0.3	-0.3	0.17	-0.26	-0.34	-0.34	0.31	0.05	0.27
bare_soil_sumvis					-0.14	-0.09	-0.08	-0.02	0.17	0.06	-0.01	0.04	-0.16	-0.1	-0.18
ECa_0_5m						0.99	0.97	0.91	-0.52	0.51	0.3	0.56	-0.39	-0.32	-0.26
ECa_1_0m							0.99	0.96	-0.59	0.54	0.23	0.56	-0.41	-0.33	-0.24
ECa_1_5m								0.98	-0.63	0.54	0.2	0.55	-0.41	-0.33	-0.22
ECa_3_0m									-0.68	0.56	0.11	0.53	-0.42	-0.33	-0.18
elevation										-0.31	0.3	-0.18	0.33	0.19	-0.03
soil_depth											0.2	0.84	-0.24	-0.22	-0.21
soil_OM												0.52	-0.05	-0.07	-0.23
soil_om_total													-0.23	-0.21	-0.26
crop_yield														0.55	0.43
crop_NDVI															0.66

Appendix 6. Bartlett's test of Sphericity and KMO MSA tests

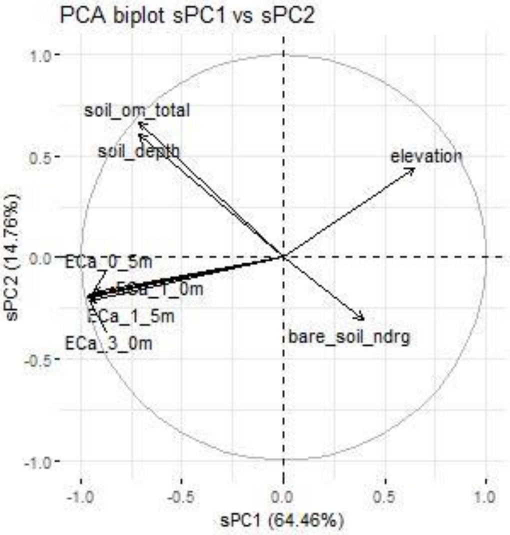
Bartlett's test of Sphericity			
bartlett_chi2	1875045.7	p-value	0.000
KMO Measure of Sample Adequacy overall			
kmo_full_overall	0.607	kmo_reduced_overall	0.766
KMO Measure of Sample Adequacy per variable			
kmo_full_bare_soil_red	0.468	kmo_reduced_bare_soil_ndrg	0.728
kmo_full_bare_soil_green	0.457	kmo_reduced_ECa_0_5m	0.775
kmo_full_bare_soil_blue	0.408	kmo_reduced_ECa_1_0m	0.828
kmo_full_bare_soil_ndrg	0.966	kmo_reduced_ECa_1_5m	0.830
kmo_full_bare_soil_sumvis	0.477	kmo_reduced_ECa_3_0m	0.795
kmo_full_ECa_0_5m	0.805	kmo_reduced_elevation	0.794
kmo_full_ECa_1_0m	0.838	kmo_reduced_soil_depth	0.653
kmo_full_ECa_1_5m	0.827	kmo_reduced_soil_om_total	0.644
kmo_full_ECa_3_0m	0.816		
kmo_full_elevation	0.770		
kmo_full_soil_depth	0.518		
kmo_full_soil_OM	0.264		
kmo_full_soil_om_total	0.548		

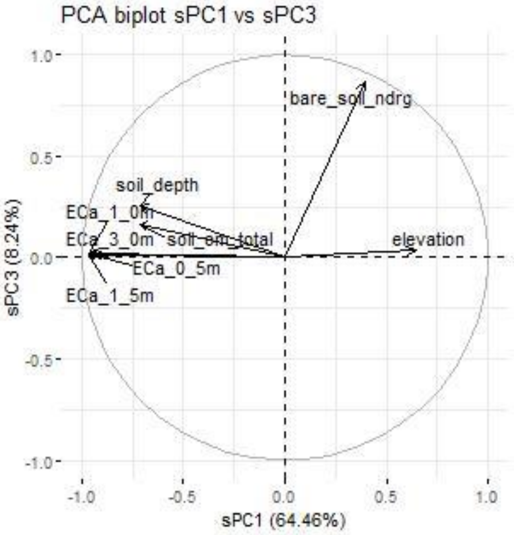
Appendix 7. Scree plots PCA and cluster analysis

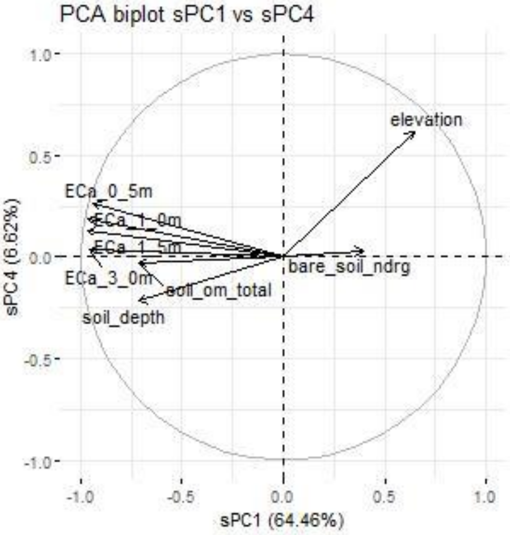


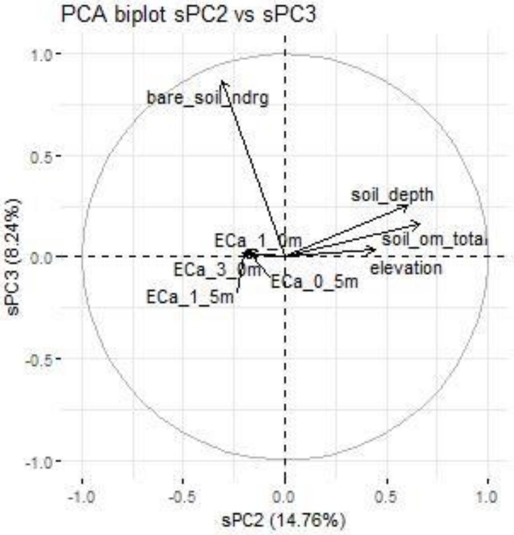


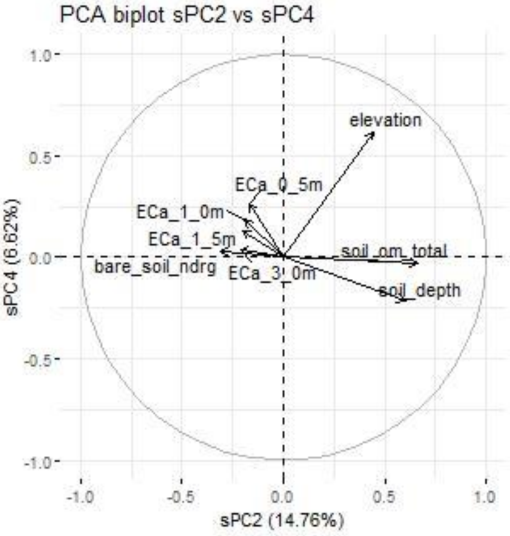
Appendix 8. Biplots PCA

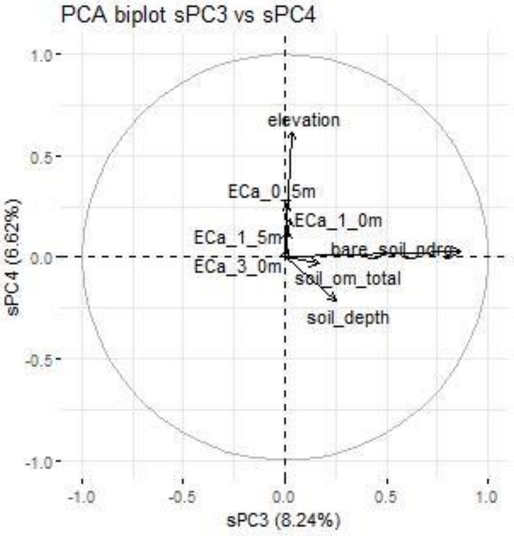












Appendix 9. Output linear models

Validation area 1: whole field excluding strip without additional fertilizer

Model structure		**AIC**			
exponential correlation with nugget		3832.8			
exponential correlation without nugget		3843.4			
spherical correlation with nugget		**3829.2**	**selected**		
spherical correlation without nugget		3858.3			
independent errors		3887.8			
ANOVA table					
	df	**F-value**	**p-value**		
(Intercept)	1	1175.710	0.000		
clusters	3	6.043	0.000		
LS means					
clusters	**lsmean**	**SE**	**df**	**lower CL**	**upper CL**
1	59.875	5.508	396	49.047	70.703
2	85.326	3.950	396	77.560	93.092
3	83.070	3.954	396	75.296	90.844
4	69.693	4.310	396	61.219	78.166
Pairwise comparisons					
contrast	**estimate**	**SE**	**df**	**t-ratio**	**p-value**
1 - 2	-25.451	6.777	396	-3.755	0.001
1 - 3	-23.195	6.755	396	-3.434	0.004
1 - 4	-9.818	6.451	396	-1.522	0.425
2 - 3	2.256	5.546	396	0.407	0.977
2 - 4	15.633	5.843	396	2.676	0.039
3 - 4	13.377	5.697	396	2.348	0.089

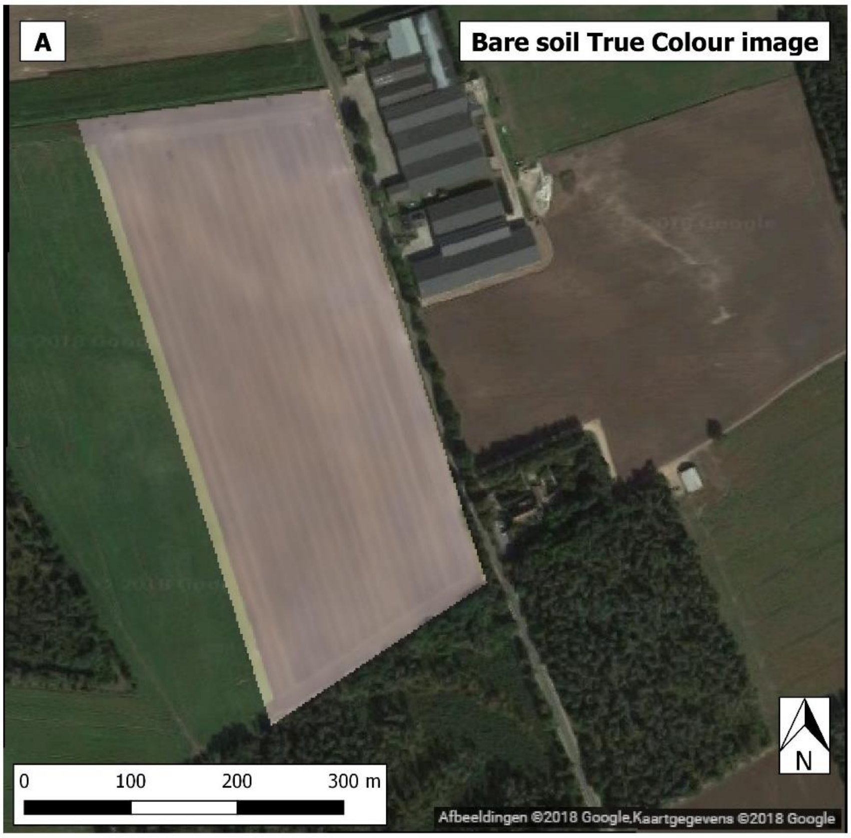
Validation area 2: upper north-west corner of the field (in fertilizer zone with 90 kg/ha initial N input)

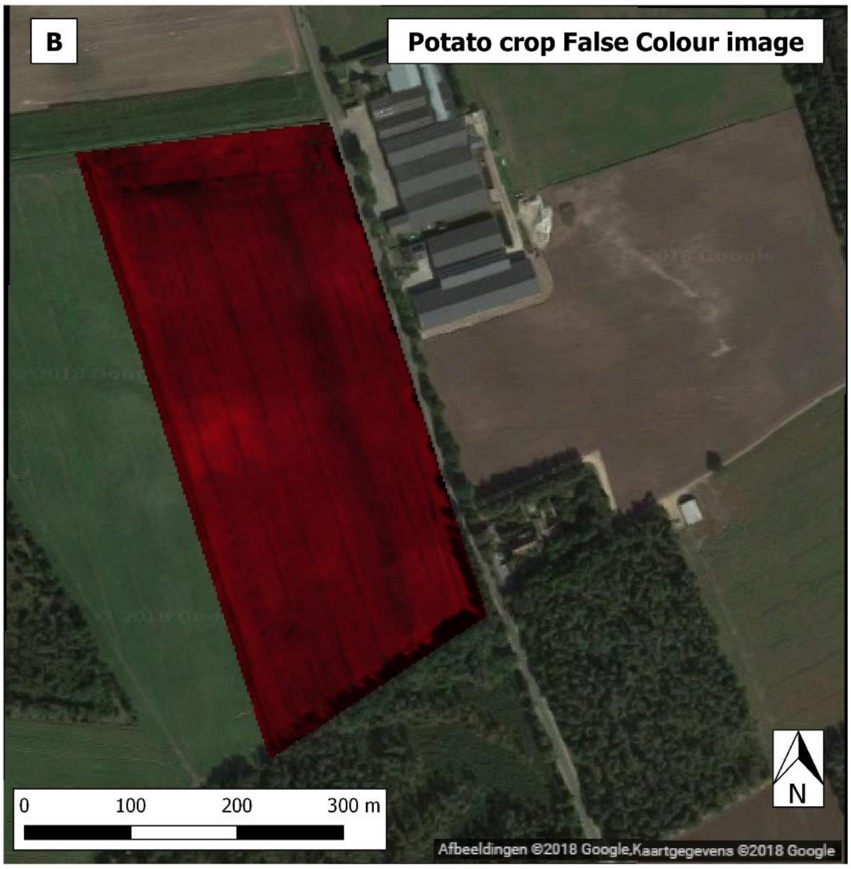
Model structure		**AIC**			
exponential correlation with nugget		2879.1			
exponential correlation without nugget		**2877.1**	**selected**		
spherical correlation with nugget		2884.7			
spherical correlation without nugget		2984.4			
independent errors		2982.4			
ANOVA table					
	df	**F-value**	**p-value**		
(Intercept)	1	552.734	0.000		
clusters	2	7.373	0.001		
LS means					
clusters	**lsmean**	**SE**	**df**	**lower CL**	**upper CL**
1	52.991	5.937	297	41.306	64.675
3	80.138	4.076	297	72.116	88.159
4	66.534	5.390	297	55.926	77.143
Pairwise comparisons					
contrast	**estimate**	**SE**	**df**	**t-ratio**	**p-value**
1 - 3	-27.147	7.178	297	-3.782	0.001
1 - 4	-13.544	7.601	297	-1.782	0.177
3 - 4	13.603	6.532	297	2.083	0.095

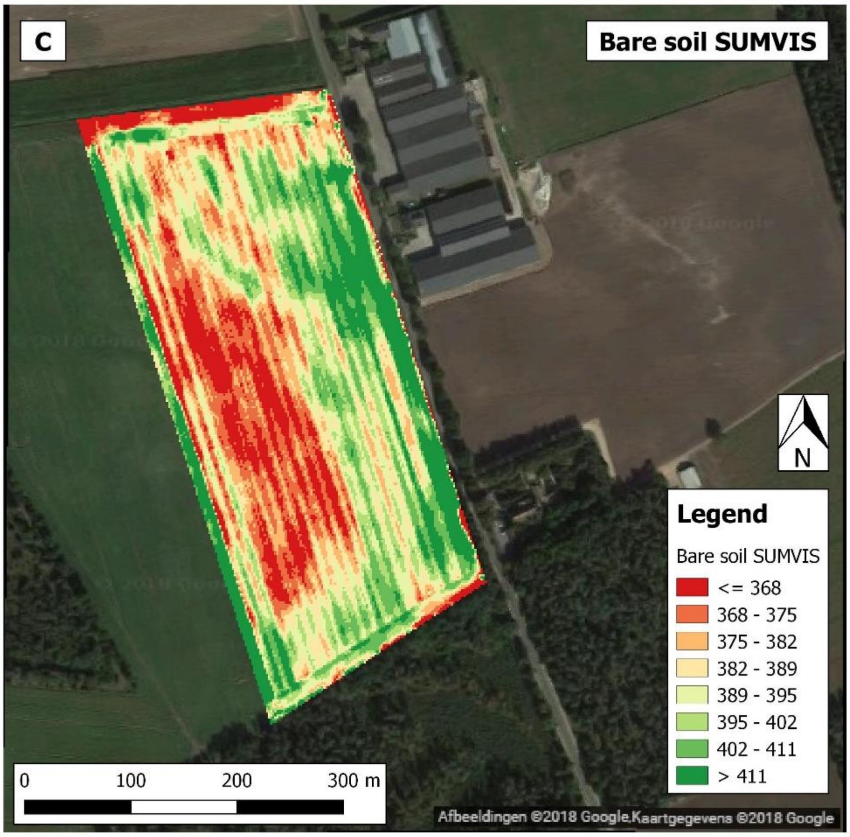
Validation area 3: mid-west of the field (in fertilizer zone with 162 kg/ha initial N input)

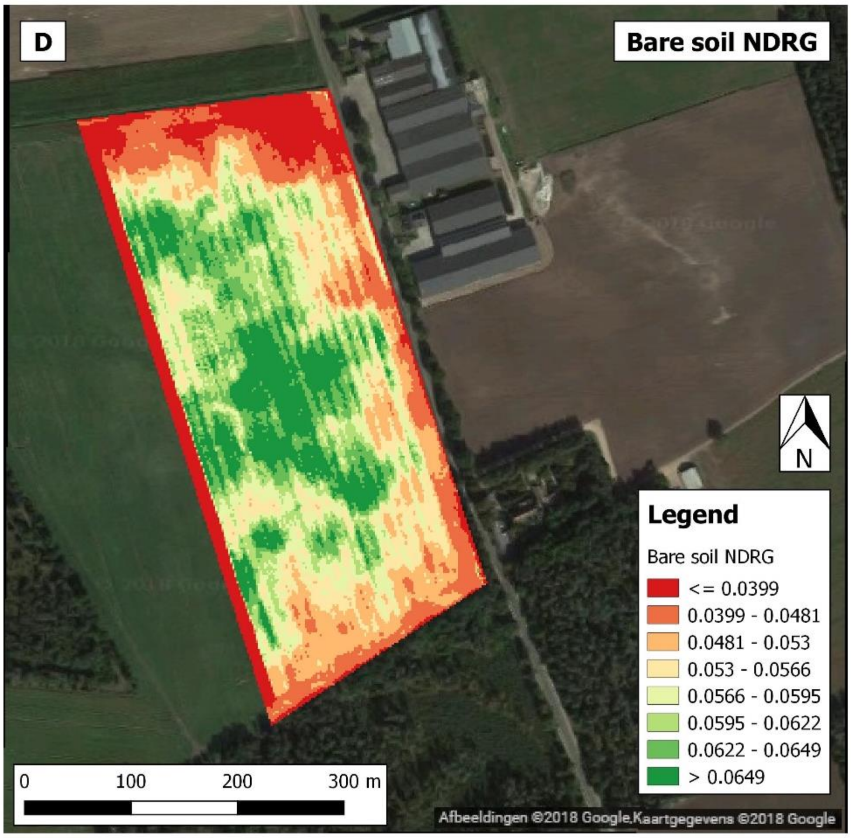
Model structure		**AIC**			
exponential correlation with nugget		1707.7			
exponential correlation without nugget		1705.7			
spherical correlation with nugget		1705.4			
spherical correlation without nugget		**1703.4**	**selected**		
independent errors		1720.1			
ANOVA table					
	df	**F-value**	**p-value**		
(Intercept)	1	3372.982	0.000		
clusters	1	2.949	0.087		
LS means					
clusters	**lsmean**	**SE**	**df**	**lower CL**	**upper CL**
2	88.251	2.103	198	84.104	92.399
3	93.580	2.304	198	89.037	98.123
Pairwise comparisons					
contrast	**estimate**	**SE**	**df**	**t-ratio**	**p-value**
2 - 3	-5.329	3.103	198	-1.717	0.087

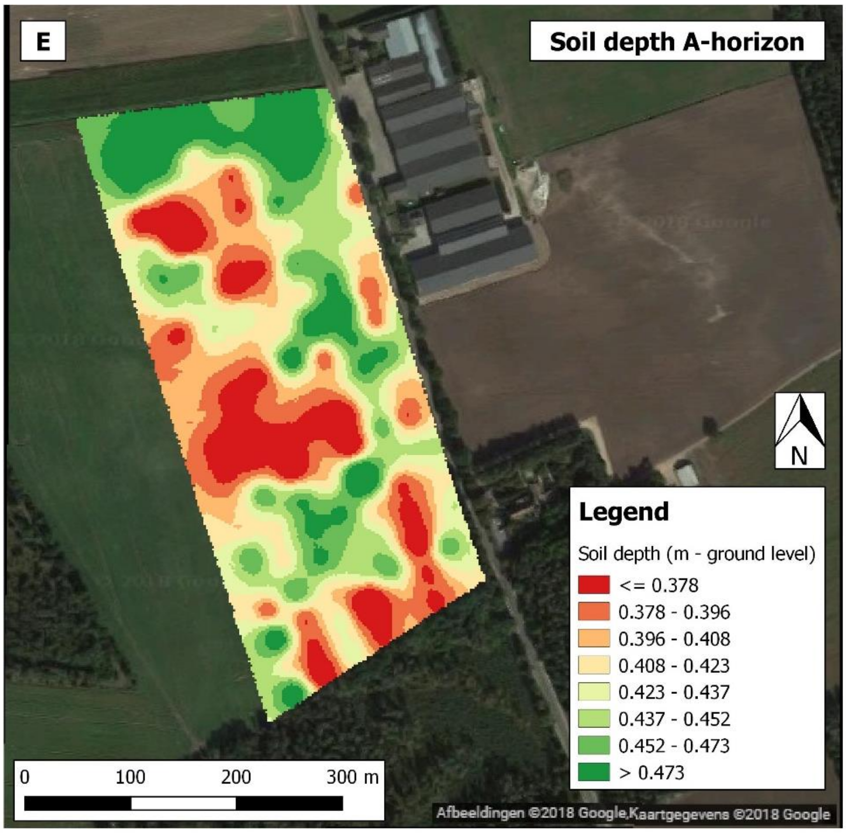
Appendix 10. Output maps soil and crop variables

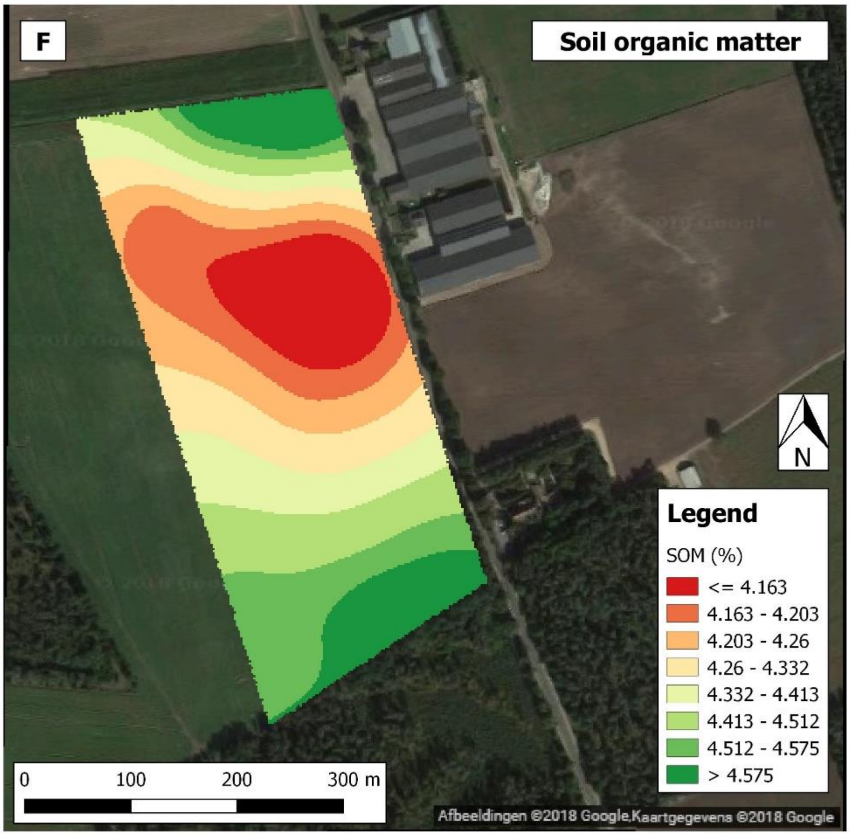

A
Bare soil True Colour image
N
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

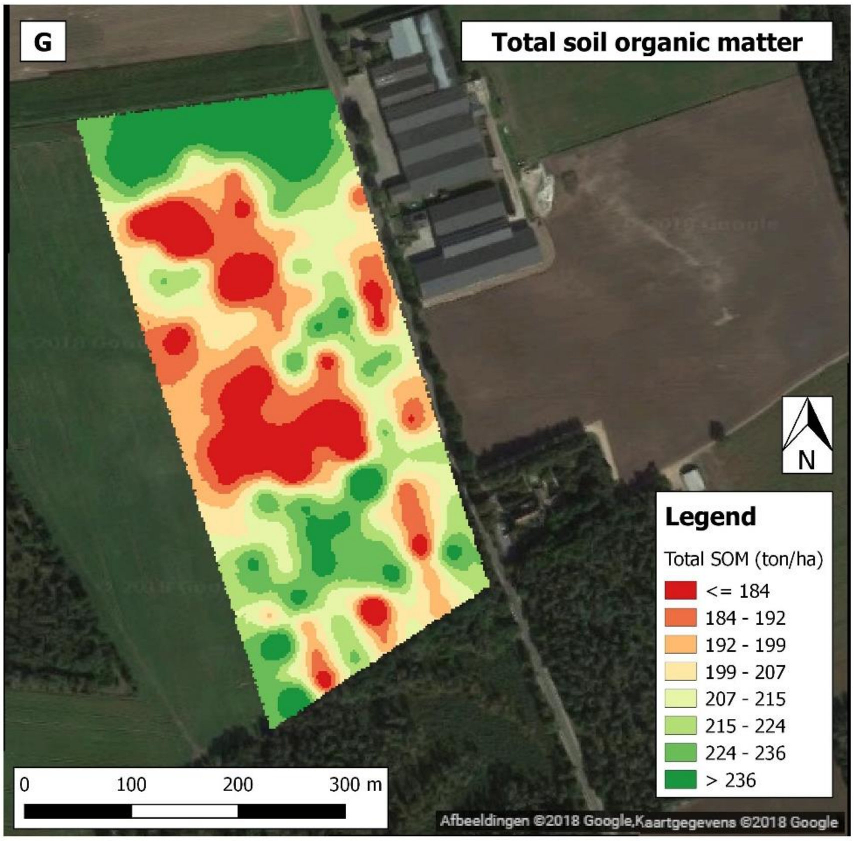

B
Potato crop False Colour image
N
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

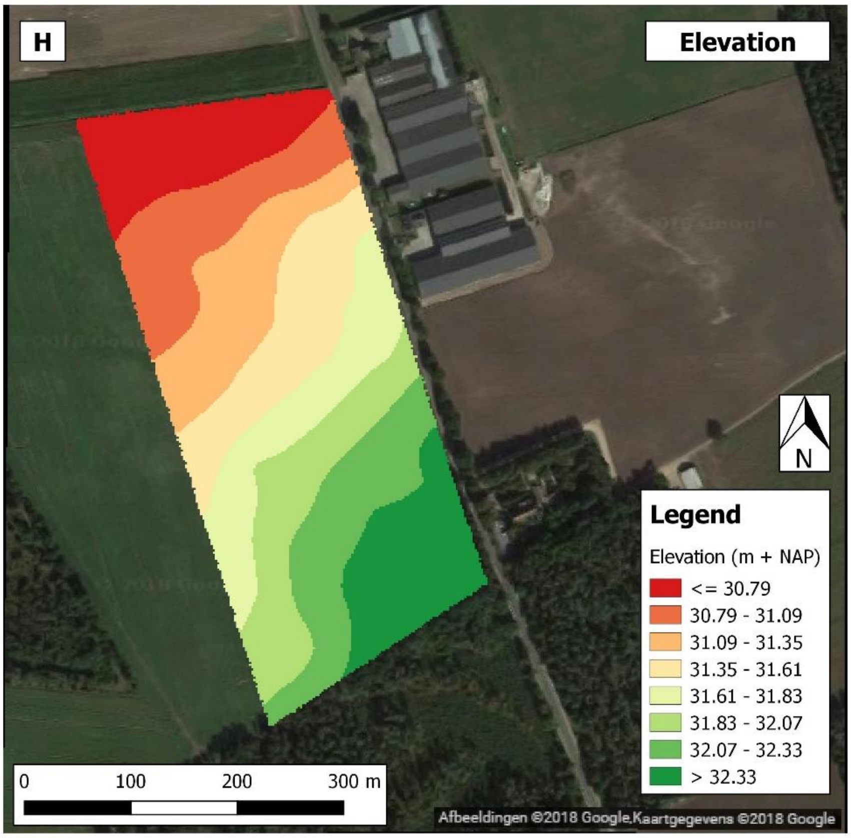

C
Bare soil SUMVIS
N
Legend
Bare soil SUMVIS
<= 368
368 - 375
375 - 382
382 - 389
389 - 395
395 - 402
402 - 411
> 411
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

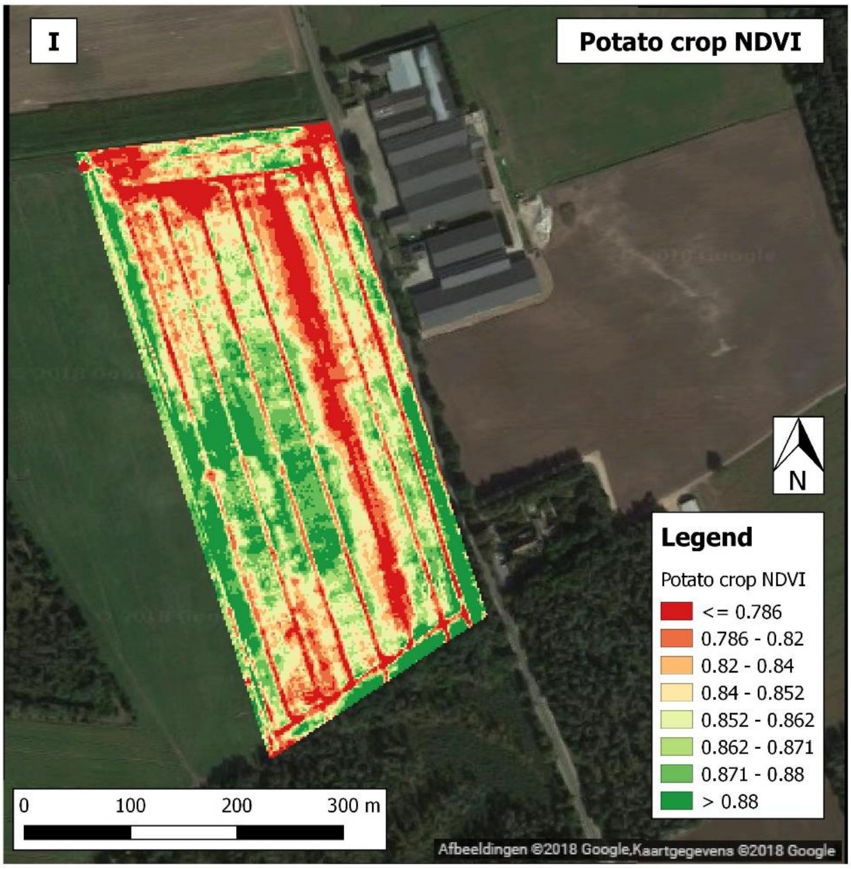

D
Bare soil NDRG
N
Legend
Bare soil NDRG
<= 0.0399
0.0399 - 0.0481
0.0481 - 0.053
0.053 - 0.0566
0.0566 - 0.0595
0.0595 - 0.0622
0.0622 - 0.0649
> 0.0649
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

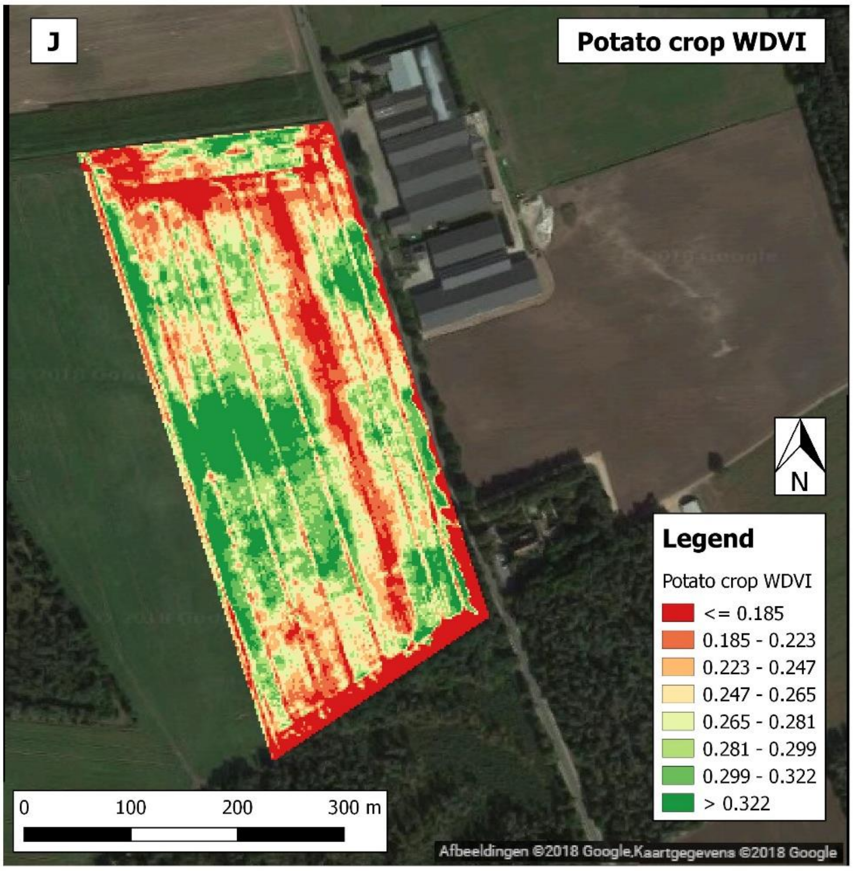

E
Soil depth A-horizon
N
Legend
Soil depth (m - ground level)
<= 0.378
0.378 - 0.396
0.396 - 0.408
0.408 - 0.423
0.423 - 0.437
0.437 - 0.452
0.452 - 0.473
> 0.473
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google


F
Soil organic matter
N
Legend
SOM (%)
<= 4.163
4.163 - 4.203
4.203 - 4.26
4.26 - 4.332
4.332 - 4.413
4.413 - 4.512
4.512 - 4.575
> 4.575
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

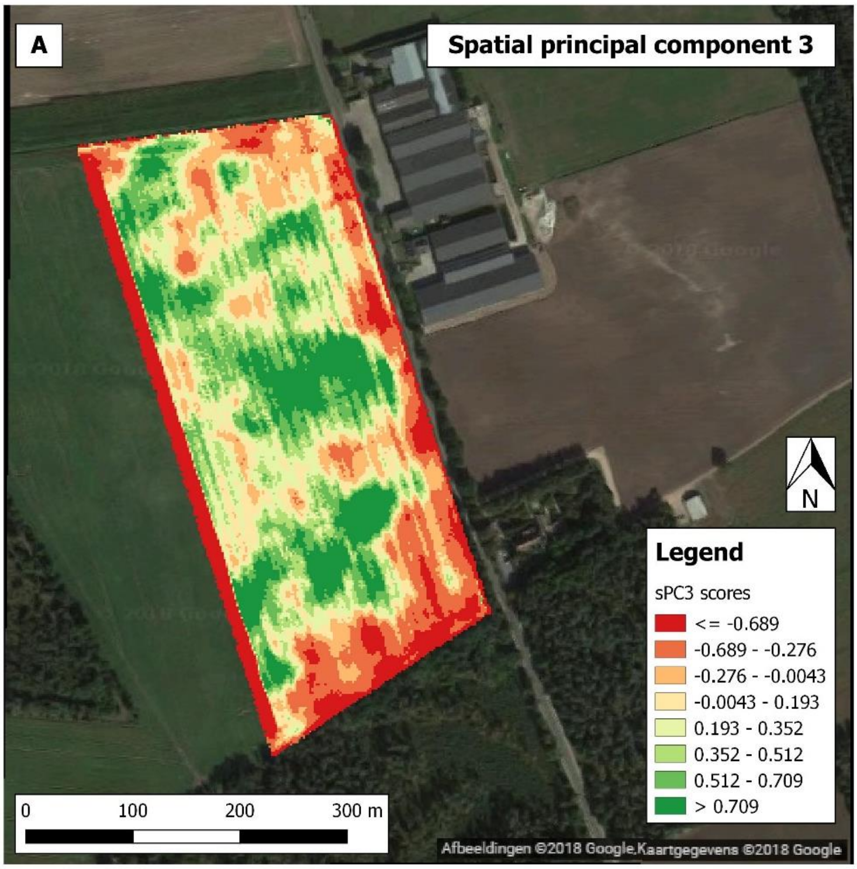

G
Total soil organic matter
N
Legend
Total SOM (ton/ha)
<= 184
184 - 192
192 - 199
199 - 207
207 - 215
215 - 224
224 - 236
> 236
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

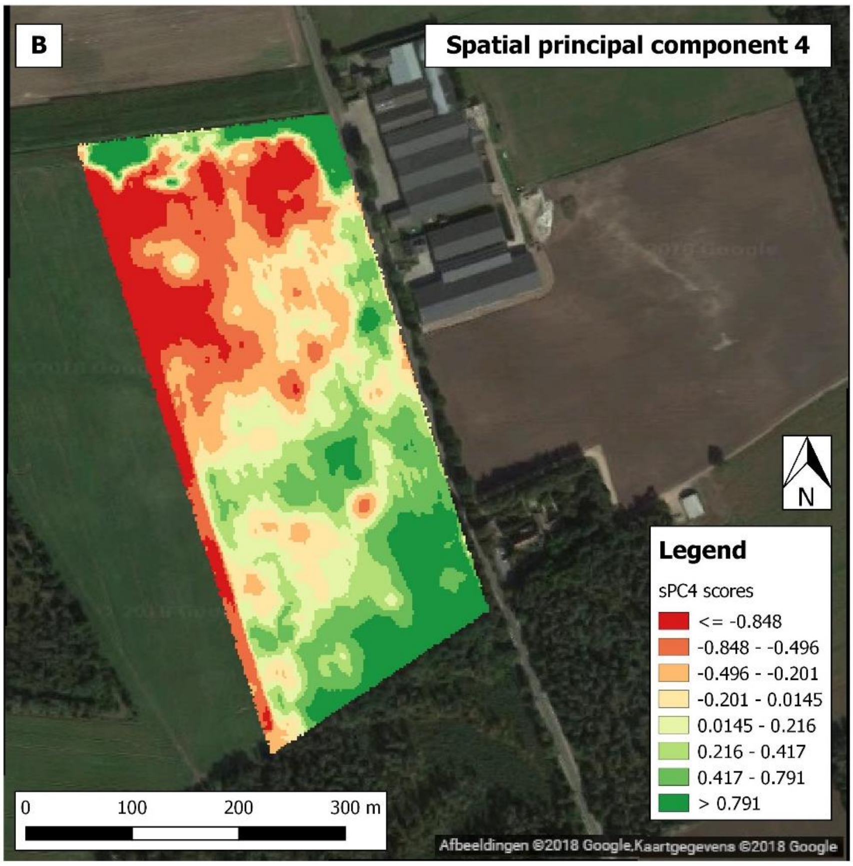

H
Elevation
N
Legend
Elevation (m + NAP)
<= 30.79
30.79 - 31.09
31.09 - 31.35
31.35 - 31.61
31.61 - 31.83
31.83 - 32.07
32.07 - 32.33
> 32.33
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google


I
Potato crop NDVI
N
Legend
Potato crop NDVI
<= 0.786
0.786 - 0.82
0.82 - 0.84
0.84 - 0.852
0.852 - 0.862
0.862 - 0.871
0.871 - 0.88
> 0.88
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

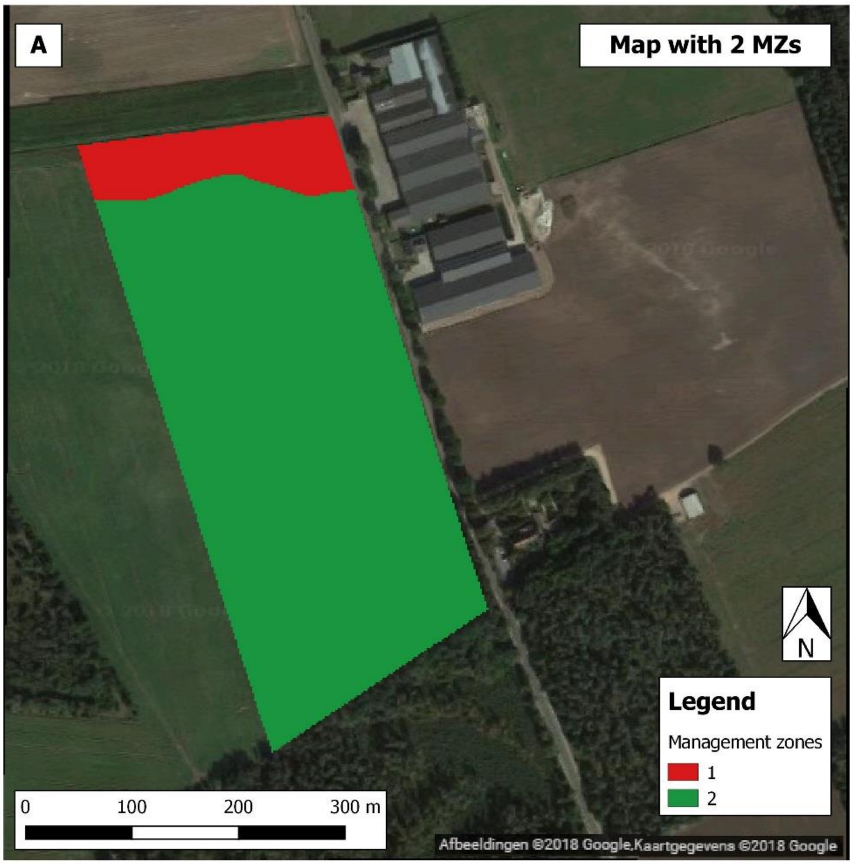

J
Potato crop WDVI
N
Legend
Potato crop WDVI
<= 0.185
0.185 - 0.223
0.223 - 0.247
0.247 - 0.265
0.265 - 0.281
0.281 - 0.299
0.299 - 0.322
> 0.322
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google

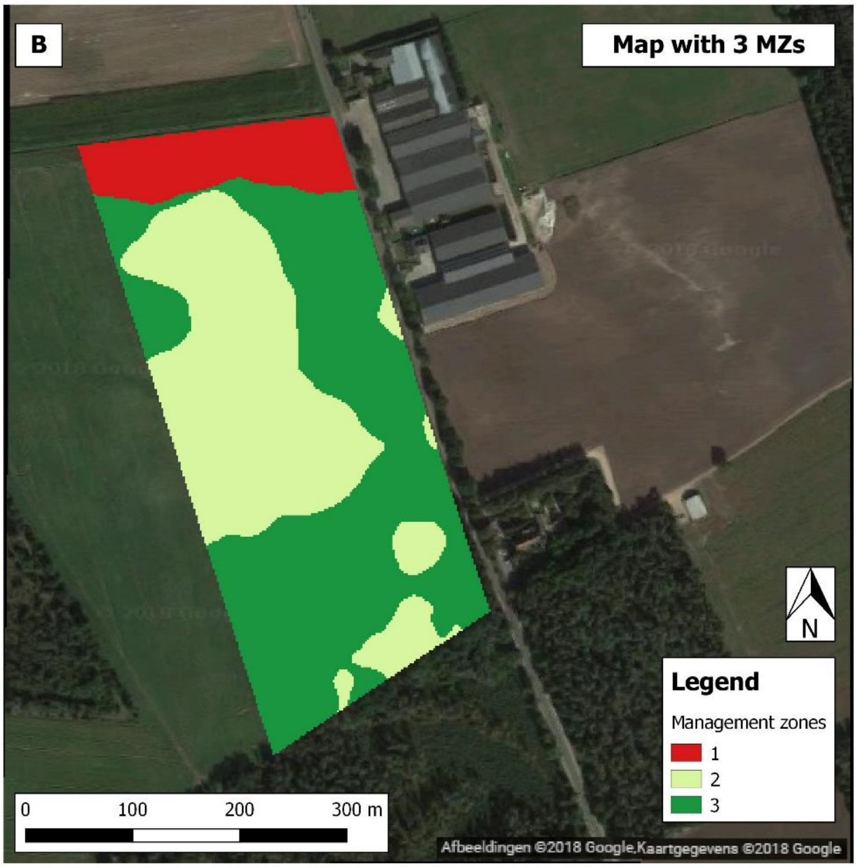
Appendix 11. Output maps sPC3 and sPC4





Appendix 12. Output maps with 2 MZs and 3 MZs


A
Map with 2 MZs
N
Legend
Management zones
1
2
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google


B
Map with 3 MZs
N
Legend
Management zones
1
2
3
0
100
200
300 m
Afbeeldingen ©2018 Google,Kaartgegevens ©2018 Google