

Spatial Data Mining in Precision Agriculture

Dissertation

zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

vorgelegt von Dipl.-Inform. Georg Ruß
geboren am 26. März 1981 in Wernigerode

Gutachter:

Prof. Dr. Rudolf Kruse
Prof. Dr. Alexander Brenning
Prof. Dr. Peter Wagner

Ort und Datum des Promotionskolloquiums:
Magdeburg, 23. Februar 2012

Ruß, Georg:

Spatial Data Mining in Precision Agriculture

Dissertation, Otto-von-Guericke-Universität Magdeburg, 2012.

Zusammenfassung

Informationstechnologie und moderne Datenverarbeitungsmethoden sind heutzutage in zunehmendem Maße die Ansatzpunkte technischer Innovationen. Davon bleibt auch die moderne Landwirtschaft nicht unberührt. Durch die Einführung neuer Geräte, GPS-gestützte Datenerfassung und räumlich hochaufgelöste Datensammlungen (engl. *spatial data*) bieten sich auf der einen Seite sehr viele Möglichkeiten zur ökonomischen und ökologischen Verbesserung vorhandener Prozesse. Auf der anderen Seite stellen diese großen und wachsenden räumlichen Datensammlungen in der Präzisionslandwirtschaft (engl. *precision agriculture*) die Datenverarbeitung vor neue Herausforderungen. Im zurückliegenden Jahrzehnt hat sich für die Anforderung, neues Wissen und neue Informationen aus vorhandenen Daten zu extrahieren, der Begriff *Data Mining* herausgebildet. Diese Arbeit beschäftigt sich auf der Grundlage von Datensammlungen der Präzisionslandwirtschaft mit zwei grundlegenden Aufgaben aus der Landwirtschaft, die mit Hilfe moderner Data-Mining-Methoden und -Algorithmen beantwortet werden können.

Die Grundlage dieser Arbeit sind Datensammlungen, die in moderner teilflächenspezifischer Bewirtschaftung anfallen. Unter anderem sind dafür Bodenleitfähigkeitsmessungen, Stickstoffdüngergaben, Bodenproben, Vegetationsindikatoren und Ertragsmessungen vorhanden. Diese Variablen und die einzelnen Ausprägungen sind georeferenziert, d.h. für jedes zugrundeliegende Feld ist mit einer gewissen räumlichen Auflösung bekannt, an welcher Stelle welche Variable welche Ausprägung besitzt. Diesen Datensammlungen werden ein zugehöriges Höhenmodell und daraus ableitbare Geländeattribut wie beispielsweise Hangneigung, Feuchtigkeitsindex und Krümmungen hinzugefügt.

Die erste der beiden Teilaufgaben befaßt sich mit der Ertragsvorhersage und setzt auf einer existierenden Arbeit auf diesem Gebiet auf. Die Ertragsvorhersage wird dabei als ein multivariates Regressionsproblem auf räumlichen Daten aufgefaßt. Die Beachtung der räumlichen Zusammenhänge erfordert die Veränderung einer herkömmlichen Kreuzvalidierung hin zu einer räumlichen Kreuzvalidierung. Ausgehend von dieser Änderung wird die Frage beantwortet, welche Regressionsmodelle sich am besten für eine Ertragsvorhersage eignen. Desweiteren kann die weitergehende Frage beantwortet werden, welche Regressionsvariablen für die Ertragsvorhersage interessant sind. Dies wird durch einen sogenannten Ansatz der räumlichen Variablenbedeutung (engl. *spatial variable importance*) evaluiert.

Die zweite Teilaufgabe befaßt sich mit dem Bestimmen von Management-Zonen (engl. *management zone delineation*). Hier ist ausgehend von einer Literaturrecherche im Bereich *precision agriculture* das Fehlen von speziell für diese Aufgabe zugeschnittenen Algorithmen festzustellen. Auch die Informatik bietet für die vorliegenden Daten keinen sofort passenden Algorithmus. Daher wird ein eigener Algorithmus (HACC-SPATIAL) entwickelt, der die aus der Literatur abgeleiteten Anforderungen erfüllt und auf hierarchischem agglomerativem Clustering mit einer räumlichen Einschränkung basiert. Insbesondere ist hier der gewünschte räumliche Zusammenhang der entstehenden Zonen einstellbar. Desweiteren bietet hierarchisches Clustering die Möglichkeit, das Ergebnis einfach zu explorieren und neues Wissen zu finden, was letzten Endes das Ziel von *Data Mining* ist.

Abstract

Technological advances are nowadays often based on improvements in information and data processing capabilities. Even modern agriculture is to a large extent based on adequate data processing, since the usage of novel information devices, GPS-based georeferenced data collection and high-resolution spatial data sets have become standard modes of operation, turning the once uniform site management into site-specific management as one of the most important sub-fields in *precision agriculture*. On the one hand, the resulting data sets clearly provide the foundations for economic and ecologic improvements. On the other hand, these data sets pose novel challenges for *spatial data mining*. Two specific tasks are explored in this study: *spatial variable importance* and *management zone delineation*.

The foundations of this thesis are data originating in site-specific management operations. They typically include electrical conductivity readings, fertilizer applications, soil sampling results, vegetation indicators and yield measurements. These variables are georeferenced, i.e. for a particular point of the site under study the variables and their values are known at a certain spatial resolution. These spatial data sets are furthermore augmented with digital elevation models from which terrain attributes such as slope, wetness index and curvatures are derived.

The first of the tasks is concerned with yield prediction and based on an existing dissertation in this area. Yield prediction is handled as a multivariate regression task using spatial data sets. However, taking the spatial relationships of the data sets into account requires some changes in the standard cross-validation to make it aware of spatial relationships in the data sets. Based on this addition, the question can be answered which of a variety of regression models are best suited for yield prediction. Eventually the regression models help to estimate which of the variables are important for yield prediction using permutation-based variable importance measures.

The second task is concerned with management zone delineation. Based on a literature review of existing approaches, a lack of exploratory algorithms for this task is concluded, in both the precision agriculture and the computer science domains. Hence, a novel algorithm (HACC-SPATIAL) is developed, fulfilling the requirements posed in the literature. It is based on hierarchical agglomerative clustering incorporating a spatial constraint. The spatial contiguity of the management zones is the key parameter in this approach. Furthermore, hierarchical clustering offers a simple and appealing way to explore the data sets under study, which is one of the main goals of *data mining*.

Danksagung

Mein Dank gilt als Erstes den drei Gutachtern meiner Dissertation. Prof. Kruse danke ich besonders für die Freiheit, letzten Endes genau das erforschen zu können, was mich interessiert, und dabei aus eigenem Antrieb zum Ziel zu kommen, die unzähligen Dienst- und Konferenzreisen eingerechnet. Prof. Brenning gilt mein Dank für die ursprüngliche fachliche Konkretisierung des Themas der Arbeit beginnend mit der GfKl-IFCS 2009 in Dresden, die hervorragende fachliche Zusammenarbeit und die beinahe unendlich vielen nützlichen kleinen und großen Hinweise für die Dissertation. Prof. Wagner danke ich für die sehr hilfreichen Treffen an seinem Lehrstuhl in Halle, die mir die landwirtschaftliche und praxisrelevante Seite meiner Arbeit klar aufgezeigt und viele wertvolle Hinweise geliefert haben. Die ursprüngliche Anregung zu dieser Arbeit stammt von Martin Schneider, meinem unmittelbaren Precision-Agriculture-Ansprechpartner und Ex-Doktoranden von Prof. Wagner, wobei auf der Fähre nach Kristiansand (N) 2005 nicht absehbar war, daß daraus eine Dissertation werden könnte. Die weitere unmittelbare Anregung wurde durch die Dissertation von Georg Weigert geliefert.

Während der Beschäftigung mit dem Thema entstanden eine Reihe von Konferenzpublikationen, deren anonymen Gutachtern ich danken möchte für die wertvollen Hinweise und Querbezüge zum Thema. Spezieller Dank gebührt Uwe Schulze für die Hinweise zu geographischen Informationssystemen und digitalen Höhenmodellen, sowie Korinna Bade für die Anregungen zum Clustering-Kapitel. Wertvolle Hinweise wurden auch in meinen eingeladenen Vorträgen am Australian Taxation Office (Australien), an der University of Waterloo (Kanada) und der NTNU Trondheim (Norwegen) zur Sprache gebracht und konnten berücksichtigt werden.

Die in dieser Arbeit verwendeten Datensätze entstammen Feldversuchen, die am Lehrstuhl für Landwirtschaftliche Betriebslehre der Martin-Luther-Universität Halle-Wittenberg bei Prof. Wagner vorgenommen wurden. Diese Datensätze wurden mit einem digitalen Höhenmodell des Landesamtes für Vermessung und Geoinformationsdienste Sachsen-Anhalt in Magdeburg angereichert. Ich danke beiden Institutionen herzlich für die Bereitstellung der Daten.

Technische Danksagungen gebühren all den Entwicklern von `vi`, `xfig` und `gentoo`, den Teilnehmern und Lesern der R-sig-geo-Mailingliste sowie den Entwicklern von R und SAGA.

Meinen beiden Töchtern Emma Charlotte und Greta Mathilde danke ich für die nicht durchwachten Nächte; ich habe mich daran gewöhnt, daß ihr beiden nachts schon immer ausgiebig schläft anstatt wach zu sein. Das Beste zum Schluß: Mimi! Ich danke Dir für die letzten Jahre, für die beiden Mädels und für das Ertragen eines nicht immer gut gelaunten und launischen Ehemanns. Ich gelobe Besserung.

Contents

1	Introduction	1
1.1	Thesis Contributions	1
1.2	Thesis Structure	2
2	Background	5
2.1	Spatial Data Mining	5
2.2	From Agriculture to Precision Agriculture	6
2.3	Winter Wheat	9
2.4	Precision Agriculture Data Sources	10
2.5	Data Available for this Study	17
2.6	Summary: Data Sets	21
2.7	Spatial Statistics and Spatial Autocorrelation	23
2.8	Temporal Relationships in the Data	29
3	Assessing the Importance of Data Sources for Yield Prediction	31
3.1	Introduction	31
3.2	Regression and Cross-Validation Made Spatial	32
3.3	Regression Models	38
3.4	Comparison: Non-Spatial vs. Spatial Regression Setting	46
3.5	Spatial Variable Importance Assessment	47
3.6	Results for Spatial Variable Importance	50
3.7	Discussion	65
3.8	Summary and Conclusion	69
4	Exploratory Spatial Clustering for Management Zone Delineation	71
4.1	Introduction	71
4.2	MZD as an Exploratory Spatial Clustering Problem	72
4.3	Literature Review in Precision Agriculture	73
4.4	Requirements	91
4.5	Spatial Clustering Algorithms	93
4.6	Hierarchical Agglomerative Clustering with a Spatial Constraint	104
4.7	Evaluation of HACC-spatial on PA Data Sets	115
4.8	Heuristic Parameter Guidelines for HACC-spatial	136
4.9	Summary	136

5	Conclusions	139
5.1	Thesis Summary	139
5.2	SVI results	140
5.3	MZD results	141
	Bibliography	153
	Appendices	
A	Data Set Details	173
A.1	F440, Variable Plots and Descriptive Statistics	174
A.2	F550, Variable Plots and Descriptive Statistics	177
A.3	F610, Variable Plots and Descriptive Statistics	182
A.4	F611, Variable Plots and Descriptive Statistics	185
A.5	F631, Variable Plots and Descriptive Statistics	189
B	Spatial Variable Importance Plots	193
C	R package details	241

Chapter 1

Introduction

Precision Agriculture is a compelling and highly active field of research typically based on large data collections and data-based decision making for agricultural operations. Among this field, site-specific management for farming operations is likely to be the most important development recently. Only decades ago, the amounts of yield or fertilizer, e.g., could not be quantified for a small area, but were rather provided for a whole site. Nowadays, however, technological advances enable small-scale statements about a crop's status, and the farming equipment, e.g. with variable rate technology, is up to the task of delivering farming inputs in precise amounts at precise locations, compared with earlier technology. Those technological advances provide technically well-equipped farm operators with a competitive edge leading to unparalleled potential benefits, both economically and ecologically.

However, those benefits depend on large georeferenced data sets and rely on the ability to adequately process them. Since the data collections are growing rapidly, it is essential to put research efforts into methods which deal with those data sets. Hence, *data mining* as an established research area increasingly takes center stage in precision agriculture (PA) when it comes to discovering novel and potentially useful knowledge. Since the data sets are georeferenced, the focus is in particular on *spatial* data mining. Naturally, the data sets facilitate numerous precision agriculture tasks, hence only two of those are further elaborated upon in this study: *yield prediction / variable importance* and *management zone delineation*. While both tasks are long known in traditional agriculture, high-resolution and geo-referenced data collection and adequate data mining techniques with the help of geostatistics will provide the basis for advancing these tasks and enabling further improvements.

1.1 Thesis Contributions

With the advent of an ever growing number of potentially useful data sources for precision agriculture, the natural question to ask is which of those sources are actually important for a particular task. Since yield prediction is one of the key tasks in agricultural farming operations, the first part of this thesis deals with assessing a variable's importance for this yield prediction task. Chapter 3 first lays the groundwork for incorporating spatial data into a yield prediction task. Afterwards, a variable's importance is assessed using a

permutation-based spatial variable importance approach. As a desired side effect, different yield prediction models are evaluated in terms of their predictive performance.

Another traditional question arising in agriculture is the so-called management zone delineation: how should a field be split up into zones for a particular purpose? An exemplary purpose could be basic fertilization, which aims to make soil and fertilizer minerals available to the plants. Chapter 4 is the second central topic of this thesis, covering this particular aspect. The basic existing approaches towards this task are shortly laid out. Based on the shortcomings of existing work, a novel approach is developed based on a special variant of clustering which takes the spatial nature of the data explicitly into account.

1.2 Thesis Structure

Figure 1.1 illustrates the data mining cycle which this thesis follows. It has been adapted for precision agriculture using the four standard steps for data mining by [Fayyad et al., 1996a]. Since this study revolves around precision agriculture and site-specific management as well as specific data sets and the area of data mining, those and related fields of research are introduced to the degree necessary in Chapter 2, covering the data acquisition and the data preprocessing. Having provided the prerequisites, the core of this thesis are the data mining models for regression and clustering for two specific PA tasks in the following two main chapters. While yield prediction and the determination of important data variables are laid out in Chapter 3, the management zone delineation approach via spatial clustering is developed in Chapter 4. Both main chapters provide detailed results and discussions on the respective topics. In Figure 1.1, they cover the data mining and the data usage steps. They are followed by a short conclusion in Chapter 5 which summarizes the main research results and contributions of this study.

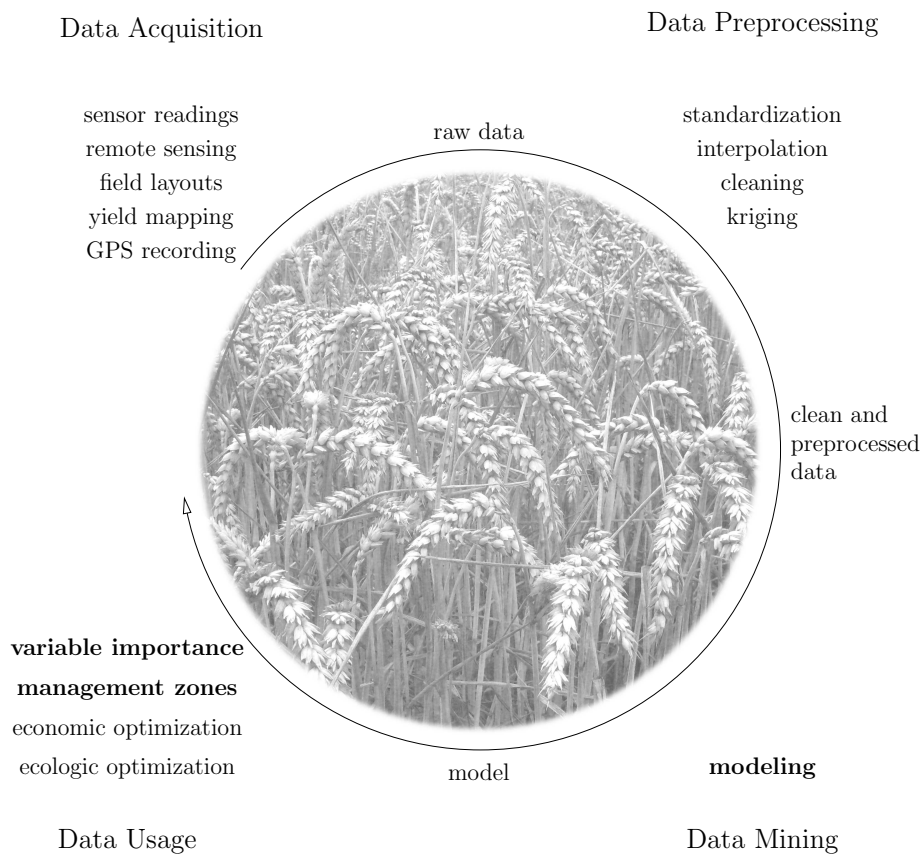


Figure 1.1: Data Mining cycle in the context of precision agriculture. This thesis is mainly concerned with the data mining modeling and data usage steps for the tasks of determining important data variables and management zone delineation.

Chapter 2

Background

This chapter provides the necessary details and insights into the data sets and methods underlying this study. In particular, the areas of data mining and precision agriculture are briefly outlined, with the main focus on site-specific management. The general data sources for precision agriculture are presented, followed by the specific data sets. The chapter closes with a brief section on spatial autocorrelation and temporal relationships in the data sets.

2.1 Spatial Data Mining

With ever growing amounts of stored data, Data Mining (DM) has become an ever more important focus of research since the end of the 1990s [Aggarwal and Yu, 1999]. It has traditionally been embedded into Knowledge Discovery in Databases (KDD), as described by [Fayyad et al., 1996a]. KDD is a larger process which also encompasses the steps before and after the actual data mining step. However, DM and KDD are nowadays often used synonymously, such that the distinction between these two areas has become rather fuzzy.

According to [Fayyad et al., 1996b], Data Mining is “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. DM can be applied to a variety of problems, including spatial and temporal data sets, in addition to classical transactional data bases. Four important problem areas are those of *association*, *clustering*, *classification* and *regression*. While association rule mining focuses on finding relationships between the different items in a transactional database, clustering focuses on finding a partition of data records into clusters such that the points within each cluster are close to one another. Classification and regression are related tasks. The first is a process in which a model learns to predict a class label from a set of training data which can then be used to predict discrete class labels on new samples. For regression, the model is trained to predict a continuous target. Regression tasks are hence often treated as classification tasks with quantitative class labels, e.g. in classification and regression trees [Breiman et al., 1984].

While traditional data mining approaches are typically focused on transactional and relational data bases, georeferenced data collections require methods that deal with spatial relations in the data. Hence, the main requirement is that attributes of the neighbors of some object of interest may have an influence on the object and therefore should be con-

sidered. The explicit location and extension of spatial objects define implicit relations of spatial neighborhood (such as topological, distance and direction relations) which are used by spatial data mining algorithms [Ester et al., 1999]. Terms which have been used synonymously are *geographic knowledge discovery* [Guo and Mennis, 2009] or *geographic data mining and knowledge discovery* [Miller and Han, 2009]. The common tasks from standard data mining are adapted for spatial data sets, hence there is spatial classification, spatial regression, spatial association rule mining, spatial clustering which obey and exploit spatial relationships in the data sets. Necessarily, geostatistical methods are employed to account for the spatial nature of the data [de Smith et al., 2009]. In the style of [Fayyad et al., 1996b], spatial data mining is understood as the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in *spatial* data.

Since this study revolves around a few precision agriculture data sets, a considerable proportion in the following is devoted to explaining the provenance of those. DM is usually embedded into a large-scale operation, consisting of at least four steps: acquisition, preprocessing, mining and usage of data [Fayyad et al., 1996a], as laid out for precision agriculture in Figure 1.1 on Page 3.

2.2 From Agriculture to Precision Agriculture

While (spatial) data mining, as explained in the previous section, is an established research area by itself, its main success results from its application in a range of fields as diverse as financial services, insurances, medical research, telecommunications and other industries. Even agriculture is turning increasingly into a data-driven operation. Hence, this section serves as a brief introduction to the area of precision agriculture (PA).

The term *precision agriculture* is often used synonymously with *precision farming* (PF), since PF is the largest subset of PA, although PA would also encompass, e.g., precision livestock farming. Among PF, one of the most important research and application areas is *site-specific (crop) management* (SSM), which is also the key topic in this study. SSM is a term also increasingly used as a synonym for PA and is defined as “the management of agricultural crops at a spatial scale smaller than that of the whole field” [Plant, 2001]. Hence, SSM is about “managing within-field heterogeneity” [Schmidhalter et al., 2008]. A closely associated term is *variable rate technology*, which is a prerequisite that enables a farmer to actually apply, e.g., different amounts of fertilizer on different field parts. The key variables in SSM are the input (mainly fertilizer) and the output (yield). However, the main target is typically not to maximize yield, but to maximize the economic profits. The economic potential of precision agriculture in general has been described in [Begiebing et al., 2007].

A historical outline of PA is given below, followed by a contemporary definition and a brief history of SSM as well as the effects of using data mining techniques in PA. With the advent of cheaper and more powerful microprocessors and more sophisticated on-farm technology, PA is in the process of becoming a wide-spread reality while it was yet a vision around the middle of the 1980’s when the digital revolution started.

2.2.1 Past

In 1985, PA was sketched rather diffusely, while computers and microchips were already emerging. The following quotes are from the first editorial of *Computers and Electronics in Agriculture* [Lambert, 1985]:

Silicon chips have been combined into devices very marvelous to behold. They speak to us if we do not fasten the seat belt. They keep time, date and notes-for-the-day on our wrist. They calculate the optimum selection of farm machinery for a given enterprise configuration. They analyze projected cash flow for the next year. They control our microwave ovens. They simulate the response of a cotton crop to its specific environment. And they determine how much feed each cow in a milking herd deserves.

People were fascinated by the technology and saw imminent use in agriculture, such as for quick calculations of equipment usage and special sensors, which partly anticipated site-specific management. Some obstacles remained, though:

The largest obstacle to implementation of data acquisition and control systems in most areas of agriculture is the lack of appropriate sensors. Sensors are necessary to provide control systems with information on the controlled variables, e.g. to measure the temperature of a room for temperature control. But now, in the same way that the well-known electrical properties of silicon have paved the way for such dramatic advances in electronic devices, microminiature devices employing the excellent mechanical properties of silicon stand poised to revolutionize sensors and sensing technology. Microminiature devices of silicon provide a number of significant advantages over current sensor technology, including smaller size, lower cost, higher performance and longer life.

Hence, the technological advances were already anticipated then, with ever-smaller and more powerful “silicon devices”, which holds true until 25 years later. However, the potential was also clear for what has since then developed into “computational intelligence” or “data mining” (called “expert systems”) in 1985:

An emerging software development, enabled by silicon technology, is expert systems. An expert system is a computerized knowledge base equipped with rules to solve a problem that usually requires an expert. Expert systems now stand poised to deliver the knowledge and experience of specialists or experts through silicon to the laity. The potential exists to combine sensor inputs from the local environment, data from a remote data base, user responses to specific questions and expert knowledge of system behavior to derive conclusions and actions that heretofore would be attributed to human intelligence.

The aim of SSM is therefore to combine the available data, such as local sensor inputs, remote sensing data and expert knowledge to aid the user in decisions and conclusions that would not be possible without this technology. Judging from nowadays’ data mining perspective, PA and SSM were destined to become data-driven applications.

With the abovementioned developments in technology, the site-specific management strategy became feasible from the beginning of the 1990s. However, to be broadly adopted, economical incentives are the main driving factor ([Khanna et al., 1999]). With, e.g., higher fertilizer prices, the adoption of SSM strategies quickly returns higher profits than traditional (uniform) field management ([Hüter et al., 2005]). Provided that significant in-field variability exists, one of the key tasks in SSM is to identify and measure the causes of this variability, along with adequately accounting for the heterogeneity.

2.2.2 Present and Future

Currently, PA is in the process of being adopted by farmers worldwide. Ground-based and remote sensors, as well as aerial or satellite images are increasingly used as a data basis for PA. The global positioning system (GPS) serves as the base for georeferencing the data which are being collected (cp. Section 2.4.1). It also serves as a standard guidance system for field equipment, such as automatic steering aids and parallel track guidance. A study conducted in conjunction with the German preagro project¹ provides results on market penetration, motivation and adoption patterns until 2006 in Germany [Reichardt and Jürgens, 2009]. According to this study, in 2001 the percentage of PA users among those interviewed was 6.65%, rising to 11.04% in 2006. More interestingly, while in 2001 46% of those interviewed were unaware of PA, this percentage dropped to 28% in 2006. Furthermore, while in 2001 the most important reason to introduce PA methods was to obtain a better knowledge of the field (49.5% of PA users), this reason came in third (28%) in 2006, outranked by the financial benefits (36%) and lowering the costs (40%). According to the study, in Germany the adoption of PA typically begins with introducing GPS technology: GPS-based area measurement is adopted first, followed by GPS-based soil sampling and GPS-based yield mapping. Site-specific basic fertilization, N-fertilization and using the N-sensor showed quite constant adoption rates. The main obstacles for adopting PA were problems with interoperability (28% in 2006) and the time it takes to get used to the technology (26%), based on PA users interviewed.

With the growing adoption of PA methods, the clear need for what was called “expert systems” in 1985 is even more imminent today. Cheaper technology simplifies the task to collect more data – with the drawback that the data are not always used to their full extent. There are certainly interesting pieces of information hidden in these data – finding these is a task for spatial data mining.

In the future, alongside many other fields, PA is bound to turn into a more and more data-driven field. Hence, the need for applying spatial data mining to the resulting data sets which is imminent today will be even more pressing tomorrow and should be addressed using suitable techniques and methods from computer science and geostatistics.

2.2.3 Practical Agricultural Effects of Data Mining

According to the *Food and Agriculture Organization of the United Nations* (FAO), in 2009 wheat ranked second worldwide in terms of total production at 683 million tonnes, followed by paddy rice at 680 million tonnes and outranked by maize (corn) at 1125 million

¹<http://www.preagro.de>

tonnes [FAO Trade and Market Division, 2010]. In the European Union, wheat is the most important crop being produced, at 150 million tonnes, followed by maize at 63 million tonnes. The two tasks covered by this thesis may influence the economic planning around wheat yield or may affect yield directly. Both tasks are equally important for crops other than wheat, and since the approaches in this thesis are not specifically tailored to winter wheat, the results may be carried over to other crops in an appropriate way. Nevertheless, the adoption of PA is typically motivated by economic advantages rather than simply increasing yield (cp. Section 2.2.2). Despite this, spatial data mining aims to extract knowledge about the data sets which can equally be used for purposes such as maximizing profits or increasing yield. Therefore, the effective application of data mining techniques on PA data serves a straightforward practical purpose.

2.3 Winter Wheat

This study is concerned with data sets resulting from five winter wheat sites in East Germany. This crop is planted before winter, enters a dormant phase during winter and breaks dormancy in spring. The phenological development stages are defined by the BBCH² scale, as described in [Meier, 2001]. For a wide range of crop species, BBCH scales have been developed. They are used in a variety of disciplines, such as crop physiology, plant pathology and plant breeding, as well as in agriculture and agriculture-related industries. The BBCH scale employs a decimal code system divided into principal and secondary growth stages and is based on the cereal code system by [Zadoks et al., 1974]. Figure 2.1 shows the BBCH stages for winter wheat. Tillering occurs from stage 10 to 30, while stem extension follows until stage 49. Afterwards, heading occurs until stage 69, ended by ripening.

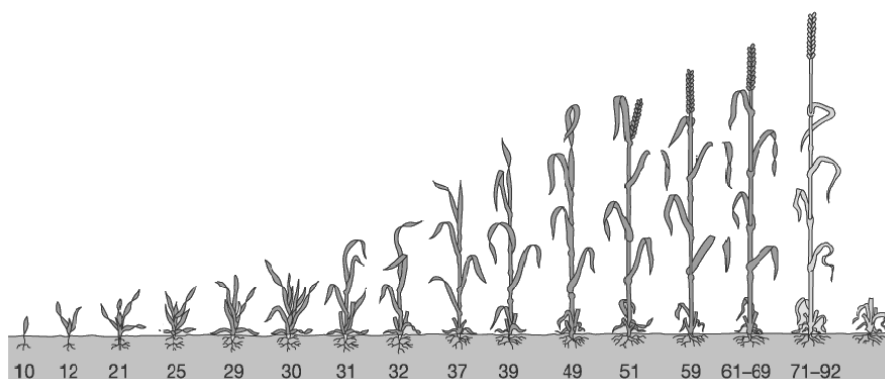


Figure 2.1: Growing stages of winter wheat (*Triticum aestivum* L.), reproduced from [Meier, 2001].

²Biologische Bundesanstalt, Bundessortenamt und Chemische Industrie

2.4 Precision Agriculture Data Sources

Before showing the specifics of the available data sets, the principal classes and categorizations of data from the precision agriculture domain are presented. Afterwards, the particular data variables available in this study are provided.

2.4.1 Global Positioning System

For site-specific management it is essential to have geo-referenced data records, i.e. each point has a precisely defined location in space. There are a number of sources for this type of locational data, of which the most commonplace one recently is the global positioning system (GPS). It consists of geostationary satellites whose locations in orbit are known. A GPS receiver can compute its geographical location based on triangulation using the satellite signals it is receiving. In practice, this allows for an accuracy of around 10-15 metres for standard consumer GPS devices. GPS was originally a military system and the publicly available signal was deliberately distorted until May 1st, 2000, when the so-called *selective availability* was decommissioned [Grewal et al., 2001]. GPS augmentation systems such as EGNOS (Europe) and WAAS (North America) are specified to provide an accuracy in the 5 m range, although in practice they often deliver higher accuracies [Alcantarilla et al., 2005]. For accuracies in the cm range, local correction signals as in differential GPS may be applied and can eventually lead to an accuracy of around 2 centimetres (cp. [Sabatini and Palmerini, 2008]). Further measurement sources, especially for digital elevation models, are laser scanners such as LiDAR (Light Detection and Ranging) or radar [Nelson et al., 2009]. The basic information for each data record being collected with a positioning system is therefore some type of point ID, as well as x-value (easting), y-value (northing) and z-value (elevation). In the data sets presented later, these data are not explicitly mentioned since it is assumed they are available.

2.4.2 Remote Sensing

Aerial and satellite remote sensing is being increasingly used to monitor a wide range of variables that affect crops, such as soil moisture, surface temperature, weed or pest infestations and photosynthetic activity. Thus, it provides support in more efficient site management. According to [Colwell, 1997], “remote sensing is the art, science & technology of obtaining reliable information about physical objects and the environment, through the process of recording, measuring and interpreting imagery and digital representations of energy patterns derived from non-contact sensor systems.” Remote sensing data are typically non-invasive, high-resolution and in some situations potentially less costly than in-situ sampling and laboratory analyses. It is therefore being investigated if they may partly replace soil sampling data (cp. [Sommer and Wehrhan, 2005]). For further introduction to the area of remote sensing, see, e.g. [Jensen, 2006].

For satellite and airborne sensors, the area under study is subject to being surveyed a few times during the vegetation period, while the spatial resolution mainly depends on the altitude of the plane or the satellite and its equipment. Once a digital raw image of the area under study has been acquired, the actual imagery is typically geometrically and

atmospherically corrected and preprocessed. Different spectral bands and sensing characteristics may be used, e.g. hyperspectral information [Cetin et al., 2005] or thermal imaging [Soliman et al., 2011]. Vegetation indicators are often calculated from the visible part of the spectrum and also available in the data sets for this study. The indicators differ mostly in the spectral bands used. Some examples of these indices are the NDVI (normalized differential vegetation index), SAVI (soil-adjusted vegetation index), VFI (vegetation fraction index), VARI (visible atmospherically resistant index), LAI (leaf area index) and REIP (red edge inflection point) of which a short comparison is, e.g., provided in [Schmidhalter et al., 2004]. Further details can also be found in geographical information system software such as GRASS [GRASS Development Team, 2010]. In the data sets in this study the REIP vegetation indicator is available and therefore further introduced below.

In this study, the *red edge inflection point* is available as a measurement in the data sets. The following provides an overview on the optical “vegetation status” sensing rationale which is behind the REIP value. In the case of nitrogen fertilization, it is certainly interesting to measure the plants’ nitrogen uptake. On the one hand, a plant’s photosynthesis is highly dependent on chlorophyll, which chemically requires nitrogen as one of the core atoms around a magnesium ion. Hence, the nitrogen uptake can be measured indirectly by the amount of chlorophyll in the plant’s leaves [see, e.g. Middleton et al., 2002]. On the other hand, nitrogen stimulates plant growth. The first effect (higher chlorophyll concentration) causes more light of the visible waveband, the so-called photosynthetically active radiation, to be absorbed by the leaves – reflectance of light from this spectrum decreases. The second effect causes the leaf area index (LAI)³ to increase. As more soil is covered by plant mass, a larger fraction of solar radiation should be reflected by the plants instead of the soil. This is, however, important only for the near-infrared region of the spectrum. Generally, the spectrum of the reflected light changes drastically with nitrogen uptake. This change can be measured by the red edge inflection point, which changes towards a higher value with higher nitrogen uptake. Figure 2.2 provides a graphical representation of this fact.

The exact calculation of the REIP value is the second derivative of the reflectance R with respect to wavelength:

$$\frac{d^2 R}{d\lambda^2} = 0 \quad (2.1)$$

An approximate formula proposed by [Guyot et al., 1988] which is used in practice is

$$REIP = 700 + 40 \frac{\frac{R_{670} + R_{730}}{2} - R_{700}}{R_{740} - R_{700}} \quad (2.2)$$

where the R_i values represent the reflectance at the respective wavelength i in nm. A discussion of further reflectance indices and measuring issues can be found in [Heege et al., 2008]. The REIP values are measured before the second and third nitrogen application, at the BBCH growing stages 32 and 49, respectively. For further information on particular types of sensors and a more detailed introduction, see [Weigert, 2006] or [Liu et al., 2004]. Plants that have less chlorophyll tend to have a lower REIP value as the red edge moves toward the blue part of the spectrum. On the other hand, plants with more chlorophyll tend to show higher REIP values as the red edge moves toward the higher wavelengths.

³calculated as $\frac{\text{leaf area}}{\text{ground area}}$

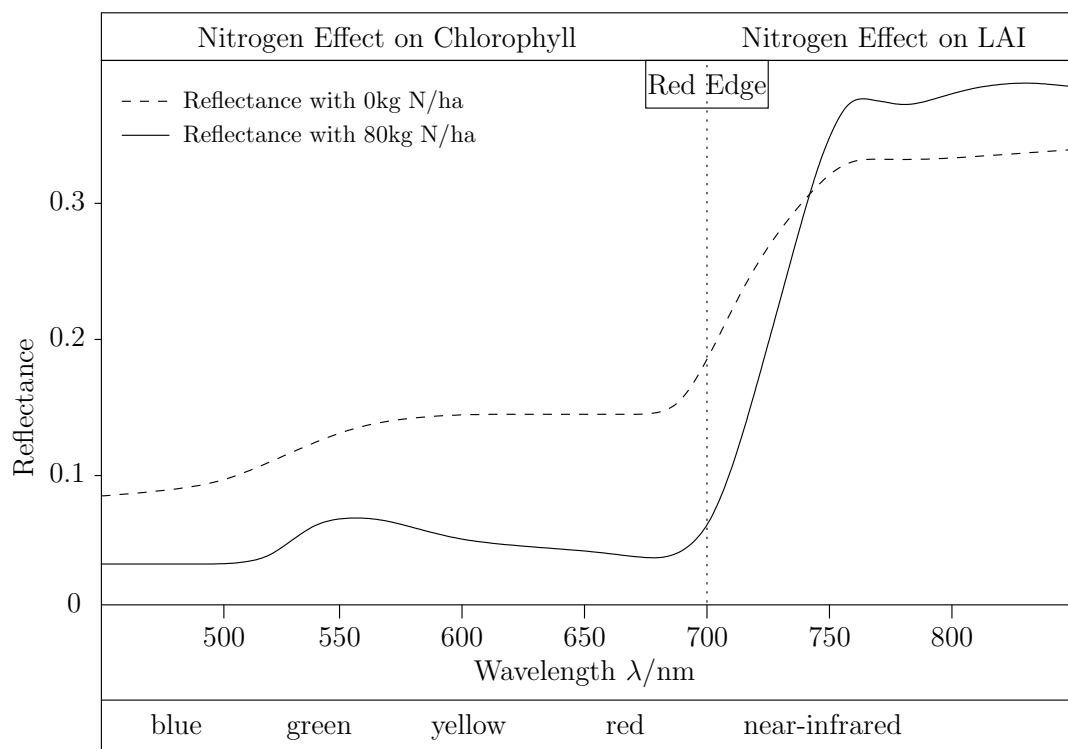


Figure 2.2: Red Edge Inflection Point Shifting, figure reproduced from [Heege et al., 2008]. The red edge inflection point shifts roughly from 710 nm to 730 nm in this figure, due to the higher amount of N fertilizer.

2.4.3 Soil Sampling

Soil sampling data, on the contrary to remote sensing data, are obtained in an invasive and often destructive way. They are collected *in situ* by taking soil probes at preset locations or in fixed intervals. These probes are laboratory-processed to reveal mineral content, organic matter content, humidity and pH value. This process is labor-intensive and time-consuming, therefore often expensive. The cost of soil sampling is mostly determined by the required spatial resolution.

Soil pH value affects the availability of nutrients, the activity of microbes and can also cause toxicity problems. Generally, soils tend to acidify by use of acidic fertilizers, leaching nitrate and basic elements as well as organic material decomposition. Furthermore, the optimal pH value varies between different crops. If the soil turns too acidic, toxic concentrations of aluminium and manganese may occur; in addition, soil microorganisms are usually affected. The availability of calcium and the cation exchange capacity may be affected. On the other hand, a high pH is likely to reduce the availability of phosphorus and other micronutrients. Hence, the pH value is likely to be an important factor in precision farming operations. Soil pH is usually set to the required levels using liming. During this process, an alkali solution is applied to the field, raising the pH value to the desired level and making certain nutrients available to the plants. For further details and a more detailed in-

production on how to measure pH values and pH effects, see, e.g. [Adamchuk and Mulliken, 2005].

According to [Wood et al., 1994], phosphorus (P) has the primary function of energy storage and transfer through the plant. Adenosine diphosphate (ADP) and adenosine triphosphate (ATP) are high-energy phosphate nucleotides that control most processes in plants including photosynthesis, respiration, protein and nucleic acid synthesis, and nutrient transport through the plant's cells. Hence, determining the appropriate amount of available phosphorus is important for modern crop production systems. Along similar lines, potassium (K) and magnesium (Mg) perform vital roles in plants, including photosynthesis, enzyme activation, stomatal control and transport of plant sugars.

2.4.4 Non-Invasive Geophysical Methods

One of the most important non-invasive geophysical methods to characterize a site's heterogeneity is to measure its apparent electrical conductivity [Corwin and Plant, 2005]. One of the key nutrients to a crop are dissolved inorganic solutes in the soil. Soil salinity refers to the presence of those solutes in the aqueous phase, including, but not limited to Na^+ , K^+ , Mg^{2+} , Ca^{2+} , Cl^- , HCO_3^- , NO_3^- , SO_4^{2-} and CO_3^{2-} . The salinity is quantified in terms of the total concentration of the solutes, measured by the electrical conductivity (EC) of the solution in dS m^{-1} . In practice, the soil's electrical conductivity is determined for an aqueous extract of a soil sample. However, the soil-sampling process is time-consuming and labor-intensive, hence expensive, when done to obtain small-scale data. Therefore, EC measurement has shifted towards determining the apparent electrical conductivity EC_a . It measures conductance through the solid soil particles and via exchangeable cations at the liquid-solid interface of clay minerals, in addition to the soil solution. It has become one of the most reliable and most frequently used, non-invasive methods to discover and map a field's heterogeneity. EC_a often exhibits empirical site-specific relationships with yield and may thus be a proxy variable to identify variables that influence yield [Kitchen et al., 2005; Ezrin et al., 2009; Corwin et al., 2003]. Basically, there are two indirect methods available for the determination of EC_a : electrical resistivity (ER) and electromagnetic induction (EM), both summarized in the following. For a more detailed discussion see, e.g., [Corwin and Lesch, 2003]. Other non-invasive geophysical methods include seismic imaging and ground-penetrating radar [Daniels, 2000] as well as mechanical, acoustic, pneumatic and electrochemical measurements (for an overview cp. [Adamchuk et al., 2004]).

ER: Electrical resistivity methods apply an electrical current into the soil via current electrodes at the surface. The difference in current flow potential is measured at potential electrodes placed near the current flow. The penetration depth of the electrical current and the volume of soil that is measured are determined by the interelectrode spacing. The larger the spacing, the deeper and the larger the volume of measurement.

EM: While ER is an invasive technique to measure EC_a , EM measures EC_a remotely. An electromagnetic transmitter coil located at one end of the appliance induces circular eddy-current loops in the soil. The loops' magnitude is directly proportional to the EC_a of the nearby soil. A fraction of the induced EM field is collected by the receiver coil, and the sum of these signals is amplified and gives an output voltage. Further

properties of the secondary magnetic field are measured, such as amplitude and phase — those differ from the primary field properties and thus allow to derive certain soil properties.

2.4.5 Yield and Input Mapping

Yield mapping is currently becoming a standard operation, since a large part of harvesting equipment is already GPS-equipped and can therefore record yields in conjunction with their geographic locations. Due to the width of the harvesting equipment and the operations taking place in the combine harvesters, there is an upper bound for the spatial resolution at around 15–25 metres [Lark et al., 1997] along the track. Nevertheless, yield mapping is rather low-cost since it requires little additional effort.

Input mapping, similar to yield mapping, is becoming a standard operation. Inputs such as fertilizer or pesticides are metered by the equipment and the allocated amount is stored along with its geolocation. While inputs have usually been uniformly distributed in the past, nowadays' technology allows for *variable rate application*, where the input is allocated based on small-scale recommendations resulting from on-line or off-line data analysis.

The most common type of input is nitrogen fertilizer. Nitrogen (N) is a constituent of amino acids, cell walls, chlorophyll, nucleic acids and proteins, among others. Therefore, nitrogen is essential for crop growth. The plants absorb nitrogen from the soil solution as a mineral, i.e., as nitrate or ammonium ions. There are three sources of mineral nitrogen in the soil: it can originate from the mineralization of organic matter, result from wet or dry atmospheric deposition and from organic or inorganic fertilizers [Ma et al., 2009]. Nitrogen losses during cropping and harvesting can usually not be avoided, hence the N mineralization is often insufficient for the crop's needs. Furthermore, the uptake of atmospheric N₂ can have an effect as well and must be taken into account when applying nitrogen fertilizer [Ledgard and Giller, 1995]. Typically, the nitrogen shortage in the soil is overcome by applying nitrogen fertilizers.

2.4.6 Topography and Digital Elevation Models

Topographic information for a site under study is often available in the form of topographic maps or digital elevation models. From these basic positional data (the (x, y, z) -triples), further terrain attributes can be derived. In precision agriculture, properties like slope and aspect are expected to have an influence on plant growth since they determine water supply and the amount of sunlight, among others. For an overview about additional attributes, see [Olaya, 2009].

Especially if precision agriculture turns from the treatment of one field towards multiple fields and larger sites, aspects like water flow [Gruber and Peckham, 2009] and a basic landform classification can come into play [MacMillan and Shary, 2009]. Those landforms can be derived from the positional data, for an overview see, e.g. [Evans et al., 2009]. It would, however, be interesting to learn which of the variables derived from basic positional information may be empirically related to yield management [Reuter and Kersebaum, 2009]. Relationships between moisture, curvature and yield have been found as far back as 1981 [Sinai et al., 1981], while in recent years the increased availability of fine-scale digital

elevation models and geographical information systems provided easy access to terrain attributes. Nevertheless, relationships between terrain attributes and yield are likely to vary between sites and years, e.g. due to weather conditions.

In summary, topographic information is nowadays rather easy to acquire using GPS technology with correction signals or other remote sensing techniques like LiDAR. Those data are non-invasive and have a high accuracy after preprocessing and corrective steps. The geospatial information contained in those data is the key to the kinds of spatial data mining which this thesis aims at. For more detailed information on topographic information and its use in areas like precision agriculture, the reader is referred to, e.g. [Hengl and Reuter, 2009]. For the data sets in this study, digital elevation models (DEMs) have been obtained.

2.4.7 Data Characteristics

There are at least two main categories for different types of data collected in precision agriculture, based on the spatial resolution and the acquisition cost.

Resolution: Depending on the method of acquiring the data and usually also the type of sensor, field data can be logged at different spatial sampling densities or resolutions. Commercial satellite imagery, for example, can range up to resolutions of $0.4 \text{ m} \times 0.4 \text{ m}$ per pixel, depending on the sensor. Aerial imagery from manned or unmanned aerial vehicles (drones) can achieve even higher spatial resolutions, depending on the height at which the equipment is flown and on the digital sensors. From a data mining point of view, it is usually desirable to acquire high-resolution imagery since lower resolutions can be generated from this material by image processing techniques.

Soil sampling and yield mapping typically show less spatial detail than the previously mentioned remote sensing techniques. For soil sampling, this is due to the extremely labor- and time-intensive work (at least for high spatial detail), while for yield mapping the scale is determined by the harvesting equipment.

The decision for a specific spatial resolution is furthermore often based on the economic feasibility and return-on-investment questions since higher resolution imagery is typically more expensive.

Remote sensors are non-invasive, while, e.g., data from soil sampling are rather invasive. Ground-based remote sensing such as electromagnetical conductivity and ground-based imagery are less invasive than soil sampling.

Acquisition cost: As far as data mining is concerned, more and higher-resolution data sets are clearly desirable. However, when it comes to deciding which data and which equipment to purchase and use, in practice the return on investment of using additional data has to be considered. With ever more affordable up-front hardware costs and service providers providing the precision farming equipment, site-specific management increasingly provides economic benefits, e.g. in N fertilization for wheat [Wagner and Schneider, 2007; Biermacher et al., 2009; Bongiovanni et al., 2007], which is the operation that the data sets in this study result from. Purchasing satellite or aerial imagery services might be appropriate, while extensive soil sampling and buying (rather than renting) expensive equipment might not be worthwhile, such

that an economic tradeoff is typically made. However, this study mainly focuses on research data sets and on the data mining process, which is why the economic benefits or drawbacks are not considered here. For the two research topics *yield prediction* and *management zone delineation* covered by this study, some information on the economic issues can be found in the respective chapters.

2.4.8 Data Preprocessing

Due to small imprecisions, different data densities and different data sources, the resulting variables are not necessarily geographically co-referenced with each other. For data analysis purposes, it is usually desirable to have the data available on a fixed grid. There are a few approaches for interpolation to a fixed grid such as nearest-neighbor interpolation, inverse distance interpolation and ordinary kriging.

Hand Contouring is a manual approach. For each variable a contour map is produced, such as those in Figures 2.7a and 2.7b. A grid can be overlaid on the variable maps and the georeferenced grid points can be sampled. When no PA data are available, these maps are subjective and require expert knowledge, but can be a useful tool for a deeper understanding of the data.

Interpolation with Nearest Neighbor For each data point to be sampled, the (single) nearest neighbor is determined. The neighbor's value is then assigned to the sampled point. For sparsely available original data, this approach is ineffective in extrapolating data in areas where there is no data. For the data densities encountered in the available data, nearest neighbor interpolation can be used; the F550, F610 and F631 data sets in this study were preprocessed using this method.

Interpolation with Inverse Distance For each data point to be sampled, the nearest neighbors are determined. An aggregate value for the sampled point is determined from these neighbors by weighting each with its inverse distance to the sampled point.

Kriging According to [Papritz and Stein, 1999] and [Stein, 1999], kriging denotes a body of techniques to predict data in Euclidean space. This is further defined as “the target quantity is estimated at an arbitrary location, given its coordinates and some observations recorded at a set of known locations”. Kriging requires a variogram (cp. Section 2.7.3) and relies on least squares to produce unbiased estimates minimizing the squared difference between the estimated and the true value. The data sets F440 and F611 in this study were preprocessed using the proprietary kriging methods in the commercially available SSTtoolbox from SST software.⁴

Stochastic Simulation Natural data usually exhibit high variability and uncertainty. Hence, a smooth and unique map for these data is unlikely to exist. Therefore, variogram-based stochastic simulation is performed, taking the uncertainty into account. This results in a number of possible maps, each of them being valid in their own right. To be able to make decisions based on this type of maps, they are often aggregated, much like climate prediction maps [Hansen, 2002].

⁴<http://www.sstsoftware.com>

2.5 Data Available for this Study

This thesis utilizes five data sets from precision agriculture research sites in Northern Germany, in the years from 2003 onwards. The sites were planted with winter wheat, which is typically treated with three applications of nitrogen fertilizer at specific growth stages during the growing season. This section provides the necessary details on these data sets. It starts with the locations of the fields of study, describes the data acquisition and the data variables. At the end of this section, the five available data sets and their variables are summarized. The data sets were acquired from Martin Schneider and Peter Wagner from Martin-Luther-Universität Halle-Wittenberg, Lehrstuhl für landwirtschaftliche Betriebslehre, Germany.

2.5.1 Site Location

The five fields of study in this thesis were located in Northern Germany, near Köthen. They were positioned in an area ranging from 51.658 to 51.713°N and from 11.838 to 12.015°E, which is depicted in Figure 2.3. The sites exhibited similar characteristics in terms of soil composition and grew the same crop.

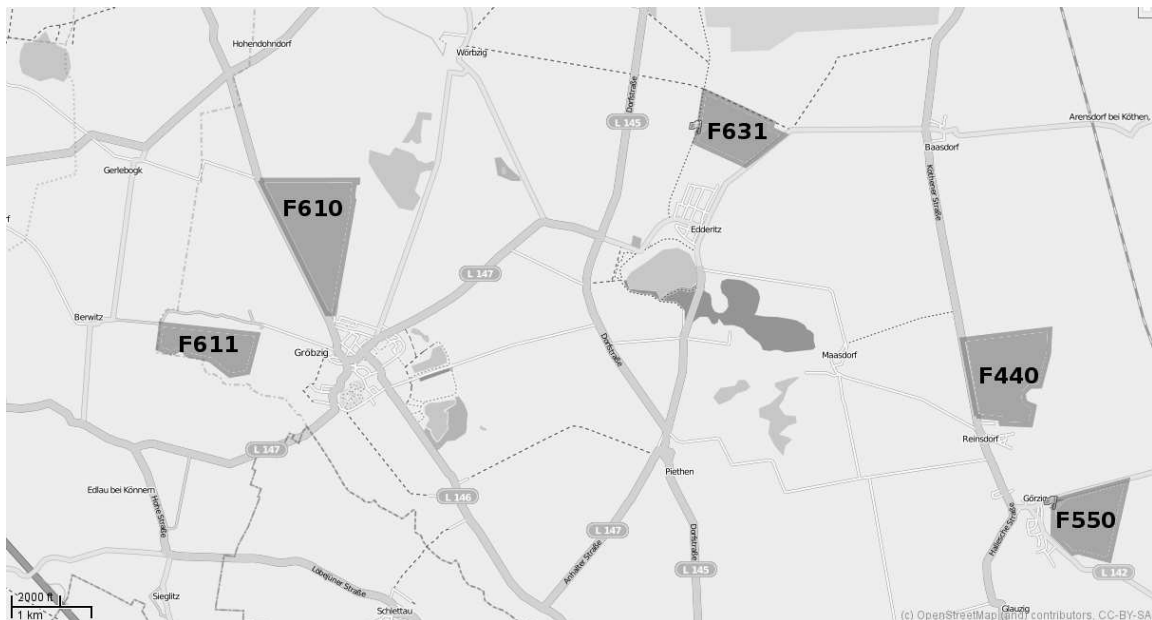


Figure 2.3: The five fields of study; the geographical area of this map covers roughly 51.658 to 51.713 degrees North and 11.838 to 12.015 degrees East. Reproduced on map material from openstreetmap.org, licensed under Creative Commons Attribution-ShareAlike 2.0 license.

2.5.2 Experimental Field Layouts

The five sites under study were experimentally partitioned into multiple test strips for comparing different fertilization strategies. This requires strategies such as blocking and split-plot design [Potvin, 2001]. In on-farm experiments, the effects of different factors can be assessed, for example the outcome of variable rate technology (VRT) [Brenning et al., 2008]. Earlier studies on this topic dealt with the econometrics of this approach (cp. [Anselin et al., 2004]) while the issues with spatial autocorrelation were mentioned in a case study for spatial regression [Lambert et al., 2003]. These studies typically relied on trial data which were collected by on-farm research operations employing different fertilization strategies. These strategies are given in the following, although not all of the strategies were executed on a single field. The variables directly affected by the strategies are the three fertilizer applications N1, N2 and N3, described in Section 2.5.4.

constant/company A constant amount of fertilizer was applied and distributed uniformly, typically $70/50/50 \text{ kg ha}^{-1}$ for N1/N2/N3.

mapping First, the yield potential of the site was determined. N1 was applied uniformly, typically at 70 kg ha^{-1} . The yield potential map with the levels *low*, *medium*, *high* determined the amount of N2 and N3, which were varied accordingly around the average nitrogen amount of the *constant* strategy.

nitrogen trial To estimate the plants' nitrogen uptake, special trials were run which consisted of applying a large range of nitrogen amounts. For N1, this traversed a range from as low as 25 kg ha^{-1} to 105 kg ha^{-1} , while for N2 and N3 the low range extended to 0 kg ha^{-1} , while the upper bound was similar.

neural network The fertilizer applications were determined via a trained neural network, as described in [Weigert, 2006].

sensor The Yara N-Sensor⁵ was used to determine N2 and N3 based on the vegetation index (REIP32 and REIP49 values) at two time points in the growing season. N1 was applied uniformly at 60 kg ha^{-1} .

Considering any of the data sets in this study, having a few strategies carried out on a particular field may lead to problems, for example in a yield prediction setup. Any of the variables may have an influence on yield, but since those variables are often determined via the abovementioned strategies, the variable STRATEGY should also be added to see its effects as a possible confounder.

Since the focus in this thesis is on exploiting the PA data sets from a data mining point of view, the actual study design and field layout are not of primary concern. Nevertheless, care must be taken in the analysis of the results due to the experimental design of the sites. Further information on this can be found in [Scheiner and Gurevitch, 2001] and [Potvin, 2001].

⁵The Yara N-Sensor is a trademark of Yara International ASA, Oslo, Norway

2.5.3 Yield

For the five fields of this study, small-scale yield data were available in different spatial resolutions from the years 2003, 2004, 2007 and 2008, in the combinations laid out in the data set details in Section 2.6. For each of the data sets except F550, the yield data also served as the reference points which the other variables are collocated with by using interpolation techniques. Yield was recorded along the harvesting lanes spaced 8 m apart. One yield record therefore covered an area of roughly 100 m². For the F550 site, the support points were not determined by yield, but rather located on a regular hexagonal grid as the hexagons' centers.

2.5.4 Nitrogen Fertilizer – N1, N2, N3

In Northwest Europe it is common practice that the nitrogen fertilizer for winter wheat, which is the crop considered in this study, is split into three doses [Neeteson, 1995]. The first dose (N1) is applied at tillering (stage 21 in Figure 2.1), the second dose (N2) at stem elongation (stage 32) and the third dose (N3) when the flag leaf appears (stage 51). The amount of nitrogen fertilizer applied was measured and geo-referenced by the available variable rate technology. Since the sites of study were designed as experimental sites for data collection, the range of N1, N2, and N3 in the data sets was typically from 0 to 100 kg ha⁻¹, while it is normally at around 60 kg ha⁻¹. The so-called *variable rate application* was the practical use of precision agriculture technology applied here. Nitrogen fertilizer mapping falls into the category of *input mapping* laid out before. The georeferenced raw point data are then interpolated and collocated with yield for further processing.

2.5.5 Vegetation – REIP32, REIP49

The *red edge inflection point* (REIP) is a vegetation indicator. Dedicated REIP sensors were used in-season to measure the plants' reflection in this spectral band. In this study, the Yara N-Sensor was used to obtain the REIP values at growing stages 32 and 49. It measured light reflectance from the crop from four different angles (45, 135, 225 and 315 degrees), covering a total area of approximately 50 m². Measurements were taken every second at normal normal working speeds. Different light conditions were compensated for by a fifth sensor positioned skywards.

2.5.6 Apparent Electromagnetic Soil Conductivity – EC25

In this work, the apparent electrical conductivity measurements were collected for the data sets using commercial sensors such as the EM-38⁶. These are designed for agricultural use and can measure small-scale conductivity to a depth of about 1.5 metres in vertical mode. There is no possibility of interpreting these sensor data directly in terms of their meaningfulness as yield-influencing factor. But in connection with other site-specific data, empirical relationships may be observed. The respective variable is called EC25.

⁶trademark of Geonics Ltd, Ontario, Canada

2.5.7 Soil Sampling – pH, P, Mg, K

In one of the data sets (F550), soil sampling variables from 2007 were available. These were obtained by taking core samples on a fixed grid at regular intervals. A circular sampling pattern was adopted, with the centers of the circles roughly 25 m apart and a circle diameter of 25 m. To account for local deviations within this circle, 15 samples were taken in a circular pattern around the center at an angle of 24 degrees. These samples were mixed and lab-analyzed for the pH, P, Mg and K measurements.

2.5.8 Terrain Attributes derived from Digital Elevation Models

In addition to the aforementioned variables which resulted from agricultural operations, a digital elevation model (DEM) at a 1-m spatial resolution for the five fields under study was provided by LVerGeo⁷. The DEM originally consisted of coordinate triples (x,y,z) on a 1-m grid covering the sites under study. To remove noise from the DEM, it was smoothed using a Gaussian filter at a standard deviation of 20 meters before the calculation of the aforementioned terrain attributes took place. Using a geographical information system⁸, terrain attributes were computed from this grid. The terrain attributes were added to the spatial data sets by a B-Spline interpolation. Hence, the existing yield points were the support points for the interpolation for all sites except F550, where the regular hexagonal grid the data were provided on was used. The actual terrain attributes are described in the following. Their descriptions are mainly based on [de Smith et al., 2009] and the SAGA documentation. Further information on the DEM processing can be found in [Olaya and Conrad, 2009].

Slope Loosely speaking, the SLOPE of a terrain surface is the amount of rise over run in the direction of the steepest descent. Slope is anisotropic. The quadratic surface method of [Zevenbergen and Thorne, 1987] was used for slope calculation.

Curvature The values for CURVATURE depend upon the line or plane along which they are calculated. At an arbitrary point, a plane drawn in the z -direction oriented in the ASPECT'S direction and passing through (x,y,z) describes the PROFILE CURVATURE. It describes the shape of the surface in the immediate neighborhood of the sample point and represents the rate of change of the slope at that point in the vertical plane. It is negative if the shape is concave, positive if the shape is convex and zero if there is no slope [de Smith et al., 2009, pp. 328–332]. At the same point, a plane constructed orthogonally to the previous plane, slicing the surface horizontally, describes PLAN CURVATURE. It is basically the curvature of a contour line at (x,y,z). While the first plane maximizes gravitational processes (maximum slope), the latter plane minimizes both.

Catchment Area/Slope The CATCHMENT AREA is the size of a point's upslope contributing area. The underlying idea to calculate the CATCHMENT AREA is that of assuming a uniform pattern of raindrops falling in the study region and analyzing the resulting water flows. This allows for finding watersheds (where water streams in adjacent cells flow away

⁷SAGA GIS, available from <http://www.saga-gis.org>

⁸Landesamt für Vermessung und Geoinformation Sachsen-Anhalt, Otto-von-Guericke-Straße 15, 39104 Magdeburg, Germany

from each other) and stream basins, which describe the CATCHMENT AREA of a stream. The CATCHMENT SLOPE is a measure describing the slope in this particular catchment area. Steep slopes may lead to higher surface drainage and soil erosion. The MODIFIED CATCHMENT AREA is a SAGA-specific variable that does not consider the water flow as a very thin film. As a result, it creates a more realistic approximation of soil moisture in valley floors [Böhner et al., 2002].

Wetness Index According to [Sørensen et al., 2005], topography affects the spatial distribution of soil moisture, and groundwater flow often follows surface topography. Topographic wetness indices, such as the basic TWI developed by [Beven and Kirkby, 1979], are therefore used to describe the topography related to soil moisture and wetness. The TWI has been found to correlate with soil pH and depth to groundwater. There are numerous deviations from and improvements to the basic TWI value. The SAGA wetness index was used here. For an in-depth discussion and overview about different TWI calculations, compare [Sørensen et al., 2005]. The direct comparison between the TWI and the SAGA WETNESS INDEX is provided in [Böhner et al., 2002].

2.6 Summary: Data Sets

This section lays out the five available data sets, which differ in size, location, field layout and the available variables. Important differences and issues about the data sets are mentioned, where appropriate. The respective variable plots can be found in Section A. The data sets are georeferenced such that each data record has a unique location on the site, expressed as a triple of (x,y,z)-values.

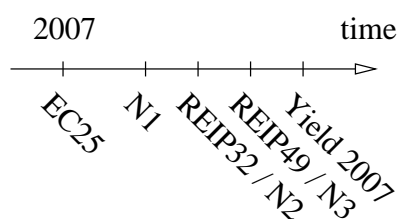
2.6.1 F440

F440 was an irregularly shaped field roughly 65 ha in size, covered by 6446 data records each roughly representing an area of 10m×10m. The data were collected during the growing season of 2007. The variables are shown in Figures A.1a to A.2c, starting on page 174. The temporal relationships in the data can be seen in Figure 2.4a. The EC25 variable visually showed distinctive spatial autocorrelation, partly also visible in the REIP32 and REIP49 variables. N1 had four distinct values (50, 57, 60, 70 kg ha⁻¹), also showing certain strips the field is divided into. N2 and N3 were differently distributed, with 45 and 50 distinct values covering a range from 0 to 95 kg ha⁻¹. A different crop variety was planted in the south of the field, accounting for the visible sharp cut in the REIP and YIELD variables. This introduced an additional variable SORTC. A summary for F440 is provided in Table A.1.

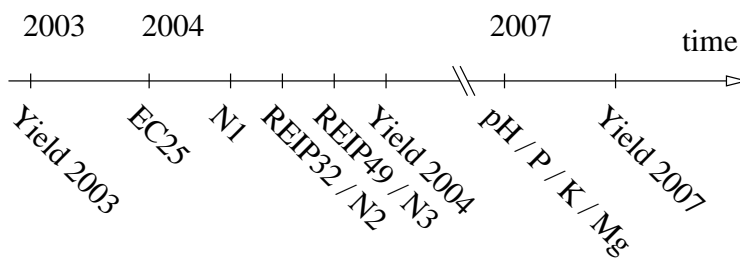
2.6.2 F550

F550 was a partly triangular shaped field located nearby the F440 field. This field was 67 ha in size, covered by 1080 data records, laid out on a triangular grid, with the grid cell centres spaced 25 metres apart. The data resolution was therefore much lower than in the other data sets. This is due to the availability of soil sampling data on this field, which are expensive to obtain at higher spatial resolutions. Data on this site were available from multiple years, provided in Figures A.3a to A.6d on pages 177–180, and in their temporal relationship in

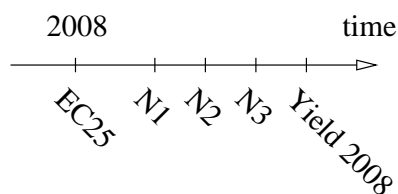
Figure 2.4b. Yield was provided in 2003, 2004 and 2007, where a pronounced zone is visible in Figure A.3c in the northern part of the field. N1 was applied almost uniformly, while N2 and N3 showed some visible strips running parallel to the northern border. Visually, EC25 showed spatial autocorrelation, which was less pronounced in the REIP variables. The soil sampling variables (PH, P, MG, K) exhibited similar characteristics, with pronounced visible zones, separated by a roughly cross-shaped area in the center of the field. A summary for F550 is provided in Table A.2.



(a) Data acquisition times F440/F611



(b) Data acquisition times for F550



(c) Data acquisition times F610/F631

Figure 2.4: Temporal relationships (data acquisition times) between variables in the five data sets under study. Different subsets of those data sets were used in the analyses in this study. Both temporal splits (subsets along the time axis) and splits according to strategies carried out were used.

2.6.3 F610

F610 was a larger field (around 100 ha) with a triangular shape consisting of 3996 data records. The available variables were EC25, current year's yield YIELD2008 and the three fertilizer applications N1, N2, N3. Although the site was in principle larger than the other sites, it exhibited a few peculiarities with missing strips of data. Those were not further cleaned or interpolated but it was rather decided to work with those data as-is. Figures A.7a to A.8c on pages 182 – 183 show those variables. A summary for F610 is provided in Table A.3.

2.6.4 F611

F611 was a field 50 ha in size, located to the west of the study area. The available 4970 data records consisted of the same variables in the same year as those for F440. See Figure 2.4a for the data variables. The variables are depicted in Figures A.9a to A.11c. N1 was applied in strips running from east to west, while the application of N2 and N3 was subject to different strategies, also visible in strips. There was visible spatial autocorrelation in EC25 and the REIP variables, as well as in the YIELD07 variable. A summary for F611 is provided in Table A.4.

2.6.5 F631

F631 was a field in the northern center of the study area, consisting of 7875 data records on a 55 ha area. In terms of available variables, this was the smallest data set (similar to F610), consisting of N1, N3, YIELD08 and EC25, shown in Figures A.12a to A.14 on pages 189f. The temporal aspects were the same as with the F440 and F611 data sets. Cp. Figure 2.4c for the data timeline. A summary for F631 is provided in Table A.5.

2.7 Spatial Statistics and Spatial Autocorrelation

Global spatial statistics look for an overall pattern between spatial proximity and the similarity of values. These statistics provide a single value that describes the spatial autocorrelation of the dataset as a whole. According to [Griffith, 2003], *spatial autocorrelation* is the correlation among values of a single variable strictly attributable to the proximity of those values in geographic space, introducing a deviation from the independent observations assumption of classical statistics. Spatial autocorrelation appears in such diverse areas as econometrics [Anselin, 2001], geostatistics [Cressie, 1993] and social sciences [Goodchild et al., 2000], among others. In most of the available precision agriculture data sets in this study, spatial autocorrelation exists, as demonstrated in Figures 2.5 and 2.6. Different measures for assessing spatial autocorrelation are briefly outlined in the following. The first of those statistics is the Moran's I coefficient, which can also be depicted in a Moran scatterplot to visualise spatial autocorrelation. The empirical semivariogram is the third tool, followed by contour maps and choropleth maps.

2.7.1 Moran's I

A classical measure for spatial autocorrelation is Moran's I [Moran, 1950], also called the Moran Coefficient (MC). Moran's I is defined as Eq.2.3, where n is the number of observations for which spatial autocorrelation should be determined. The c_{ij} are the weights in an n -by- n binary geographic connectivity matrix: if two locations are neighbors, then $c_{ij} = 1$, otherwise $c_{ij} = 0$. y is the variable for which spatial autocorrelation should be determined (cp. [de Smith et al., 2009]).

$$MC = \frac{n \sum_{i=1}^n \sum_{j=1}^n c_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n \sum_{j=1}^n c_{ij}} \quad (2.3)$$

The Moran Coefficient's values range from -1 , indicating negative autocorrelation, to $+1$, meaning perfect positive correlation. A value of 0 indicates a random spatial pattern. Generally, negative values of MC indicate negative spatial autocorrelation, and vice versa for positive values.

2.7.2 Moran Scatter Plot

The Moran scatter plot visualizes type and strength of spatial autocorrelation in a data distribution. It uses the Moran's I value to determine the extent of linear association between the values in a given location (x-axis) with values of the same variable in neighboring locations (y-axis). For calculating the Moran scatter plot, a spatially lagged transformation of a variable (y-axis) on the original standardized variable (x-axis) is regressed. This allows to compare a location's values with its neighboring values. For the EC25 variable of the F440 data set, a Moran scatter plot is depicted in Figure 2.5.

2.7.3 Empirical Semivariograms

While autocorrelation statistics such as the Moran's I value provide an indication of the local homogeneity of a dataset, it is sometimes interesting to understand how autocorrelation changes with distance. This can be examined using the semivariogram.

In the theory of geostatistics, the variable of interest is treated as a random variable. Hence, at each point \mathbf{x} in space there is a series of values for a property $Z(\mathbf{x})$. The observed value $z(\mathbf{x})$ is then drawn at random according to some law, from some probability distribution. Therefore, at location \mathbf{x} a property $Z(\mathbf{x})$ is a random variable with mean μ and variance σ^2 . With the spatial data sets in this study, variables at places near to one another tend to be spatially dependent. This spatial dependence can be described by the spatial covariance for a random variable, as in Equation 2.4.

$$C(\mathbf{x}_1, \mathbf{x}_2) = E[\{Z(\mathbf{x}_1) - \mu(\mathbf{x}_1)\}\{Z(\mathbf{x}_2) - \mu(\mathbf{x}_2)\}] \quad (2.4)$$

Assuming that the mean $\mu = E[Z(\mathbf{x})]$ is constant for all \mathbf{x} , and assuming $\mathbf{x}_1 = \mathbf{x}_2$, Equation 2.4 defines the variance $\sigma^2 = E[\{Z(\mathbf{x}) - \mu\}^2]$, which is the same everywhere. When $\mathbf{x}_1 \neq \mathbf{x}_2$, their covariance depends only on their separation $\mathbf{h} = \mathbf{x}_i - \mathbf{x}_j$, which is the

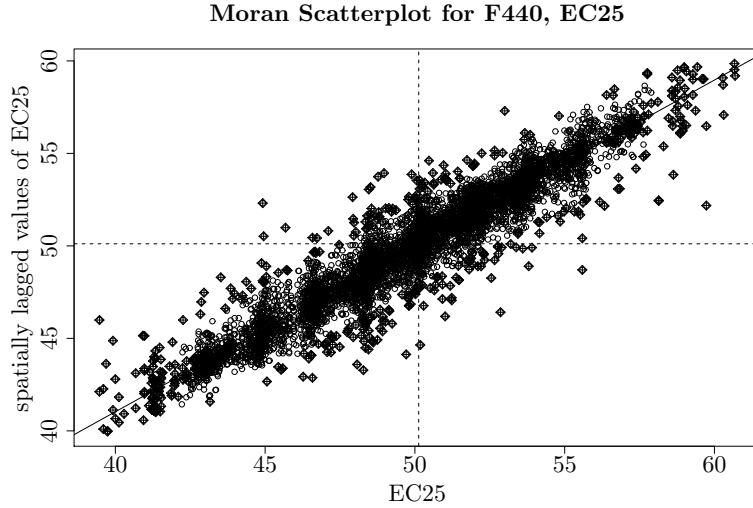


Figure 2.5: Moran scatter plot for EC25 for the F440 site. The scatter plot's slope corresponds to the value for Moran's I. The four quadrants of the scatter plot describe an observed value in relation to its neighbors. The top right and bottom left quadrants represent positive spatial autocorrelation, while the top left and bottom right quadrants represent negative spatial autocorrelation.

lag vector consisting of distance and direction. Hence, the covariance $C(\mathbf{x}_i, \mathbf{x}_j)$ is defined as in Equation 2.5. Independence of the first and second moments of the process with respect to location constitutes second-order stationarity [Oliver, 2010a].

$$\begin{aligned}
 C(\mathbf{x}_i, \mathbf{x}_j) &= E[\{Z(\mathbf{x}_i) - \mu\}\{Z(\mathbf{x}_j) - \mu\}] \\
 &= E[\{Z(\mathbf{x})\}\{Z(\mathbf{x} + \mathbf{h})\} - \mu^2] \\
 &= C(\mathbf{h})
 \end{aligned}
 \tag{2.5}$$

However, mean and covariance typically deviate from this assumption. Hence, second-order stationarity may be too strong an assumption under practical circumstances. Assuming instead that for small distances the general mean is constant (intrinsic hypothesis, cp. [Matheron, 1963; Oliver, 2010a]), the expected differences are zero (Equation 2.6).

$$E[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] = 0 \tag{2.6}$$

Replacing the covariances by the variances of differences, which (like the covariance) also depend only on the lag, leads to the semivariance (Equation 2.7) [Webster and Oliver, 2007].

$$\begin{aligned}
 \text{var}[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] &= E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2] \\
 &= 2\gamma(\mathbf{h})
 \end{aligned}
 \tag{2.7}$$

$\gamma(\mathbf{h})$ is termed the *semivariogram*. In practice, the semivariogram is typically estimated using the Method-of-Moments estimator [Oliver, 2010a] yielding the semivariances, shown in Equation 2.8.

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\}^2 \quad (2.8)$$

In Equation 2.8, the $z(\mathbf{x}_i)$ and $z(\mathbf{x}_i + \mathbf{h})$ values are the realizations of Z at places \mathbf{x}_i and $\mathbf{x}_i + \mathbf{h}$, and $m(\mathbf{h})$ is the number of paired comparisons at lag \mathbf{h} [Oliver, 2010a]. Changing \mathbf{h} yields the *experimental* or *sample variogram*. Its visualization consists of the semivariance calculated at various lag distances displayed against the lag. For further information regarding the semivariogram, it is referred to [Clark, 1979; Cressie, 1993].

If the semivariance depends exclusively on $\|\mathbf{h}\|$, i.e. the length of \mathbf{h} , the result is an omnidirectional variogram. However, in most natural settings, the change in a variable of interest is expected to be different when observing it in different geographical directions. This behaviour is called anisotropy. As an example, consider Figure A.1d (page 174): the visible change in the distribution of the EC25 variable is much higher in a north-east to south-west direction than in a north-west to south-east direction. Therefore, a directional variogram can be calculated which captures the effects of direction on the variance in the data set.

An illustrative example of two omnidirectional experimental variograms from the F440 and F611 data sets is shown in Figure 2.6. For the same variable, the two fields exhibit different autocorrelation behaviour. For EC25 on the F440 field, the variogram shows strong positive spatial autocorrelation, while for EC25 on the F611 field it appears to show negative spatial autocorrelation. Empirical semivariograms are used to fit a theoretical variogram which can then be used for other geostatistical techniques such as kriging (cp. Section 2.4.8).

2.7.4 Contour Map

Each data record of a georeferenced data set can be attached to a single point on a two-dimensional surface. Connecting nearby values of equal quantity with lines yields a generalized continuous surface. These *contour lines* have the property that all values on one side of the line are greater than the line's value, whereas all values on the other side are smaller than its value. Contour lines can be drawn on maps with a uniform interval of vertical distance separating them – this allows a hill or valley to be visualized as a series of concentric loops converging towards a point. For basic data analysis tasks and a quick visualization, these maps are useful. Contour maps may also occur during the preprocessing, which is further elaborated upon in Section 2.4.8. Two contour maps are shown in Figures 2.7a and 2.7b. In Chapter 4, maps similar to contour maps are produced in a management zone delineation approach.

2.7.5 Choropleth Map

In contrast to the previously shown contour map, a choropleth map is a thematic map of an area under consideration. The main difference to a contour map is that its region boundaries are not defined by the data set itself, but rather by an appropriate manual division into

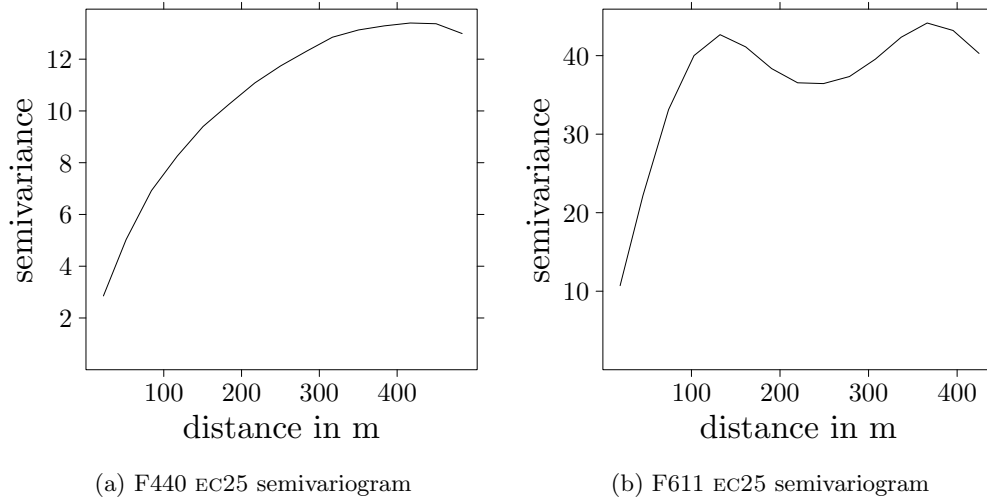
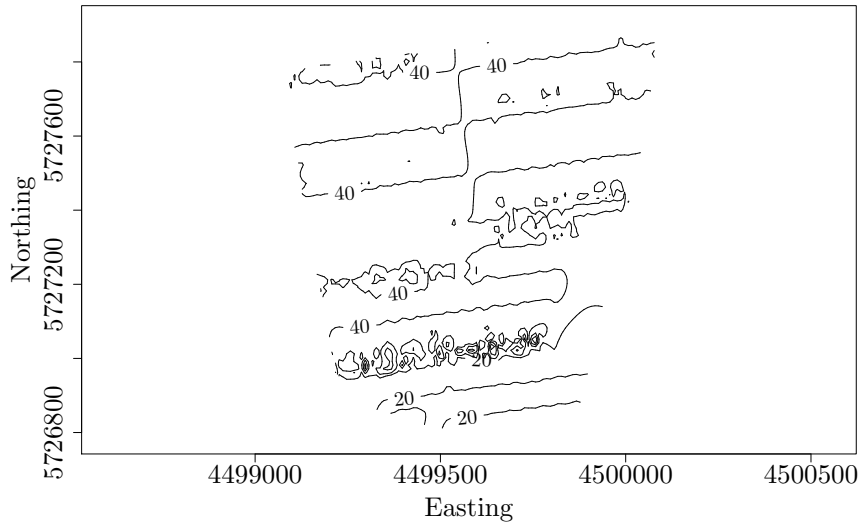
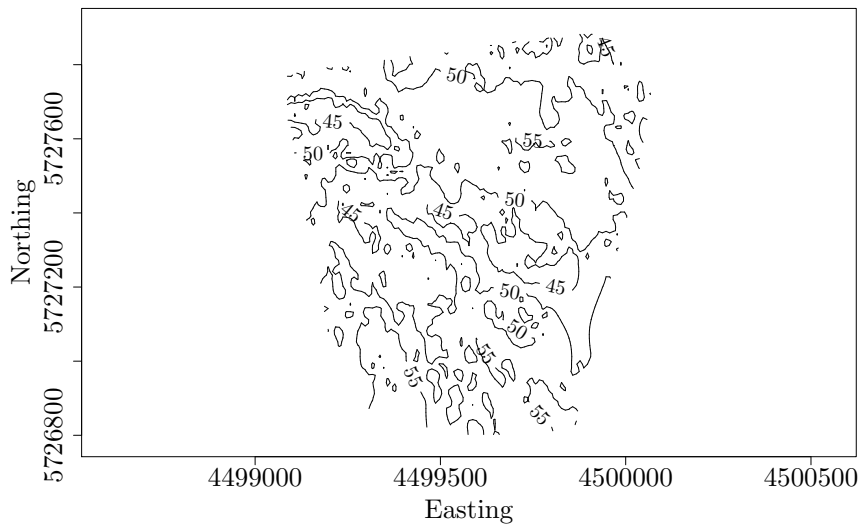


Figure 2.6: Variograms for (a) EC25 on F440 and (b) EC25 on F611. Both graphs depict distance vs. semivariance. (a) exhibits positive spatial autocorrelation whereas (b) shows negative spatial autocorrelation. Cp. Figures A.1d and A.10a in the Appendix.

sub-areas, which is mostly done by considering external circumstances or expert knowledge. In the case of the currently considered agriculture data, this could be a partitioning of the field according to farmer's experience, e.g. by rule-of-thumb considering high, medium and low yield areas. Choropleth maps are the result of a novel management zone delineation approach laid out in in Chapter 4.



(a) F440 N3 values, contour map



(b) F440 EC25 values, contour map

Figure 2.7: Contour maps for N3 and EC25 of the F440 site. (cp. Figures A.1c and A.1d on page 174f). Spatial autocorrelation is (visually) much less pronounced in the N3 figure with sharp visible zone borders, while it is rather strong in the EC25 figure. Figures were produced using R code from [Bivand et al., 2008].

2.8 Temporal Relationships in the Data

Precision agriculture data sets usually consist of multiple variables, which are recorded by different devices. If those devices record the data simultaneously, a snapshot of the field's current state is acquired. However, this snapshot is normally repeated during the growing season which provides more data and introduces *temporal* relationships into the data set, similar to the acquisition of a timeseries. For example, a vegetation indicator which is measured at one stage of crop growth is likely to be autocorrelated with the same indicator measured a certain time period later. This is especially true for multi-year data sets, which result from one field being observed during precision agriculture operations in multiple (consecutive) years. It would therefore be definitely advantageous to perform temporal data mining based on these aspects, from a computer science point of view.

From a precision agriculture point of view, however, this discipline is currently in the early adoption phase, which leads to the issue that multi-year data sets are not available by default. Therefore, methods of temporal data mining might certainly be desirable and should be investigated, but would take rather long to be transferred into practice on most fields.

Further practical reasons for neglecting the temporal aspects are crop rotation and weather effects. Small-scale yield data for different crops planted in different years at the same site are typically not comparable. Even in monocultural setups, yields typically differ considerably due to different weather and precipitation conditions. Hence, temporal aspects in the multivariate data sets which are encountered in this work are not taken into account. In other words, this means that usually less than two site-years are being used in the analysis, although more may be available.

Chapter 3

Assessing the Importance of Data Sources for Yield Prediction

3.1 Introduction

A core task in agriculture is yield prediction, i.e., using in-season predictive modeling to estimate the yield level at the end of the growing season. Traditionally, due to the lack of small-scale data, yield has usually been estimated uniformly for a whole site, neglecting most fields' spatial heterogeneity. With the advent of precision agriculture and small-scale data, this task may be advanced towards predicting yield on a much smaller scale. While a lot of uncertainties remain, especially due to unpredictable weather and precipitation conditions, small-scale yield prediction is turning into a data-driven regression task using PA data sets. Nevertheless, the spatial nature of the data sets is often neglected and leads to serious issues with non-spatial regression models. Furthermore, taking yield prediction one step further results in the question of *which of the available data variables are actually important for yield prediction*. Although most of the data variables are well reasoned for in terms of their usefulness for yield prediction from the agriculture point of view, this has not yet been determined from a data mining point of view.

This chapter provides insights into answering the above main question. The first part develops a regression setting which enables the spatial evaluation of advanced regression models. The second part takes advantage of this regression approach and provide answers to the question of which of the available data variables are relevant for yield prediction. Figure 3.1 outlines this chapter.

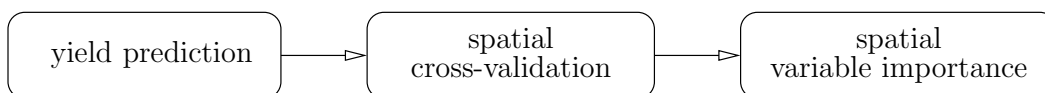


Figure 3.1: Chapter outline

The main motivation for this work is based on the experiences of [Weigert, 2006]. Therefore, the context of [Weigert, 2006] is briefly summarized in the following, while a shortened version is also available in [Weigert and Wagner, 2003].

In his work, the author aims at finding algorithms to evaluate site-specific data with respect to yield prediction. His work is a showcase of collecting experiences and solutions to the specific problem of small-scale nitrogen fertilization. The main target of his work is to generate decision rules for fertilization, which, based on available data and their algorithmic evaluation, have been economically optimized. He describes this task as a problem of supervised learning, which is common in the area of data mining. Before the data mining step, the author tackles the issue of data preparation, which begins with the raw field data and, after some steps, ends with data that can be fed to data mining algorithms. For the modeling stage an artificial neural network (multi-layer perceptron) and a decision tree are used. The models are verified using cross-validation. From those regression models, decision rules are generated which are optimized economically. Furthermore, patterns that could describe agricultural relationships are derived, such as decision rules [Schneider et al., 2006]. In short, the author's work covers a full data mining process, which includes the data preprocessing and cleaning, the mining itself and the usage of the mining results. This process is also laid out in two differently focused articles, [Wagner and Weigert, 2003] and [Weigert et al., 2003]. The economic advantages are quite clear, as laid out in [Bachmaier and Gandorfer, 2009; Biermacher et al., 2009]. It is clearly pointed out by [Weigert, 2006] that the need for further research into the data mining stage is necessary. The follow-up work of [Ruß et al., 2008a,b,c] elaborated upon the specifics of the chosen network model in the work of [Weigert, 2006]. Further regression models were introduced in [Ruß, 2009] and [Ruß et al., 2010b]. Nevertheless, the approaches that are nowadays most common to decide the amount of fertilizer do not use the available data to their full extent. Furthermore, given the available spatial data sets from precision agriculture, there exist a few modeling pitfalls which should be avoided. This issue is elaborated upon further in Section 3.2.3.

In summary, the current study extends the work of [Weigert, 2006] by evaluating the usage of regression models other than neural networks for yield prediction. It also extends existing work by presenting a unified approach to identify important yield prediction variables using arbitrary regression models.

3.2 Regression and Cross-Validation Made Spatial

3.2.1 Regression

[Hand et al., 2001] defines regression as follows:

The aim is to use a sample of objects, for which both the response variable and the predictor variables are known, to construct a model that will allow prediction of the numerical value of the response variable for a new case for which only the predictor variables are known.

From the machine learning and data mining perspective, the basis of regression is the inductive learning hypothesis [Mitchell, 1997]:

Any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples.

Since the areas of statistics and machine learning are in some respects closely related, but often employ different terms for the same aspects, a clarification of these terms is needed. For the data sets in this thesis, the response variable is YIELD at the end of the respective season, unless stated otherwise. The remaining variables are the predictors, although not all of them have to be used. In other contexts, the terms *dependent* or *target* for the response variable are used, while *independent*, *explanatory* or *regressor* are used for the predictor variables. Those can be numerical, but they need not be. Predictive accuracy is one of the most substantial properties of such regression models, therefore various measures of accuracy have been devised. In this work, the term (*modeling*) *error* is used for the deviation between the sample and the estimated model, while the term *residual* is used in statistics.

Although predictive accuracy is a critical aspect of models, it is not the only aspect. A regression model might provide insights into which of the predictor variables are most important. From expert knowledge it may be known that some predictors must be included, while others should not be. Different regression models often show different predictive accuracy, of which some may be attributed to the way the models work and some may be caused by the data themselves. Another aspect is interactions where the effect that one predictor has on the response variable depends on the values taken by other predictors.

For the regression models used in this thesis, the respective R implementations and specific model settings are provided in Appendix C.

3.2.2 Cross-Validation

The question asked in regression is how well a certain model built on a specific data set performs on an independent validation set. However, in practical setups a dedicated validation set is often not available. Furthermore, it is rather simple to obtain a model which perfectly fits the data it is built from, but which performs poorly on independent data. This issue is known as *overlearning* or *overfitting*. In order to avoid this, cross-validation is a technique for estimating the performance of a predictive model in case that a dedicated validation data set is unavailable. The underlying idea for cross-validation is that the data set which the model's performance is supposed to be evaluated on is partitioned into a training and test set. A predictive model is then trained using the training set and its performance is reported on the independently sampled test set.

Two standard procedures for partitioning the data set are *random sub-sampling* and *k-fold cross-validation*. For random sub-sampling, the data set is split randomly into training and test sets. The model is trained on the training set and the prediction error is reported on the test set. This is repeated a sufficient number of times and the errors are averaged. During the repetitions, the same samples may end up multiple times or not at all in either test or training set. In *k-fold cross-validation*, the parameter k determines the size of the split. The data set is randomly split into k partitions and $k - 1$ parts are used for training, while the remaining partition is used for testing. This is repeated k times such that each split is used exactly once for validation. The results are usually combined to yield a single estimate. $k = 10$ is commonly used.

Independence of the data records in the data set under study is assumed. Problems arise when this assumption is violated. For spatial data the sampling procedure should be

aware of spatial autocorrelation in the data sets. Otherwise, two data records which are very similar due to being spatially adjacent may end up in the training and the test set and lead to overfitting of the underlying model. This issue is discussed further in Section 3.2.3 where the cross-validation sampling procedure is adapted for spatial data sets.

For comparing the different models' performance, the root mean squared error is used, since it is the most common measure for estimating a predictor's performance. For a data set consisting of n observations (y_n, \mathbf{x}_n) , it is computed as shown in Equation 3.1. The difference between the actual response and the predicted response is squared. These squared errors are averaged and the root of the average is taken.

$$rmse = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_{i_{predicted}})^2} \quad (3.1)$$

3.2.3 Spatial Sampling for Regression

In order to account for the spatial nature of the data in a cross-validation setup, there are two main starting points. Those are shown in Figure 3.2. Given a spatial data set, performing a regression task in a cross-validation setting requires at least a sampling and a regression technique. While each of the regression techniques outlined in Section 3.3 may be modified in such a way as to incorporate spatial data sets, this is a rather uncommon approach. There are certainly specialized approaches for some of the regression techniques to make them account for spatial data sets, such as geographically weighted regression, as mentioned in Section 3.3.3. However, the idea here is to keep the standard regression techniques as-is and incorporate them into a spatial setup by changing the cross-validation sampling accordingly. This essentially leads to a sampling procedure adapted to the specifics of the available spatial data in precision agriculture. Nevertheless, it may be transferred to other disciplines where similar tasks with similar spatial data sets are encountered.

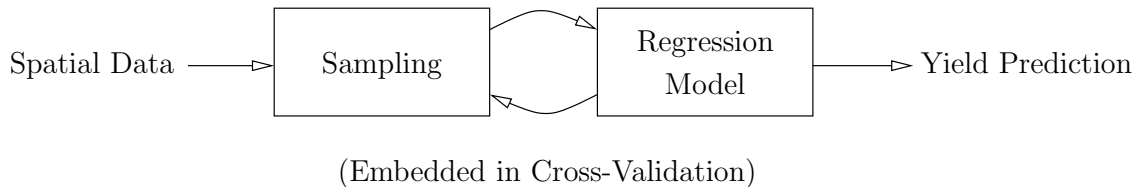


Figure 3.2: Regression steps; there are two spots where the spatial component in the data sets may be considered in a yield prediction setup; classically, each of the *regression* models would be adapted for the spatial case; in this work, the spatial component is accommodated in the *sampling* step where the data set is split up into training and testing sets for the cross-validation task.

3.2.4 Spatial Cross-Validation

Since data records in spatial data are likely to be spatially autocorrelated, data records must not be considered as independent. Adjacent data records are likely to be very similar, even identical data records are likely to exist without being the result of an erroneous preprocessing or data acquisition. In random sampling, very similar or even identical data records may end up regularly in training and test set, even in a *sampling without replacement* setup. If those data subsets are used as a regression training and test set, an arbitrary regression model is trained on data records which appear similarly in the test set. The model should normally try to generalize from the training records, while in this setting it may be sufficient to memorize the training records to obtain a good predictive performance.

Since the regression models are to be kept as-is, the sampling procedure is to be adapted for spatial data sets. In a k -fold cross-validation approach, the data set is partitioned into k parts of roughly equal size, of which $k - 1$ are used for training and the k -th is used as a test set. This random partitioning should now be turned into a spatial partitioning, i.e. a tessellation of the underlying agriculture field. Due to the data being usually preprocessed and therefore often on a fixed grid, a grid-based tessellation approach seems to suffice for this task. Considering the data closer, however, reveals several drawbacks to this tessellation approach:

skewness of the data grid Although the data are usually sampled in regular distances, this does not necessarily lead to a rectangular or hexagonal grid. Furthermore, there might be gaps between harvesting lanes and the grid might change from one field part to another. Therefore, a traditional grid would have to be tailored (rotation, size of the cells) anew for each particular field, hampering automatic data processing. A depiction can be found in Figure 3.3a.

points on grid borders The probability is high that a few points in each partition are situated exactly on the grid borders. This leads to ambiguities in point assignment. This issue may also be inferred from Figure 3.3a.

points on field borders Since the field borders are typically irregular there would also be irregular grid cuts at those outer borders. Furthermore, the partition sizes should be roughly equal which would not hold true for these outer grid parts. Further processing (i.e. merging certain partitions) would be necessary to overcome this issue. As an example, Figure A.2a exhibits the non-rectangular and irregular field shape of F440.

in-field irregularities Due to the nature of the data, there may be “holes” in the data – areas in the field for which no data records exist due to power poles, buildings or incomplete data acquisition. Interpolation of these holes should also only be done where a sufficient number of surrounding points exist. In a grid-based tessellation approach, these holes would have to be filtered manually after applying the grid. For example, cp. Figure A.10b, which shows such a “hole” in the northern half of the field.

For the above reasons, the grid-based partitioning approach is insufficient for the precision agriculture data since those are not typically as regular as they might seem. However,

a standard voronoi tessellation of the field would overcome the abovementioned drawbacks of a grid-based approach if it could be constructed flexibly, assuming that the spatial data density is similar throughout the field. This is essentially a task of spatial clustering, which is treated in more detail in Chapter 4. The input for a spatial clustering algorithm here are the data records' coordinates, while the predictor and response variables are of no concern for a random spatial tessellation.

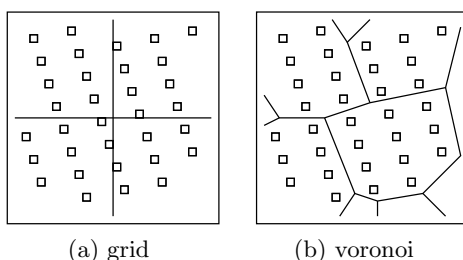


Figure 3.3: Comparison of grid vs. voronoi approach

An algorithm fit for this task is k -means clustering [MacQueen, 1967], shortly described in Algorithm 1. It provides a voronoi tessellation of the plane when given the data records' coordinates and a number k as an input. However, it is not guaranteed to find an optimal tessellation and is sensitive to initialization. While these properties are usually considered unsatisfactory for a clustering purpose, they are exploited in the setup presented here. Cross-validation requires a random partitioning to obtain a reliable estimator during multiple runs. Due to the instability of k -means, the resulting spatial partitioning is in principle unstable as well and therefore leads to slightly different tessellations. Figure 3.4a shows the F440 site spatially tessellated into ten clusters with the k -means procedure. Changing the cluster number from ten to nine (Figures 3.4a and 3.4b) leads to visible changes in the clustering. Therefore, even if k -means showed stable behavior, it would be rather simple to notably change the underlying tessellation by adapting the parameter k . Larger k are also possible: for 10-fold cross-validation and $k = 20$, 18 clusters could be used for training, while the remaining two are used for reporting the predictive error. For $k \rightarrow n$ (with n being the data set size), this procedure converges towards behaving like a non-spatial cross-validation setup.

Algorithm 1 k -means, adapted from [MacQueen, 1967].

- (i) Initialize cluster centroids: place k points into the space represented by the data records that are being clustered
 - (ii) Assign each record to the cluster that has the closest centroid.
 - (iii) Once all records have been assigned, recalculate the positions of the k centroids.
 - (iv) Repeat steps (ii) and (iii) until the centroids are stable. This creates a separation of the records into clusters from which the metric to be minimized can be calculated.
-

The parameter k for the k -means algorithm has to be determined empirically and tailored to the specific needs of the following cross-validation regression task. An upper bound

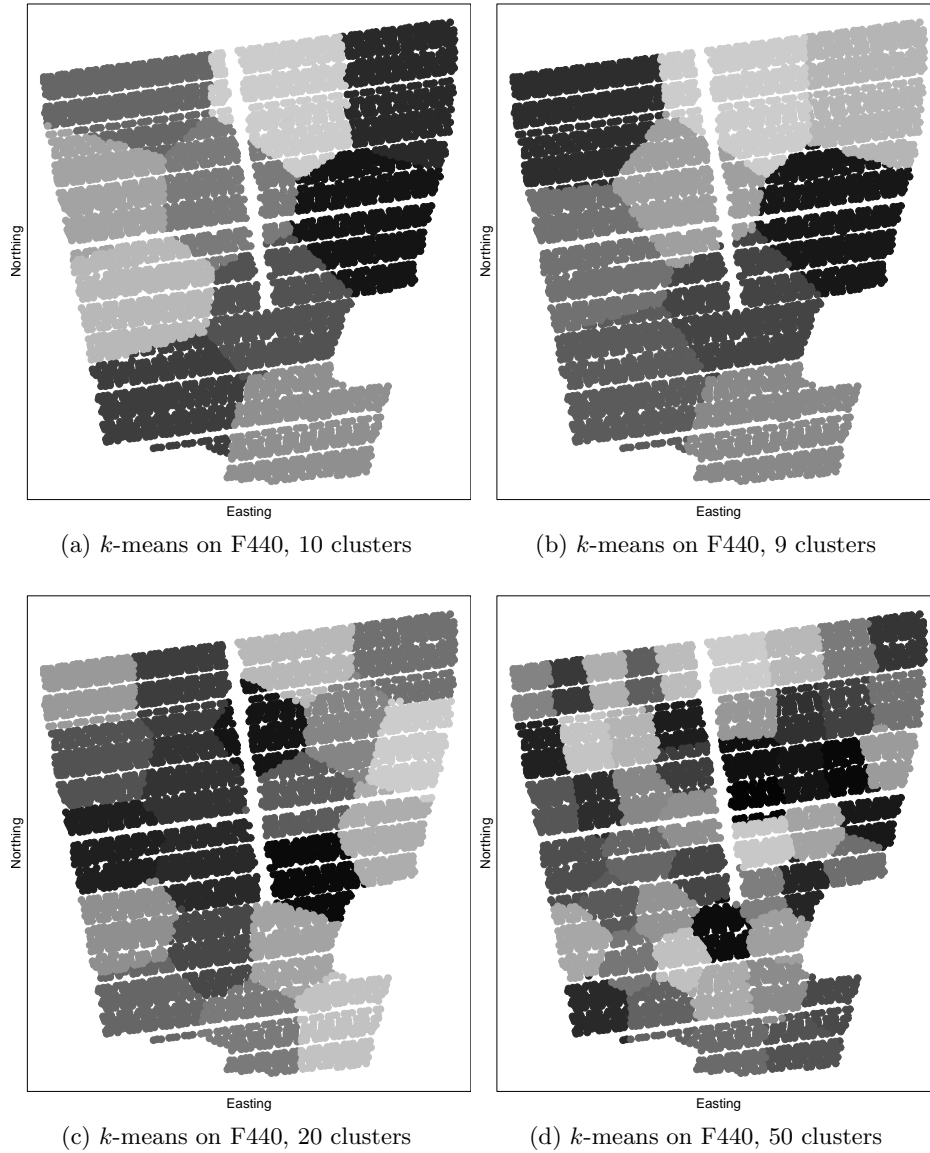


Figure 3.4: Different k -means results on F440. Note that the plot point sizes have been increased for better visual cluster recognition. Figures (a) and (b) show the notable difference between the results of changing the cluster number by 1. The higher k is set, the more the spatial cross-validation procedure shown here converges towards a random (non-spatial) sampling.

(smaller clusters) may be determined by the precision of the available farming equipment, whereas a lower bound (larger clusters) may be set to 10 for the cross-validation regression task. The objective function minimized by the k -means algorithm is the following:

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2 \quad (3.2)$$

where the distance measure between a point $x_i^{(j)}$ and the respective cluster centroid c_j is the Euclidean distance here, since the data records' coordinates are given in Euclidean space.

3.3 Regression Models

The spatial cross-validation approach shown in the previous section wraps around a generic regression technique requiring a training and a test mode. In training mode, it takes as inputs data vectors consisting of one or more predictor variables' values and an associated response value. In test mode, it takes as inputs data vectors consisting of one or more predictor variables' values and computes a response. Seven regression techniques are outlined in the following. These are used in the spatial variable importance approach shown in Section 3.5.

3.3.1 Linear Regression

Given a set of n data records $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^n$, linear regression assumes that the relationship between the response variable y_i and the vector of predictors x_i is approximately linear. For this purpose, a disturbance term ε_i is introduced. It adds noise to the linear relationship between the predictors and the response variable. For n data records, this results in n equations, which can be written in vector form as:

$$y = X\beta + \varepsilon \quad (3.3)$$

where y is a vector of responses, X a predictor matrix, β a vector of coefficients and ε a vector of disturbance terms.

Assume that b is a candidate value for the parameter β . Then the quantity $y_i - Xb$ is called the residual for the i -th observation. It measures the vertical distance between the data point (x_i, y_i) and the hyperplane $y = Xb$. Therefore, it assesses the degree of fit between the actual data and the model. The sum of squared residuals (SSR) is a measure of the overall model fit:

$$S(b) = (y - Xb)^T (y - Xb) \quad (3.4)$$

The value of b minimizing this equation is the ordinary least squares estimator (OLS) for β . $S(b)$ possesses a unique global optimum:

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^p} S(b) = (X^T X)^{-1} X^T y \quad (3.5)$$

Linear regression is used here as a baseline model for comparing advanced regression techniques against.

3.3.2 Generalized Additive Models

Generalized Additive Models (GAMs) are an extension to the aforementioned linear models. Developed by [Hastie, 1991], they replace the weighted sum $\sum x_j\beta_j$ in logistic regression (a generalized linear model) by $\sum f_j(x_j)$, where f_j is a non-parametric function. This function is estimated using a scatterplot smoother and can reveal possible nonlinearities in the different x_j 's effects. The basic idea behind the scatterplot smoother is the tradeoff between the goodness of fit of the estimated regression function and the smoothness of the function. A perfectly fit regression function is typically not smooth and generalizes rather badly due to the effect of overlearning (cp. Section 3.3.5 on artificial neural networks).

Mathematically, a cubic spline smoother is introduced which ensures the smoothness of $f(x)$. This leads to the function to minimize shown in Equation 3.6.

$$\sum (y_i - f(x_i))^2 + \lambda \int f''(x^2)dx \quad (3.6)$$

The solution to Equation 3.6 is a piecewise cubic polynomial joined at the observed values of x in the data set. However, λ is not set directly, but rather determined in reverse by setting a bound on the degrees of freedom on the cubic spline smoother and searching numerically for the appropriate value for λ . For multiple regressor variables, the additive model to fit is (Equation 3.7):

$$\hat{y}_i \approx \sum_j f_j(x_{ij}) \quad (3.7)$$

for which an analogous numerical solution as in the version with one regressor variable can be calculated in a so-called “backfitting” procedure.

3.3.3 k-Nearest-Neighbor

Suppose the training data in a regression setup are simply stored, without further evaluation. When an instance from the test set has to be predicted, the k most similar instances are retrieved from the training set. These k instances are aggregated and produce an estimate for the response variable in the test instance. This behavior is at the foundation of k -Nearest-Neighbor (kNN) learning methods. They are an example of *lazy learning*: the decision of how to generalize beyond the training data is deferred until each new test set instance is encountered [Mitchell, 1997]. Algorithm 2 shows the basic idea where the k nearest neighbors of the test set instance are averaged. The nearest neighbors are typically determined via the Euclidean distance between the instances.

While the idea behind kNN is rather simple, a straightforward question to ask is whether all of the k nearest neighbors should have the same influence on the value of the (averaged) response variable. This leads to *distance-weighted* kNN: here, the influence a single nearest neighbor has on the test set instance is weighted by the inverse distance between the respective training instances and the test set instance. The closer the two instances are, the more the response value is determined by the training instance. This requires a simple change to Algorithm 2, replacing Equation 3.8 by the following Equation 3.9:

$$\hat{f}(x_q) = \frac{\sum_{i=1}^k w_i f(x_i)}{\sum_{i=1}^k w_i} \quad (3.9)$$

Algorithm 2 k -Nearest-Neighbor Regression, adapted from [Mitchell, 1997]

/ Training phase:*/*

Add each training example $\{y_i, x_{i1}, \dots, x_{ip}\}$ to the list *training-examples*

/ Regression phase:*/*

Given a test set instance x_q to be predicted:

Let x_1, \dots, x_k denote the k instances from *training-examples* that are closest to x_q

Return

$$\hat{f}(x_q) = \frac{\sum_{i=1}^k f(x_i)}{k} \quad (3.8)$$

where the w_i are defined as $w_i = \frac{1}{d(x_q, x_i)^2}$, with d being an appropriate distance measure. In the case of distance-weighted kNN, the choice between a local and a global model can be made. In a local model, only the k nearest neighbors are considered, as in the standard kNN. For obtaining a global model, all training instances are considered for predicting a response. Nevertheless, standard kNN and distance-weighted kNN do not explicitly consider spatial relationships in the data sets. While locally weighted regression is a further generalization of the kNN approach, it does not straightforwardly apply to spatial data either [Cleveland and Devlin, 1988].

Therefore, for spatial data geographically weighted regression (GWR) has been developed [Fotheringham et al., 2002]. The idea with GWR is that for a regression model not only the actual predictors at one point have an influence on the response, but also neighboring points (weighted by their inverse distance – “geographically weighted”). This is similar to the kNN approach, while the distance between instances is not calculated in feature space (between the predictor vectors), but rather in geographical space. This assumes that the underlying process is not stationary, but spatially autocorrelated. The residuals of a global model are assumed to be (a) rather high and (b) positively spatially autocorrelated. The local model of GWR returns a better prediction which is locally adjusted and where the residuals are not spatially autocorrelated. However, GWR is specifically tailored to ordinary least-squares regression and can not easily be modified to fit other regression techniques, hence it is not used here.

For the data sets encountered in this thesis, the number of predictors is rather small. For data sets with higher numbers, the *curse of dimensionality* might come into play. Many of the predictors may be irrelevant for the response variable, but are still considered in the calculation of the response. Therefore, instances which should be considered close to each other in kNN regression are still rather distant when all predictors are considered. Two approaches towards tackling this issue consist of choosing weights for each of the predictors or otherwise determining the predictors’ importance, such as in Section 3.5.

3.3.4 Regression Trees

Learning decision trees is a paradigm of *inductive learning*: a model is built from data or observations according to some criteria. The model aims to learn a general rule from the observed instances. Decision trees can therefore accomplish two different tasks, depending

on whether the target attribute is discrete or continuous. In the first case, a classification tree would result, whereas in the second case a regression tree would be constructed. Since the focus is on solving a regression task, the regression tree is explained in the following.

Regression trees approximate learning instances by sorting them down the tree from the root to some leaf node, which provides the value of the target attribute. Each node in the tree represents a split of some attribute of the instance and each branch descending from that node corresponds to one part left or right of the split. The value of the target attribute for an instance is determined by starting at the root node of the tree and testing the attribute specified by this node. This determines whether to proceed left or right of the split. Then the algorithm moves down the tree and repeats the procedure with the respective subtree. In principle, there could be more than one split in a tree node, which would result in more than two subtrees per node. However, in this application scenario, regression trees with more than two subtrees per split node are not taken into consideration.

Regression as well as decision trees are usually constructed in a top-down, greedy search approach through the space of possible trees [Mitchell, 1997]. The basic algorithms for constructing such trees are CART [Breiman et al., 1984], ID3 [Quinlan, 1986] and its successor C4.5 [Quinlan, 1993]. The idea here is to ask the question “which attribute should be tested at the top of the tree?” To answer this question, each attribute is evaluated to determine how well it is suited to split the data. The best attribute is selected and used as the test node. This procedure is repeated for the subtrees. An attribute selection criterion that is employed by ID3 and C4.5 is the entropy and, resulting from it, the information gain. Entropy is a measure from information theory that describes the variety in a collection of data points: the higher the entropy, the higher the variety. An attribute split aims to lower the entropy of the two resulting split data sets. This reduction in entropy is called the information gain. For further information it is referred to [Mitchell, 1997].

However, if the addition of nodes is continued without a specific stopping criterion, the depth of the tree continues to grow until each tree leaf covers one instance of the training data set. This is certainly a perfect tree for the training data but is likely to be too specific – the problem of overlearning occurs. For new, unseen data, such a specific tree is likely to have a high predictive error. Therefore, regression trees are usually pruned to a particular depth which is a trade-off between high accuracy and high generality. This can be achieved by setting a lower bound for the number of instances covered by a single node below which no split should occur.

3.3.5 Neural Networks

According to [Mitchell, 1997], “neural networks provide a general, practical method for learning [...] vector-valued functions from examples.” Artificial neural networks (ANNs) are inspired by the biological foundations of a human brain’s inner workings. Highly interconnected neurons are assumed to learn by establishing and reinforcing their connections. ANNs emulate this neural structure and the learning process itself. For training, gradient descent is typically used, in conjunction with a backpropagation procedure. Multi-layer perceptrons (MLPs) are regularly used in regression and are laid out in the following.

Figure 3.5 illustrates a simple feedforward artificial neural network. The ANN learns an input-output mapping function from data. For this purpose, on the left side the inputs are

fed into the network, i.e. the values of the predictors. Each hidden unit is similar to a basic perceptron unit depicted in Figure 3.6. This perceptron aggregates the inputs, typically by a weighted sum. The aggregate value is fed into a function, which may be *tangens hyperbolicus*, a logistic function or similar functions. An additional output function may be applied. The perceptron returns a single output value which may then be fed into another perceptron for processing. In this manner, the input vector of predictors is propagated through the network. In the output layer, the estimated value of the response variable is obtained. The error between the network response and the actual response value is calculated. Based on this error, a procedure called *backpropagation* starts, whereby the error is sent layer-wise backwards through the network. This allows for a layer-wise adaptation of the connection weights and the bias values. Once the input layer is reached, another input vector is propagated through the network in the same way. This procedure follows the principle of *gradient descent*: the error is iteratively minimized along the steepest descent, but not necessarily towards the global minimum error. Generally, MLPs can be seen as a practical vehicle for performing a non-linear input-output mapping [Haykin, 1998]. Further details on this network type may also be obtained from [Hagan, 1995] and [Nauck et al., 2003].

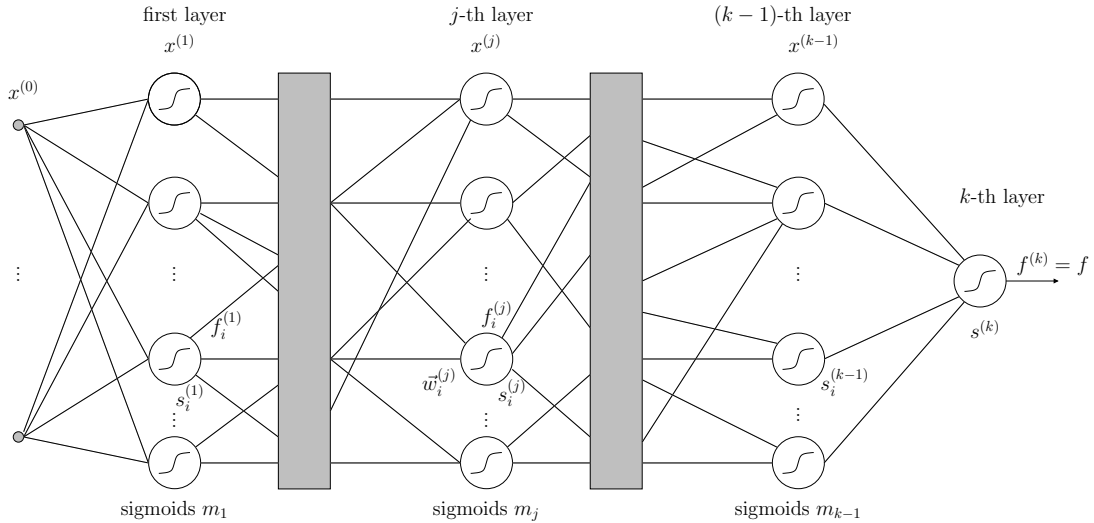


Figure 3.5: Schematic of a feedforward neural network. Each sigmoid is essentially a perceptron unit as depicted in Figure 3.6.

In previous work MLPs with backpropagation learning have been used to learn from agricultural data and predict yield, albeit in a non-spatial setup (see Section 3.1 for details). Regarding the question of network size, the results obtained in [Ruß et al., 2008a,c] lead to assume that the extension to more than one hidden layer only marginally increases the generalization performance of MLPs, but rather drastically increases the computation time for the backpropagation algorithm. Hence, here it is assumed that one hidden layer is sufficient to approximate the underlying function sufficiently well. Empirical evaluation shows that the size of this layer should be set such that around $\frac{1}{20}$ of the data set size (data records \times

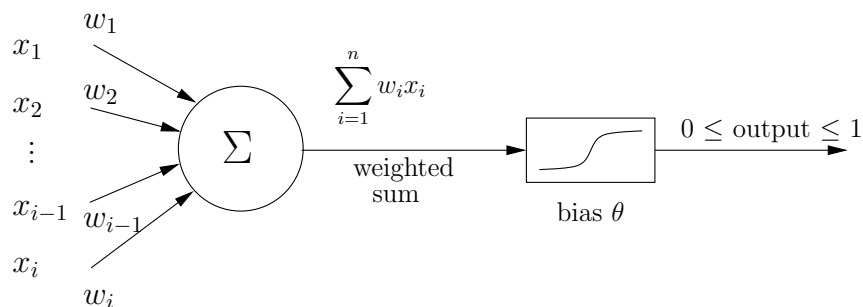


Figure 3.6: Schematic of a single perceptron unit with a weighted sum as the input function, a sigmoid function as the activation function and identity as the output function. This perceptron is depicted in Figure 3.5 as a single sigmoid for simplicity.

data record length) is reached for the number of internal network connections, which leads to around 20 hidden neurons. Further internal parameters are *tangens hyperbolicus* as the activation function and a minimum gradient of 10^{-3} , while the learning rate does not seem to have a major effect on the error rate as long as it is set conservatively.

3.3.6 Support Vector Regression

Support Vector Machines (SVMs) are a supervised learning method initially described by [Boser et al., 1992]. Although SVMs are mainly used in classification tasks, they have been used in regression setups as well, e.g. in electricity load forecasting [Chang et al., 2001], in chemistry [Ivanciuc, 2007], landslide hazard prediction [Brenning, 2005] and modeling fish-habitat relationships [Knudby et al., 2010b]. Yield prediction is a regression task, so the focus is on support vector regression (SVR) in the following.

Given the training set, the goal of SVR is to approximate a linear function $f(x) = \langle w, x \rangle + b$ with $w \in \mathbb{R}^N$ and $b \in \mathbb{R}$. This function minimizes an empirical risk function defined as

$$R_{emp} = \frac{1}{N} \sum_{i=1}^N L_{\varepsilon}(\hat{y} - f(x)), \quad (3.10)$$

where $L_{\varepsilon}(\hat{y} - f(x)) = \max(|\xi| - \varepsilon, 0)$. $|\xi|$ is the so-called slack variable, which has mainly been introduced to deal with otherwise infeasible constraints of the optimization problem (cp. [Smola and Schölkopf, 1998]). By using this variable, errors are basically ignored as long as they are smaller than a properly selected ε . The function shown here is called ε -insensitive loss function. Other kinds of functions can be used, some of which are presented in chapter 5 of [Gunn, 1998]. To estimate $f(x)$, a quadratic problem must be solved, of which the dual form, according to [Mejía-Guevara and Kuri-Morales, 2007] is as follows:

$$\text{maximize} : -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) - \varepsilon \sum_{i=1}^N (\alpha_i + \alpha_i^*) + \sum_{i=1}^N y_i (\alpha_i - \alpha_i^*) \quad (3.11)$$

with the constraint that $\sum_{j=1}^N (\alpha_j - \alpha_j^*) = 0, \alpha_j, \alpha_j^* \in [0, C]$. The regularization parameter $C > 0$ determines the tradeoff between the flatness of $f(x)$ and the allowed number of points with deviations larger than ε . As mentioned in [Gunn, 1998], the value of ε is inversely proportional to the number of support vectors. An adequate setting of C and ε is necessary for obtaining a suitable solution to the regression problem.

Furthermore, $K(x_i, x_j)$ is known as a kernel function which allows to project the original data into a higher-dimensional feature space where it is much more likely to be linearly separable. Some of the most popular kernels are radial basis functions (Equation 3.12) and a polynomial kernel (Equation 3.13):

$$K(x, x_i) = e^{-\frac{\|x-x_i\|^2}{2\sigma^2}} \quad (3.12)$$

$$K(x, x_i) = (\langle x, x_i \rangle + 1)^\rho \quad (3.13)$$

The parameters σ and ρ have to be determined appropriately for the SVM to generalize well. This is usually done experimentally. Once the solution for the above optimization problem in equation 3.11 is obtained, the support vectors can be used to construct the regression function:

$$f(x) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(x, x_i) + b \quad (3.14)$$

Further discussion can be found in [Gunn, 1998], more recently in [Frag and Mohamed, 2004], with a practical guide to support vector classification in [Hsu et al., 2008]. Research into determining SVM parameters for regression shows that this can be achieved faster by genetic algorithms but that the standard cross-validation approach yields the same results [Mejía-Guevara and Kuri-Morales, 2007].

3.3.7 Bagging / Random Forests

Bootstrap aggregating (or bagging) has first been described in [Breiman, 1994] and [Breiman, 1996]. It is generally described as a method for generating multiple versions of a predictor and using these for obtaining an aggregate predictor. In the regression case, the prediction outcomes are averaged. Multiple versions of the predictor are constructed by taking bootstrap samples of the learning set and using these as new learning sets. Bagging is typically considered useful in regression setups where small changes in the training data set can cause large perturbations in the responses. Hence, bagging is one method of a family of resampling ensemble methods.

Bagging wraps around the actual regression technique, which may be a regression tree, a neural network, support vector regression or a k-nearest-neighbor technique (kNN). Bagging has been shown to work rather well for unstable procedures, such as regression trees and neural networks. It is not expected to lead to large improvements in rather stable regression techniques such as kNN. In this thesis, bagging is employed in conjunction with a regression tree prediction technique, leading to *random forests*.

According to [Breiman, 2001], random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with

the same distribution for all trees in the forest. In the version used here, the random forest is used as a regression technique, while it also applies to classification trees. Basically, a random forest is an ensemble method that consists of many regression trees and outputs a combined result of those trees as a prediction for the target variable. Usually, the generalization error for forests converges towards a lower bound as the number of trees in the forest becomes large. The random forest algorithm is described in Algorithm 3. The parameters k and r determine the number of inner and outer validation runs and are set to the values described in [Breiman, 2001]. Depending on the available computational power, these parameters may be set to higher values. Note also that the error estimate calculation has been changed from the original mean squared error to the root mean squared error used in this thesis.

Algorithm 3 Random Forest Algorithm, adapted from [Breiman, 2001]

repeat

Divide data set randomly into learning (\mathcal{L}) and test set (\mathcal{T})

repeat

Select a bootstrap sample \mathcal{L}_b from \mathcal{L} (sampling with replacement)

Grow a regression tree from \mathcal{L}_b

Select pruned subtree ϕ_i based on test set \mathcal{L}

until $k=25$

Apply the k trees to each $\mathbf{x}_n \in \mathcal{T}$

$\hat{y}_n = \frac{1}{k} \sum_{i=1}^k \phi_i(\mathbf{x}_n)$ /*average tree results for test set examples*/

$e_B(\mathcal{L}, \mathcal{T}) = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$ /*compute test set root mean squared error*/

until $r = 100$

$\bar{e}_B = \frac{1}{r} \sum_{i=1}^r e_{Bi}$ /*compute average bootstrap error estimate*/

3.4 Comparison: Non-Spatial vs. Spatial Regression Setting

From the theoretical point of view, as laid out before, non-spatial regression modeling used with spatial data sets should lead to overfitting of the regression model and to underestimation of the prediction error. In order to validate this hypothesis, a few of the more advanced regression models were chosen and run twice on two of the spatial data sets: in a spatial and a non-spatial cross-validation setup. The k parameter determines the number of folds in both the spatial and the non-spatial cross-validation setups. The results are shown in Table 3.1 and also available from [Ruß and Kruse, 2010; Ruß and Brenning, 2010b]. The key result is that the underestimation of the modeling error does occur in practice – therefore, spatial autocorrelation has a major influence on those models. Another expected behavior is that the error levels in the spatial cross-validation setup tend to decrease with rising k – spatial cross-validation is expected to be the same as non-spatial cross-validation as soon as $k = N$, where N is the number of data records in the data set.

	k	F440		F611	
		spatial	non-spatial	spatial	non-spatial
Regression Tree	10	1.09	0.56	0.69	0.40
	20	0.99	0.56	0.68	0.42
	50	0.91	0.55	0.66	0.40
Support Vector Regression	10	1.06	0.54	0.73	0.40
	20	1.00	0.54	0.71	0.40
	50	0.91	0.53	0.67	0.38
Bagging	10	0.99	0.50	0.65	0.41
	20	0.92	0.50	0.64	0.41
	50	0.85	0.48	0.63	0.39

Table 3.1: Results of running different cross-validation setups on the data sets F440 and F611; comparison of spatial vs. non-spatial treatment of data sets; root mean squared error is shown, averaged over clusters/folds; k is either the number of k -means clusters in the spatial setup or the number of folds in the non-spatial setup. Table reproduced from [Ruß and Brenning, 2010b].

An aspect in the data which has not been considered so far are the temporal relationships between yield and possible predictors. The basic idea here is to confirm that yield prediction is really influenced by the predictors through the use of different subsets of predictors. In [Ruß and Brenning, 2010a], the full data sets are split temporally as more predictors become available throughout the season. The expected result is that, as more variables become available closer to harvest (especially vegetation indicators), the yield prediction tends to improve. This has also been found in [Ruß et al., 2008a], albeit only for neural networks as a regression model.

3.5 Spatial Variable Importance Assessment

3.5.1 Introduction to Spatial Variable Importance

The underlying question to ask in a yield prediction setup is whether a predictive variable is important or not for the regression model that aims to predict yield. One way of doing this is to apply standard feature selection approaches, with possibly large search spaces. Nevertheless, the interdependencies and possible interactions between variables in the data sets may lead the feature selection approaches such as *forward selection* [Langley, 1994] or *backward elimination* [Dash and Liu, 1997] into local optima.

While feature selection requires the computation of a regression model for each generated subset of features and is therefore computationally rather heavy, variable importance follows a different approach. The cross-validation with regression setup is left unchanged, i.e. the data are subdivided into training and test sets, using the spatial sampling procedure introduced in Section 3.2.3. The regression model is trained as usual on the training data and the prediction error is computed using the test set.

Given a trained regression model, the intuitive computational approach for assessing variable importance is based on measuring the increase in prediction error associated with permuting a predictor variable [Strobl et al., 2007]. If the prediction error on the permuted test set deviates significantly from the prediction error on the original test set, the variable whose values have been permuted is likely to bear significance for the regression model. This process is schematically depicted in Figure 3.7. This rather novel approach has been shown to be successful in [Knudby et al., 2010a] and is carried over to further regression models and different types of data sets here. For neural networks, a similar procedure called *sensitivity analysis* can be performed: single (input or hidden) neurons are removed and the effect this has on the predictive accuracy is recorded (see, e.g., [Nauck et al., 2003]). However, this can not be carried over to other types of regression models. For bagging and random forests, further information on marginal and conditional permutation importance is provided in [Strobl et al., 2008].

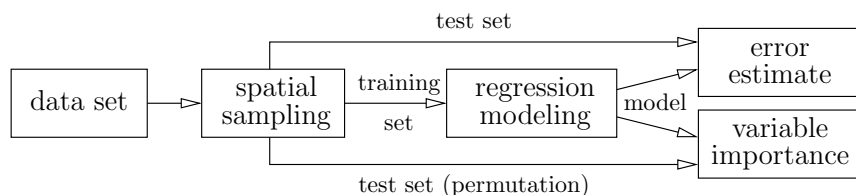


Figure 3.7: Spatial variable importance approach for regression

3.5.2 Algorithm for Spatial Variable Importance

Algorithm 4 presents the spatial variable importance approach more precisely. First, the data are spatially sampled, as described in Section 3.2.3. A regression model M is trained on the training data L and the prediction error e is computed. In nested inner loops, the predictor variables in the test set are selected one by one and each is permuted repeatedly

($k_{permutmax}$ times). After each permutation, the trained model M is run on the permuted test set and the prediction error is computed. The deviations of the respective error values from e are returned.

3.5.3 Experimental Setup for Spatial Variable Importance

The five data sets under study had different sets of available variables which were used for yield prediction. Table 3.2 shows the regression formulae which were used. The regression models described in Section 3.3 are available in R and mainly used in their standard settings. Where changes and parameter adaptations are required, these are listed in Appendix C.

Furthermore, the sites had different fertilization strategies carried out, such that for each site numerous data subsets could be generated, each containing a single strategy. The SVI approach was also applied to those subsets. In addition, the variable STRATEGY was added as a possibly confounding variable into the formulae for the full data sets.

Two different varieties of winter wheat were grown on F440, resulting in an additional SORTE variable. Since neglecting this predictive variable might confound the spatial variable importance (SVI) results, it was decided to split the F440 data set into two distinct and spatially disjoint subsets such that each of those contained one crop variety. Another option would have been to add the SORTE variable as a predictive variable, but experiments showed the SORTE variable to have a dominating effect on SVI (cp. Figure 3.8a).

For spatial cross-validation, it was decided to use a setting of $k = 20$ for the k -means cluster algorithm. Higher numbers of clusters tended to smooth the results of the spatial variable importance, but the overall behaviour converged towards the non-spatial cross-validation procedure, as shown in [Ruß and Brenning, 2010b].

From the 20 spatial clusters (contiguous sub-areas of the site), 90% (18) are sampled randomly to create the training set. The remaining two clusters represent the test set, such that $\binom{20}{2} = 190$ possible combinations of training/test clusters result. The random sampling from these clusters was repeated 100 times. For each of the training/test set combinations, each of the regression models was trained and the root mean squared error (RMSE) was computed on the test set. Additionally, for each variable in the test set, 200 permutations of this variable were created while leaving the remaining variables as-is. The trained model's prediction based on this changed test set was recorded and the prediction's deviation from the original test set's prediction was taken as a measure of spatial variable importance (SVI).

The results are provided in two ways. To enable a comparison between the different models, their RMSEs were directly compared, grouped by the different data sets. A naive prediction model was added, which, for each test set example, simply outputs the mean value of the response variable in the respective training set. More advanced regression models were expected to outperform this naive prediction, which would show as a lower RMSE value.

The actual spatial variable importance values were provided again for each data set and its subsets separately. To account for the different characteristics of the models and possibly resulting different SVI values for single variables, the figures show the different models separately.

Algorithm 4 Spatial Variable Importance

```
1: /*input: spatial data set D, parameters  $k_{means}, k_{sampling}, k_{permutmax}$  */
2: /*output: list of predictors with average error deviances*/
3:
4: repeat
5:    $j++$ 
6:   /*sample data spatially into training set L(earning) and test set E(valuation)*/
7:    $(L, E) \leftarrow \text{spatialsampling}(D, k_{means})$ 
8:    $M \leftarrow \text{train}(\text{model}, L)$ 
9:    $e \leftarrow \text{predict}(M, E)$ 
10:
11:   for all predictors  $p_i$  in  $E$  do
12:      $k_{permut} \leftarrow 0$ 
13:
14:     repeat
15:        $k_{permut}++$ 
16:        $E_p \leftarrow \text{permute}(E, p_i)$  /*permute one predictor in test set E*/
17:        $e_{p_i}[k_{permut}] \leftarrow \text{predict}(M, E_p)$  /*calculate error of M on permuted test set*/
18:
19:     until  $k_{permut} = k_{permutmax}$ 
20:     /*compute average error deviation with regard to unpermuted E*/
21:      $dev_{p_i}[j] \leftarrow \frac{1}{k_{permutmax}} \sum_{i=1}^{k_{permutmax}} |e - e_{p_i}[i]|$ 
22:
23:   end for
24: until  $j = k_{sampling}$ 
25:
26: for all predictors  $p_i$  do
27:    $avrdev_{p_i} \leftarrow \frac{1}{k_{sampling}} \sum_{m=1}^{k_{sampling}} dev_{p_i}[m]$ 
28:
29: end for
30:
31: return all  $avrdev_{p_i}$  /*return average deviations for further processing*/
```

data set	predictors	response
F440	EC25, N3, N2, N1, REIP32, REIP49, STRATEGY, DEM, (SORTE)	YIELD07
F550	EC25, N3, N2, N1, REIP32, REIP49, STRATEGY, DEM	YIELD04
F550	EC25, N3, N2, N1, REIP32, REIP49, STRATEGY, DEM, YIELD03	YIELD04
F610	EC25, N3, N2, N1, STRATEGY, DEM	YIELD07
F611	EC25, N3, N2, N1, REIP32, REIP49, STRATEGY, DEM	YIELD07
F631	EC25, N3, N2, N1, STRATEGY, DEM	YIELD07

Table 3.2: Regression formulae for the different data sets. Since F550 had the previous year’s yield YIELD03 as an additional variable, it was decided to generate experimental results with two formulae on this site. Furthermore, the DEM variable consists of the following variables generated from the digital elevation model: CATCHMENT.AREA, CATCHMENT.SLOPE, CATCHMENT.AREA.MOD, WETNESS.INDEX, SLOPE, CURVATURE, CURVATURE.PLAN, CURVATURE.PROFILE. For the full data sets, the confounding variable STRATEGY was also added.

3.6 Results for Spatial Variable Importance

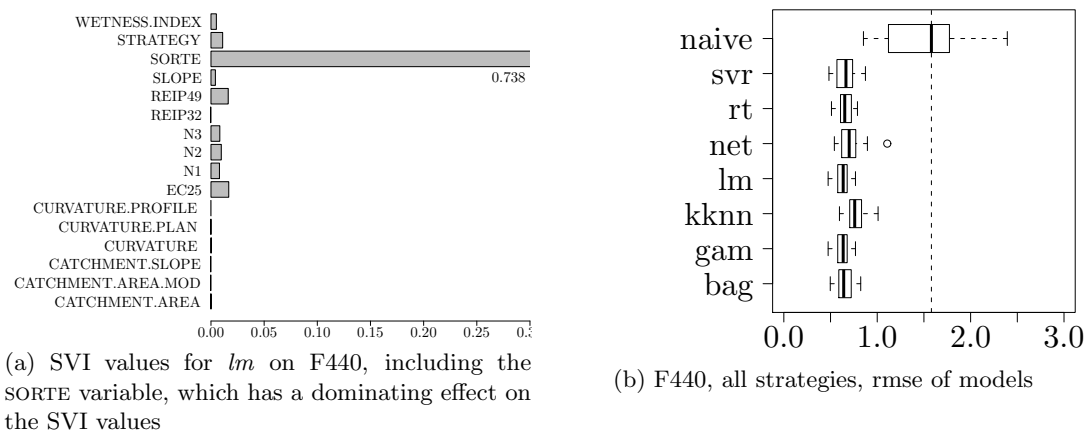
The results in this section are grouped according to the data sets from which they were generated. Each data set available was analyzed first with all strategies combined. This also required taking the predictor variable STRATEGY into account as a possible confounder. After this analysis, the single strategies were analyzed further by running the same spatial variable importance approach on the respective subsets of the data set. An overview about the most important results is given in the following Section 3.6. Detailed results are provided in the appendix part, starting on Page 193.

The results in the following section are summarized using starplots. The seven models are aligned on seven axes radiating from a common centre in a 180 degrees angle. For each combination of data (sub)set vs. predictor variable, a star is generated, showing this particular variable’s importance in the data set. The SVI is assumed as the mean of the combination’s RMSE values. The results are normalized per data set to a range of [0,1] such that SVI comparisons among the respective data set are possible.

3.6.1 Results for F440

The F440 site consisted of two spatially disjoint areas where two different varieties of crop were grown (cp. Figure A.2d on page 175). This led to an additional SORTE variable which was added as a possibly confounding variable. In this study, however, this particular variable had a large variable importance compared to the remaining variables (cp. Figure 3.8a). This result is consistent with a similar study on variable importance for corn [Miao et al., 2006], where the influence of the crop variety (variable: HYBRID) was highest, while other variables were rather unimportant in its presence.

The permutation-based variable importance approach was in principle able to deal with this and still provided meaningful SVI values for the remaining variables. However, the question arose whether the two crop varieties might be fundamentally different. This would lead to each variety’s variables having different SVI values. Therefore, applying the SVI

Figure 3.8: F440, showing the effect of the *SORTE* variable, both in SVI and RMSE values

approach to the complete F440 data set would only have provided an aggregate SVI for the variables on the complete site, without enabling any distinction between the crop varieties.

Furthermore, using the *SORTE* variable lead to a wrong overall idea of the different models' performance. As a baseline model, the naive predictor was used: it simply predicted for each test set sample the average yield in the training set. However, this naive predictor did not use any of the variables and hence provided a poor overall performance (cp. Figure 3.8b for an example of the *lm* model on this site). This partially distorted the results and conveyed the result that any of the “true” regression models is considerably better than *naive*, which is not true when simply considering the *SORTE* variable as a predictor.

Therefore it was decided to split the F440 data set into two subsets each containing only one of the two possible values for *SORTE* and extend the SVI approach to those two subsets, called F440sorte1 and F440sorte2, respectively. This was consistent with the general idea of the conditional variable importance approach for tree-based models shown in [Strobl et al., 2008]. Nevertheless, in the following the results both for the complete F440 site as well as for the two spatially disjoint subsets are provided.

For the complete F440 site, the results summary is provided in Table 3.3. The SVI results for the two subsets of F440, F440sorte1 and F440sorte2, are provided in Tables 3.4 and 3.5. As mentioned before, the summary for F440 mainly shows the dominance of the *SORTE* variable, regardless of the model and data strategy subset used. Furthermore, the regression models' performance summary, shown in Figure B.1, conveys the idea that any of the non-naive regression models perform considerably better than the naive predictor. This is mainly an effect of the *SORTE* variable. Judging the five RMSE comparisons for F440 and its strategy subsets (cp. Fig. B.1), *regtree*, *lm*, *gam* and *bagging* appear to be the regression models with the best performance. *net*, *kknn* and *svr* are less consistent in their performance. The SVI values are depicted in Figure B.2 to B.6. Since the field comprises two different varieties of crop, the SVI values should be seen as a mix of both crop varieties. Keeping this restriction in mind, the single most important variable throughout the strategies and the models is the REIP49 predictor, except for the “constant” strategy,

where the REIP32 predictor is the most important variable instead. This difference in the REIP importance remains when judging both crop varieties separately and can thus not be attributed to those. The most likely conclusion is that with the “constant” strategy sufficient amounts of N1 are applied, making latter N applications somewhat less important. On the other hand, with the “low, constant” strategy the amount of N1 seems to be insufficient, judging indirectly from the low importance of the REIP32 variable, such that the latter applications of N become more important, at least indirectly via the REIP49 variable.

Apart from the mentioned differences between the REIP SVI values in the two “constant” strategies, there is a notable SVI difference between the varieties in the “neural network” strategy where the F440sorte1 subset exhibits a clear importance for both values, while the F440sorte2 subset does not. Furthermore, in the “sensor” strategy for F440sorte2, which bases its fertilizer amounts on a REIP sensor, the SVI values for the REIP values are low, but those for the N2 and N3 variables are quite high, which is plausible. However, this result is not valid for F440sorte2, where instead a high SVI for the EC25 variable is provided.

An interesting result for the comparison between the crop varieties are the results for the WETNESS INDEX. While for F440sorte1 the SVI for WETNESS INDEX is relatively high in all but the “neural network” strategies, this holds vice versa for the F440sorte2 subfield. On the one hand, this may indicate different water uptake of the crop varieties. On the other hand, since the two varieties are planted in spatially disjoint areas, the topography of those areas is likely to be different leading to different SVI values. This latter hypothesis is also confirmed by the difference in the SVI values between both crop varieties in the remaining terrain attributes, where F440sorte2 often shows significant SVI values (e.g. CURVATURE and CATCHMENT variables in the subsets) while F440sorte1 does not.

Throughout the different models and strategies of F440, the EC25 variable has typically a positive SVI value. As pointed out in Section 2.4.4, EC25 often exhibits empirical relationships with yield, such that the SVI findings for EC25 are consistent with this observation.

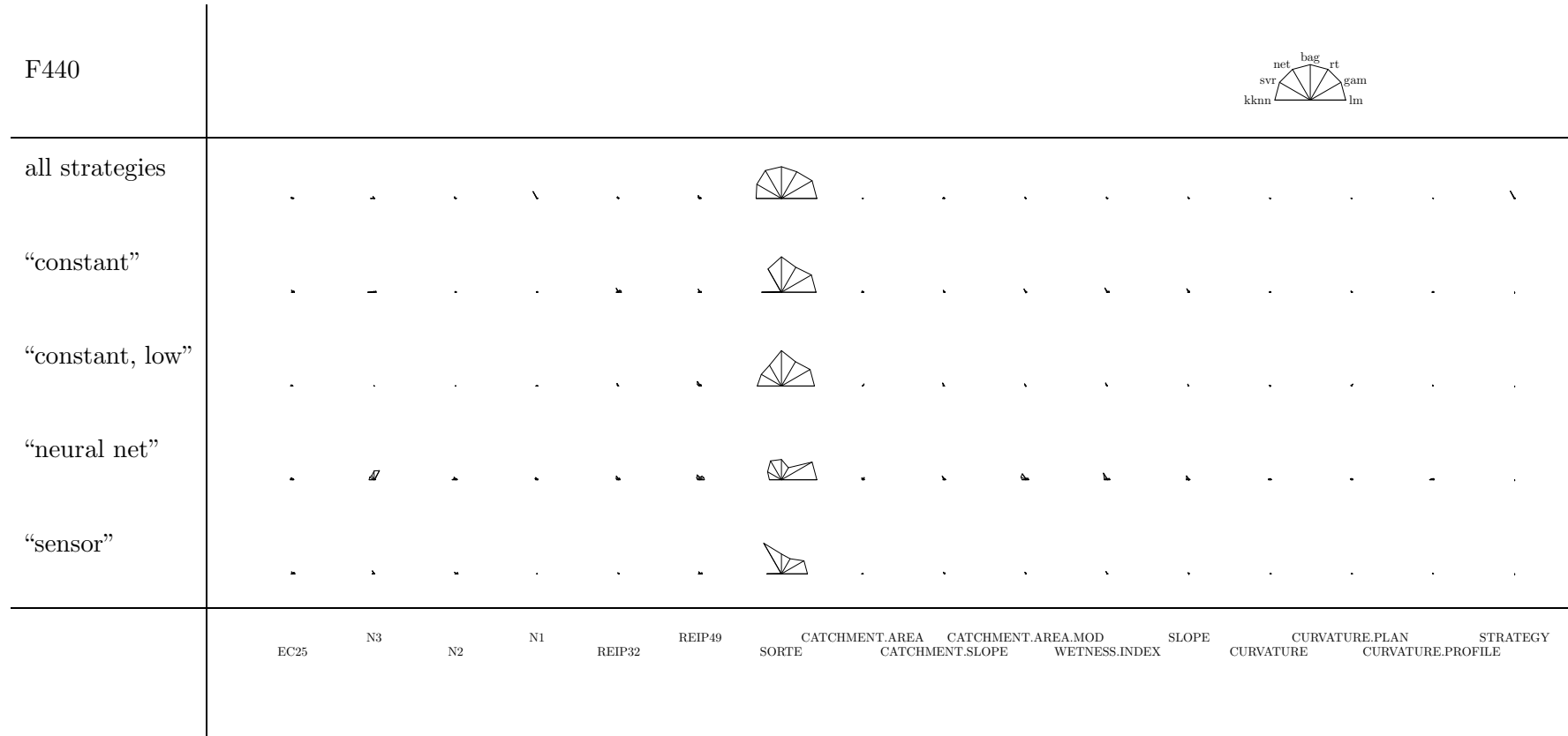


Table 3.3: F440, SVI results. SVI means of seven regression models plotted versus variables and subdivided into strategies. The SVI means were standardized to the interval [0,1] per subset (= per row).

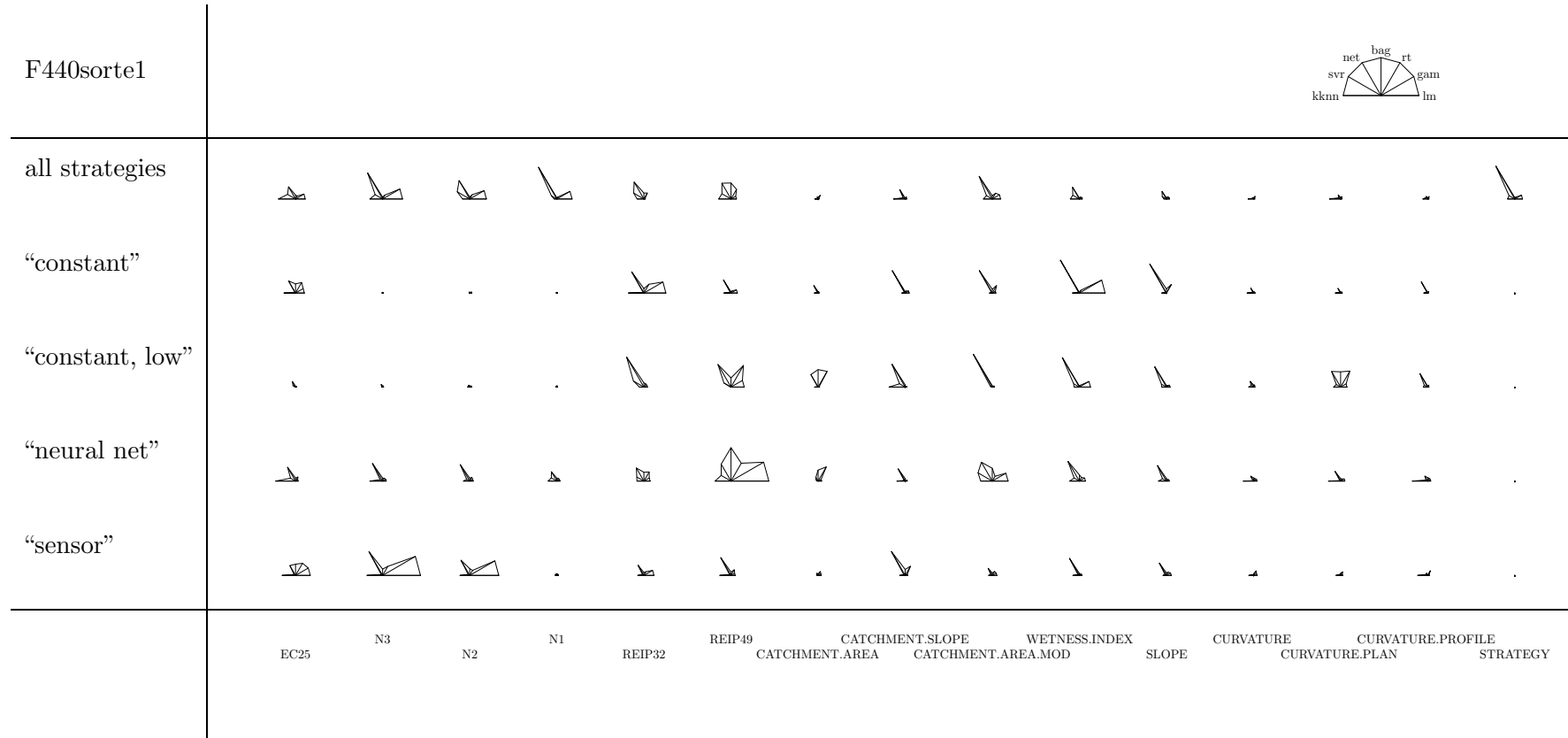


Table 3.4: F440sorte1, SVI results. SVI means of seven regression models plotted versus variables and subdivided into strategies. The SVI means were standardized to the interval [0,1] per subset (= per row).

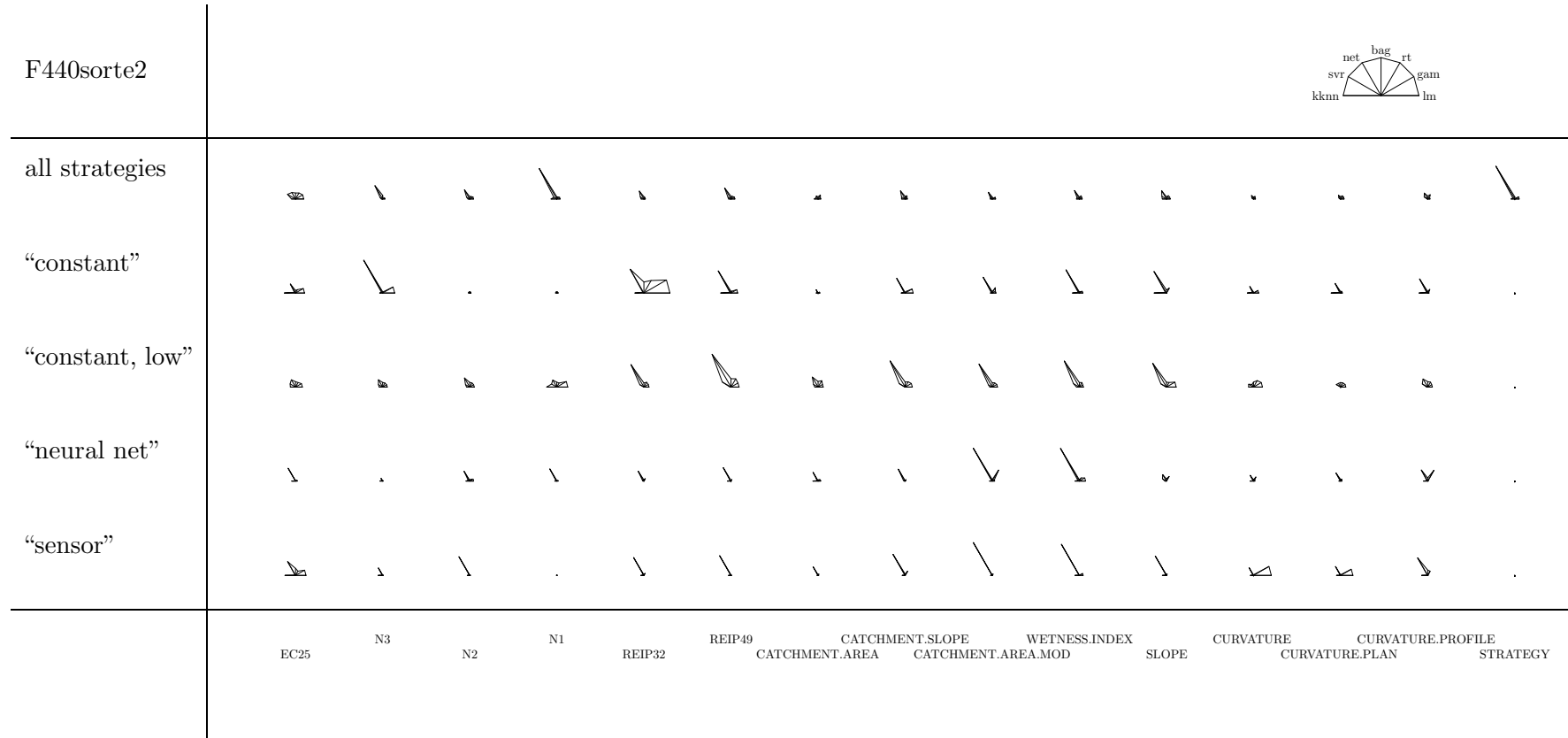


Table 3.5: F440sorte2, SVI results. SVI means of seven regression models plotted versus variables and subdivided into strategies. The SVI means were standardized to the interval [0,1] per subset (= per row).

3.6.2 Results for F550

For the F550 site, in addition to the variables for F440, the yield in 2003 was available. The relationship between previous year's and current year's yield is not as straightforward as it may be expected, e.g. it is not necessarily true that a field patch that generated a previous high yield is bound to grow a comparable high yield in the current year. Hence, to keep the results comparable to the remaining sites, it was decided to use the F550 data set in two configurations: without/with YIELD2003 as a predictive variable.

Figures B.19 and B.25 show the RMSE for the F550 site and its substrategies, without/with YIELD2003. Judging from those results, it can be seen that adding the predictor YIELD2003 only slightly improves the naive predictor's baseline, but does not have a major effect on the regression models. Most importantly, the neural network performs worst, throughout the two data sets and throughout the different strategies. The remaining regression models are typically quite close to each other, with the most consistent good performance for the *lm*, *svr* and *bagging* models.

Tables 3.6 and 3.7 provide a summary of the SVI results for the F550 site. The detailed results are available in Figures B.20 (Page 213) to B.30 (Page 223). The most obvious effect of directly comparing the SVI results is the importance of YIELD2003, which appears consistently throughout the models and strategies.

Interestingly, for the "company" and "mapping" subsets, the CATCHMENT variables, WETNESS INDEX and SLOPE have significant SVI values, with and without YIELD2003, but not in all models. This is a consistent difference between the "mapping"/"company" and "sensor" strategies. Hence, in the first two subsets, the yield's heterogeneity seems to be explained partly by terrain attributes, since those have relatively high SVI values. In comparison, in the latter subset yield heterogeneity seems to be rather explained by actually different N applications and the REIP values. Since the "company" and "mapping" approaches apply constant, predetermined N amounts and do not take REIP values into account for fertilization guidance, the SVI values confirm this difference quite clearly, however, not for all regression models. This difference may partly also be attributed to the small number of distinct values that the N variables take in the "company" and "mapping" subsets. Another plausible result, similar to the F440 site, is the SVI of the N2, N3, REIP32 and REIP49 variables in the "sensor" strategy. Furthermore, when YIELD2003 is added as a predictor, it is clearly least important in the "sensor" subset when comparing the four subsets. This confirms the basic idea of the "sensor" approach, which bases its fertilization on REIP values. Furthermore, in the "sensor" strategy, the REIP and N2/N3 variables show most strongly in *lm*, which may lead to the assumption that a linear relationship between the REIP and N values has been implemented in the "sensor" strategy.

In the presence of YIELD2003, the "N-trial" subset shows significant SVI values for the N variables, which should be expected. However, in the absence of YIELD2003, terrain attributes such as WETNESS INDEX, CATCHMENT variables and even CURVATURE have relatively high SVI values. This may hint to a quite strong relationship between N1 and the previous year's yield YIELD2003, at least in this regression setting.

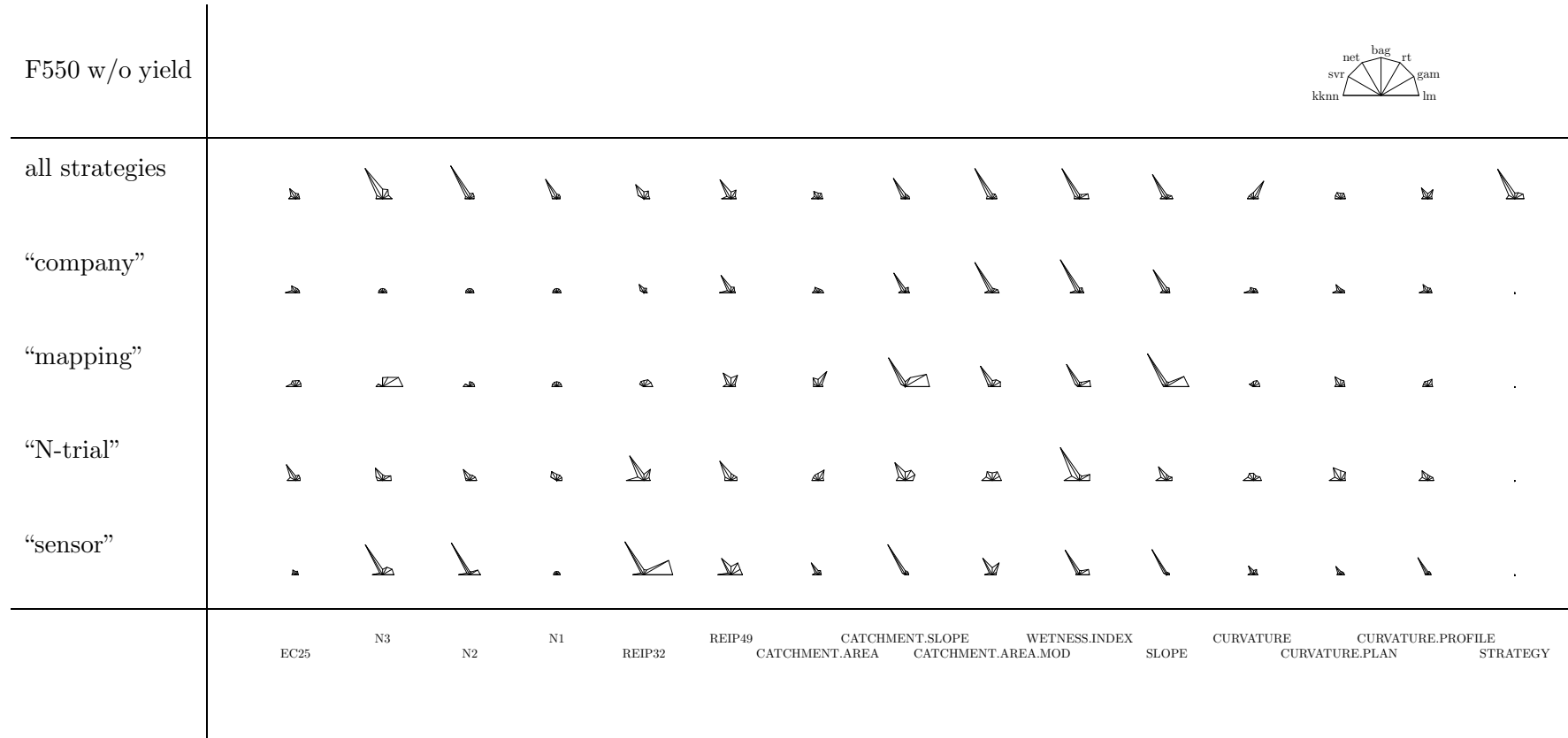


Table 3.6: F550 without YIELD2003, SVI results. SVI means of seven regression models plotted versus variables and subdivided into strategies. The SVI means were standardized to the interval [0,1] per subset (= per row).

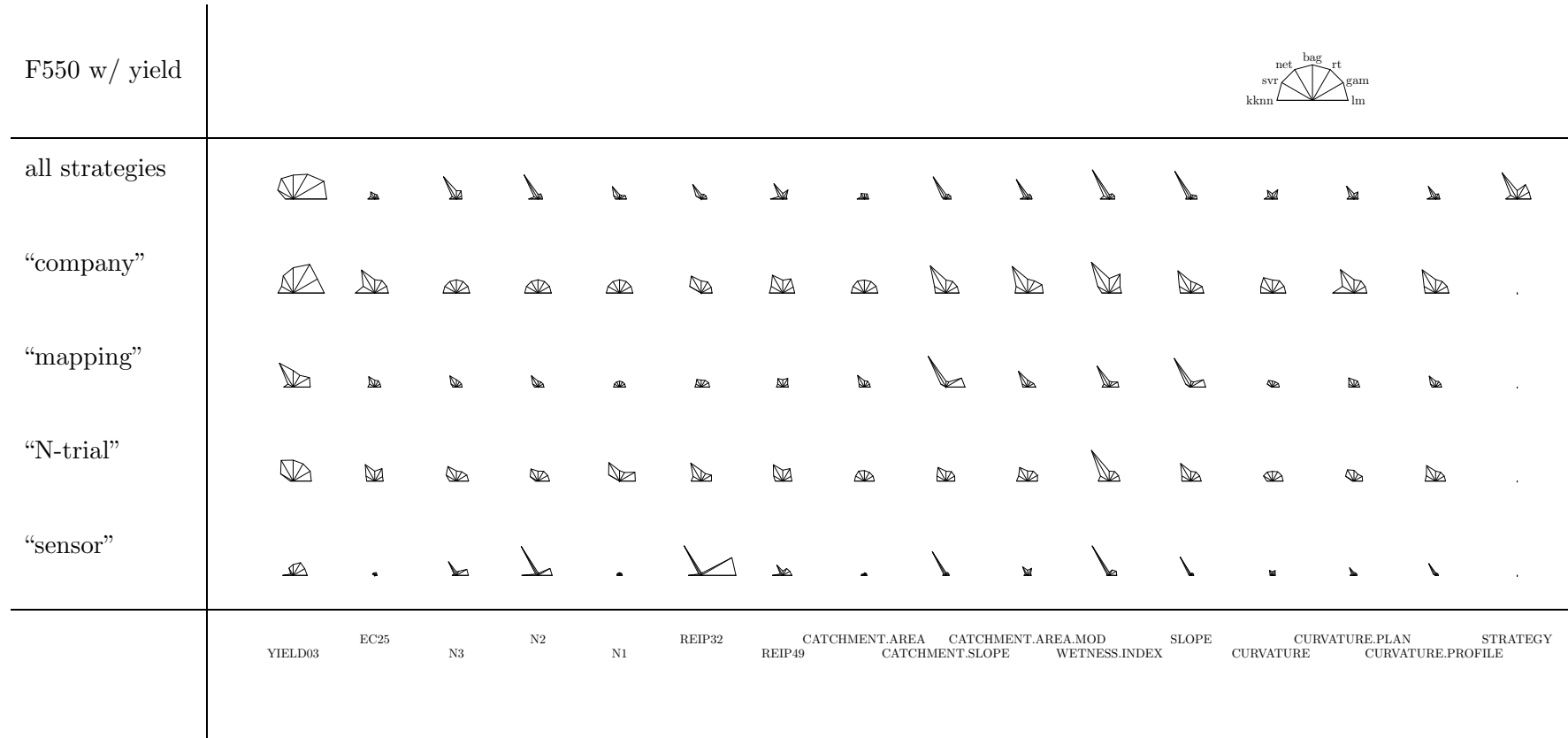


Table 3.7: F550 with YIELD2003, SVI results. SVI means of seven regression models plotted versus variables and subdivided into strategies. The SVI means were standardized to the interval [0,1] per subset (= per row).

3.6.3 Results for F610

The RMSE comparison for F610, presented in Figure B.31, shows that *svr* is one of the best models for this site (in the full data set), closely followed by *lm* and *gam*, all at values around 0.8. However, this is only slightly better than the *naive* predictor at around 0.85. The worst performing models are *net*, *regtree* and *kknn* at RMSE levels around 1.0. *net* exhibits especially low performance in the F610 subsets, such that the SVI values for this model should be taken with care.

The SVI results for F610 are shown in Table 3.8 and in detail in Figures B.32 to B.37. Throughout the different strategies and models, EC25 is the single variable which most consistently exhibits a high SVI value. Apart from the confounding variable STRATEGY, the three fertilizer applications have relatively high SVI values, but not in all models.

For the “constant” strategy subset and *lm/gam*, the CURVATURE, WETNESS INDEX and SLOPE variables exhibit a relatively high SVI, and among *regtree/bagging*, the CATCHMENT variables show high SVI values. This may be attributed to the effect that in the absence of site-specific fertilization in the “constant” strategy, the terrain attributes have a high influence on YIELD. This hypothesis is confirmed by the high SVI for EC25, but also disproved by the high SVI for N3 in *lm/gam*.

Different results were obtained for the “neural network” subset (Figure B.34). In addition to a slight SVI value for WETNESS INDEX and CURVATURE in *lm/gam*, there is a notable SVI for PLAN CURVATURE in *regtree/bagging*. For the “N-trial” subset, the only variables to seemingly have an effect in the SVI approach are N1/N2, followed by small SVI values for SLOPE in *lm/gam*, as well as PLAN CURVATURE in *regtree/bagging*.

The “sensor 1” and “sensor 2” strategies differ strongly in their SVI results, but have in common that EC25 is often present with high SVI values. While the “sensor 1” subset has a strong emphasis on the terrain attributes in the linear models, this effect is much less pronounced in the “sensor 2” subset. Furthermore, in the latter subset the N2 variable has often higher SVI values than N1/N2, while this effect is reversed in the “sensor 1” subset.

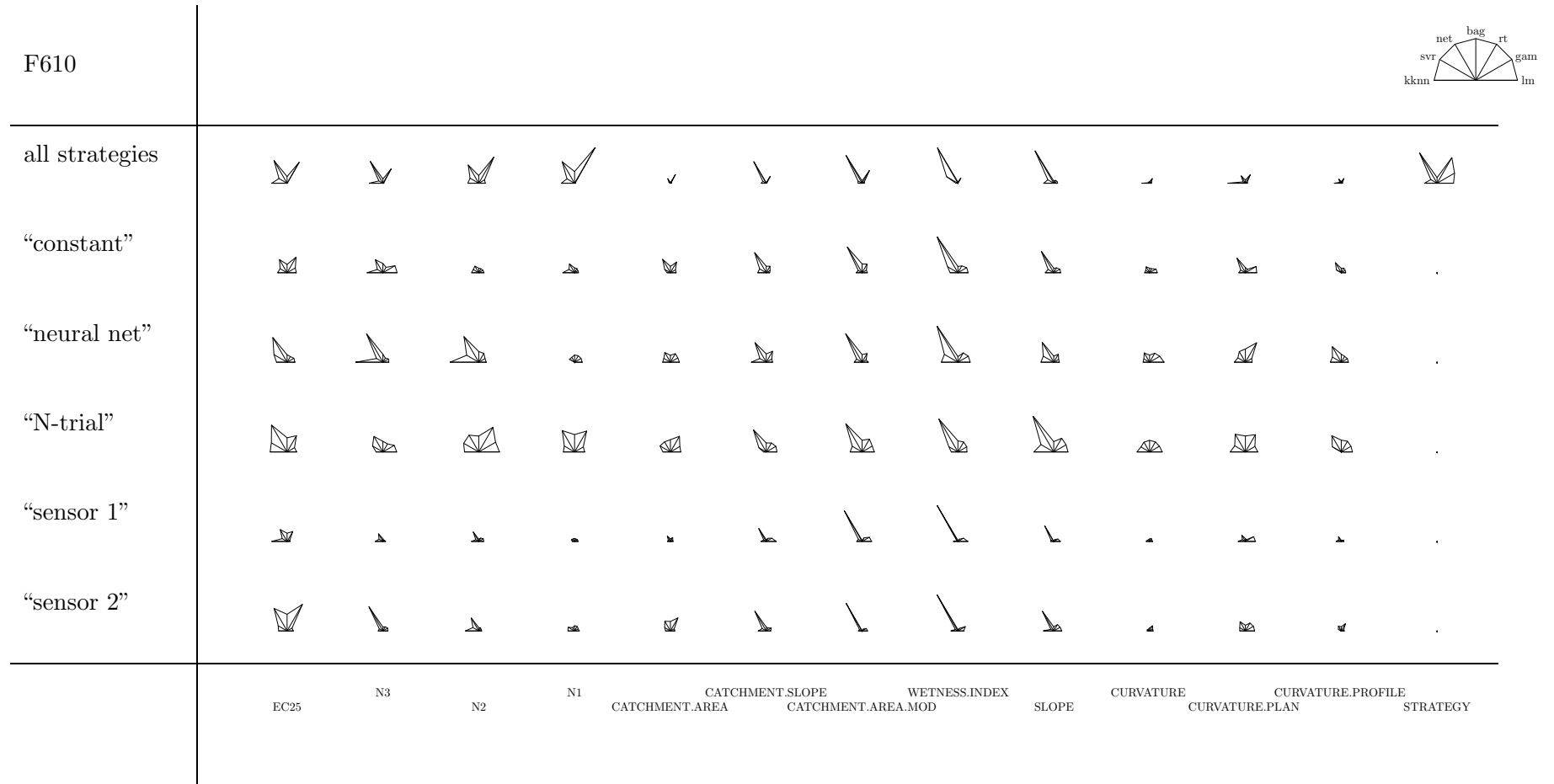


Table 3.8: F610, SVI results. SVI means of seven regression models plotted versus variables and subdivided into strategies. The SVI means were standardized to the interval $[0,1]$ per subset (= per row).

3.6.4 Results for F611

The RMSE comparison for F611 in Figure B.38 shows that any regression model performs better than *naive* on this data set. Performance-wise the linear models, *svr* and *bagging* hover around the same value of 0.65, followed by *regtree* and *net* and *kknn*, in a range from 0.7 to 0.75. The SVI results summary is provided in Table 3.9, while the details are shown in Figures B.39 to B.42.

The most straightforward result in Table 3.9 is the importance of REIP49 as well as WETNESS INDEX and SLOPE throughout the strategies and often among a wide range of different models. For the whole site, the SVI summary shows an outstanding influence of REIP49 in every regression model. This is followed by the SVI values for SLOPE. Since *net* shows a comparative performance, its results are also similar, with SVI values for SLOPE, REIP49 and WETNESS INDEX showing pronouncedly, albeit also showing the confounding variable STRATEGY.

For the “constant” strategy, any regression model is better than *naive*, also showing *lm/gam* with the best performance, followed by *bagging*. The neural network performs worst. In terms of SVI, shown in Figure B.40, *lm/gam* emphasize the REIP49 variable as well as the WETNESS INDEX and SLOPE. The fertilizer variables do not have a high SVI value, which is on the one hand due to the fertilizer not being applied site-specifically in this strategy. On the other hand, this leads to an insufficient number of different levels to be recognized by permutation-based SVI. The REIP49 variable is also recognized by the tree-based models as being the most important, while *svr* shows a completely different picture by exposing MODIFIED CATCHMENT AREA and CATCHMENT AREA as the most important variables. *net* agrees with the linear models on SLOPE, WETNESS INDEX and REIP49, but also shows MODIFIED CATCHMENT AREA and CATCHMENT SLOPE. This may hint to both linear and non-linear relationships in this data subset and an influence of a few DEM variables.

The “neural network” strategy subset of F611 (RMSE shown in Figure B.38c) still shows *lm/gam* to perform best, followed by *svr* and *bagging*. *net* performs worst, even worse than *naive*. In Figure B.41, the linear models expose SLOPE and REIP49 to be the most important variables, which is also discovered by the tree-based models, although at a lower level. *net* also agrees on SLOPE, but also shows WETNESS INDEX to be of importance. *svr*, if any, shows SLOPE and REIP49, but at a very low SVI.

The “sensor” strategy on F611 (cp. Fig. B.38d) shows about the same ranking as the other subsets on this site, with *net* performing worst and the other models (except *regtree*) being better than *naive*. *lm/gam* perform best, closely followed by *svr* and *bagging*. The SVI values show the REIP and N2/N3 variables in different models, as well as the WETNESS INDEX and SLOPE. This is different from, e.g., the F440 site, where the “sensor” strategy typically minimized the SVI of terrain attributes while showing the highest SVI values for the REIP and N2/N3 variables.

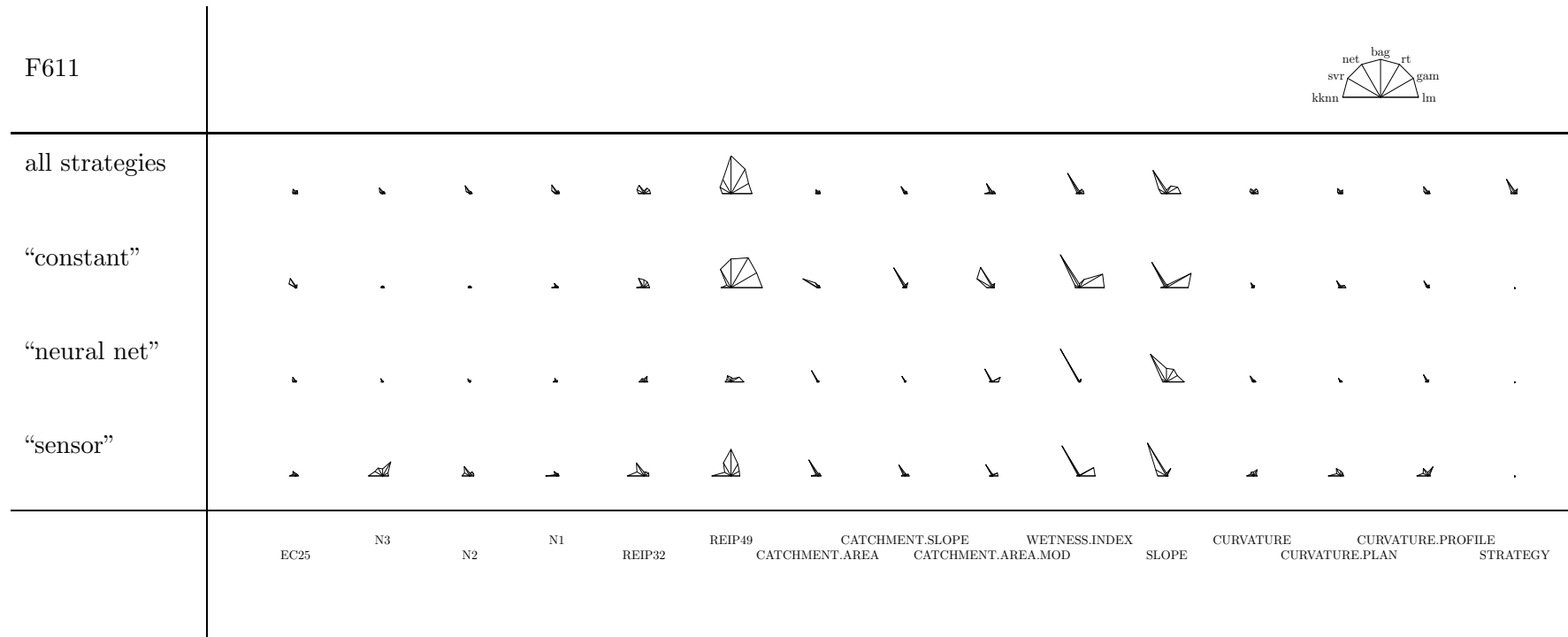


Table 3.9: F611, SVI results. SVI means of seven regression models plotted versus variables and subdivided into strategies. The SVI means were standardized to the interval [0,1] per subset (= per row).

3.6.5 Results for F631

The RMSE comparison for F631 and its subsets is shown in Figure B.43, while the SVI summary is provided in Table 3.10, with the SVI details in Figures B.44 to B.47. In contrast to the previously described data sets (except F610), the F631 site does not have small-scale REIP variables, which may be a reason for the overall worse yield prediction performance than for the comparable F611 site. The SVI summary mainly shows the importance of the EC25 and CURVATURE variables with different strategies and models, while also exhibiting WETNESS INDEX and SLOPE to be of importance. As with the previous data sets, the differences in the N variables' importance between the different strategies can be seen.

For the complete data set (cp. Fig. B.43a), *svr* and *lm/gam* perform best at around 1.1, followed by *bagging* at 1.2, which is still better than *naive*, *net* and *regtree*, which are all around 1.3. For the SVI values depicted in Figure B.44, the most influential variables in the linear and the tree-based models are CURVATURE and EC25. The N2 variable also ranks favorably for *lm* and *gam*, comparable to SLOPE and WETNESS INDEX. The latter variable is the most important for *net*, being followed by CATCHMENT AREA, SLOPE and N2, albeit confounded by STRATEGY.

For the “constant” strategy subset, the overall variance in the RMSE values is considerably higher than for the full data set. Only *lm*, *gam*, *bagging* and *svr* perform better than *naive*, while the overall RMSE level is typically higher than with the full data set. In terms of SVI, the linear and the tree-based models agree on CURVATURE as an important variable. *lm* and *gam* also show SLOPE and WETNESS INDEX to be of importance, while the tree-based models expose EC25 as the second most important variable. *svr* shows MODIFIED CATCHMENT AREA and CATCHMENT AREA to be of importance.

The “neural network” fertilization strategy (RMSE depicted in Fig. B.43c) shows rather mixed results. While *net* performs worst, *lm/gam* and *svr* again perform best, significantly better than *naive*. The SVI results (cp. Figure B.46) seem to show an effect for PLAN CURVATURE for the linear and tree-based models, as well as N2 and EC25. Although *net* shows a large importance value of WETNESS INDEX and a few other variables, due to its overall dissatisfactory performance those results should be considered with care. *svr* shows most of the variables as having some importance, but there are no truly outstanding variables.

The “N-trial” strategy subset (RMSE shown in Figure B.43d) shows a result similar to the aforementioned subsets. The worst model is again *net*, while *lm*, *gam*, *svr* and *bagging* perform best, the latter four significantly better than *naive*. In the SVI computations in Figure B.47, the linear models show an importance for CURVATURE and WETNESS INDEX, with no further variables reported as important ones. *regtree* shows most of the variables to have some importance, but the more stable tree-based model *bagging* only shows CURVATURE to have a low importance. Again, due to the dissatisfactory performance of *net*, its SVI results should be considered with care.

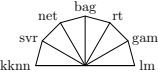
























































F631															
all strategies															
“constant”															
“neural net”															
“N-trial”															
	EC25	N3	N2	N1	CATCHMENT.SLOPE CATCHMENT.AREA	CATCHMENT.SLOPE CATCHMENT.AREA.MOD	WETNESS.INDEX	SLOPE	CURVATURE	CURVATURE.PLAN	CURVATURE.PROFILE	STRATEGY			

Table 3.10: F631, SVI results. SVI means of seven regression models plotted versus variables and subdivided into strategies. The SVI means were standardized to the interval $[0,1]$ per subset (= per row).

3.7 Discussion

The spatial variable importance approach developed and evaluated in this chapter consists of a few steps to account for the spatial nature of the data sets. The first step was to perform a spatial cross-validation instead of using non-spatial methods to perform a spatial prediction. In this study, a spatial clustering of the data sets under study was employed. While this is not the only approach to use with spatial data sets, it has the clear advantage to allow for the usage of arbitrary regression models, as long as those can be wrapped around in a cross-validation approach. For most of the regression models used in this study, special developments exist that make them fit for spatial data sets, e.g. geographically weighted regression [Fotheringham et al., 2002] for linear models and a case study on classification and regression trees for environmental and ecological data by [Bel et al., 2009]. Neural networks and support vector machines have been specially prepared for spatial environmental data by [Kanevski et al., 2009] and applied to remote sensing data by [Mountrakis et al., 2011] as well as in a comparison study by [Brenning et al., 2006]. Random forests have, e.g., been applied to predict species distribution in a marine wildlife environment by [Oppel and Huettmann, 2009] and in a comparative study by [Brenning, 2009] integrating terrain analysis and multispectral remote sensing data. Even though each of these approaches succeeds in providing a special version of the respective regression technique, those can hardly be compared against each other without implicitly favoring one over the other. Cross-validation has also been modified to fit other types of dependent data, such as in [Bühlmann, 2001] for bootstrapping models and time series data and in [Zhu and Morgan, 2004] for spatio-temporal data. Related work for clinical data from paired organs by [Brenning and Lausen, 2008] also uses bootstrap aggregating to achieve misclassification error rates by taking autocorrelation into account. Follow-up work compares different ensemble classification models on paired data [Adler et al., 2011] and concludes that resampling significantly reduces the misclassification rate.

An argument similar to one above for spatial regression and yield prediction holds for the spatial variable importance assessment. There are numerous approaches dealing with environmental data in regression settings and trying to assess variable importance. For neural networks, most of the variable importance literature points towards sensitivity analysis, which essentially removes connections (nodes or edges) from the network and assesses the change in predictive capability (cp. [Olden et al., 2004; Weigert, 2006]). In recent years, the usage of tree-based models has increased rapidly. For simple classification and regression trees, variable importance results have been described, e.g. in [Sandri and Zuccolotto, 2010] and [Clemencon et al., 2011], typically basing the variable importance on some measure of node impurity. Variable importance in random forests and bagging have seen wide usage, e.g. in a theoretical introduction by [Strobl et al., 2009] and with empirical results in [Archer and Kimes, 2008] and [Auret and Aldrich, 2011]. Permutation-based variable importance approaches have been described, e.g. by [Lemaire and Clairot, 2006] and [Nicodemus et al., 2010]. However, the above approaches typically handle only specific regression techniques rather than employing a wider framework which could incorporate a multitude of techniques. Therefore, the approach presented in this study can be regarded as a generic framework to assess variable importance for spatially autocorrelated data sets in a regression setting, regardless of having to use a specific regression technique.

rank	F440	F550	F610	F611	F631
1.	<i>svr, bagging</i>	<i>bagging</i>	<i>svr</i>	<i>lm, gam</i>	<i>svr</i>
2.	<i>lm, gam</i>	<i>naive</i>	<i>lm, gam</i>	<i>svr, bagging</i>	<i>lm, gam</i>
3.	<i>net</i>	<i>svr</i>	<i>naive</i>	<i>regtree</i>	<i>bagging</i>
4.	<i>regtree, kknn</i>	<i>lm, gam</i>	<i>bagging</i>	<i>net</i>	<i>regtree, net, naive</i>
5.	<i>naive</i>	<i>regtree, kknn</i>	<i>regtree</i>	<i>kknn</i>	<i>kknn</i>
6.	–	<i>net</i>	<i>net, kknn</i>	<i>naive</i>	–

Table 3.11: Model rankings for the spatial cross-validation approach; the best models have the lowest RMSE and the highest ranking. Models that exhibit similar RMSE values have been combined in the table.

3.7.1 Yield Prediction and Model Comparison

The first question of this chapter is which of a variety of yield prediction models should be chosen. Section 3.6 show a rather mixed picture in that regard. Choosing an appropriate regression model based purely on the RMSE values may be appropriate in a practical environment when the only interest lies in predicting yield most precisely. Table 3.11 presents a model ranking for the five data sets. Overall, *svr* ranks in the top three models consistently. This is followed by the linear models *lm* and *gam* and also by *bagging*, although their rankings differ. *regtree* and *net* consistently rank in among the last places, while *kknn* is clearly the worst model, probably since it does not generalize at all.

The models have not been fully optimized via a grid search in the parameter space or similar procedures developed for this purpose [Jiménez et al., 2007; Frohlich and Zell, 2005]. Therefore, the results may differ when particular optimizations for single models are applied. Results for a classification setting in [Brenning, 2009] have shown an internal cross-validation for parameter tuning of a support vector machine to yield slightly worse results than applying the technique in its default settings. In the current study, especially the dissatisfactory overall performance of *net* may be attributed to the huge parameter search space for the neural network which must be typically fine-tuned to a particular task. Even if *net* were to be fully optimized, it would typically require another part of the training data as a validation subset, further reducing the size of the actual training data which often impairs its predictive performance. Nevertheless, *net* would remain rather fragile to perturbations in the test data, as shown by the results of the SVI approach where the SVI values often fluctuate strongly.

For practical purposes, the previously discussed yield prediction setting allows for an objective model comparison. Using the models' performance as a proxy, it can serve as guideline for selecting a model for a particular spatial (yield) prediction task. In a comparable classification setting, the usage of tree-based models instead of linear models could be supported by this model comparison [Knudby et al., 2010a]. In the results obtained on the data sets in this study, it is suggested to use *svr* and/or *bagging* (cp. Table 3.11).

3.7.2 Spatial Variable Importance

In general, the permutation-based SVI approach succeeded in identifying important variables in the yield prediction setting presented in this chapter. This section serves as a discussion of two groups of results. The first group of results describes those which could be expected from the extensive knowledge about the PA data sets such as fertilization strategies, N applications and sensor usage. The second group is concerned with the results that are potentially novel and useful.

From an agriculture point of view, it certainly would be interesting to clearly determine the effect that N fertilization has on the final yield. Aside from the site's circumstances (such as precipitation, solar irradiation and other natural influences), this can partly be answered by the SVI approach. However, since it is based on measuring a variable's importance by permuting it in the test set and measuring the increase in RMSE of the associated (trained) regression model caused by this permutation (cp. [Strobl et al., 2007]), a low number of distinct values for one variable interferes with this approach. Unfortunately, in this study the N variables typically exhibited only a few distinct levels, unless special fertilization strategies were carried out. However, in the absence of site-specific fertilization, the importance of terrain attributes and EC_a could be established.

In particular, the SVI approach showed expected results on subsets of the data sets where specific fertilization strategies were carried out. The "sensor" strategies which employed a vegetation sensor that makes use of the REIP values used these sensor data for their fertilization recommendations. Therefore it was expected that the REIP values have a high SVI value. This was clearly shown in the results for the "sensor" strategies on F440 (Fig. B.6), F550 (Fig. B.24) and F611 (Fig. B.42), albeit not always in a linear model, but also in *bagging*. The importance of vegetation indicators is likely to be transferable to other vegetation indicators, of which a few are given in Section 2.4. This result is consistent with the study of [Pettersson et al., 2006], where it was found that different vegetation indicators and apparent electrical conductivity values explained much of the yield variation for a particular grain site. From a data mining point of view, with those results the SVI approach also allows for analyzing particular fertilization strategies.

For the F550 site, the SVI approach was carried out with and without YIELD2003 as a predictor. While the overall RMSE values for these two data sets did not differ significantly, YIELD2003 had an influence on the SVI values. With YIELD2003 included as a predictor, no further variables showed an outstanding SVI value in any of the models. This led to the assumption that the previous year's yield may be an important variable hinting towards current year's yield. Without YIELD2003, variables like WETNESS INDEX, REIP and N3 seem to be of importance in both the linear and the tree-based models. Yield-based management zoning approaches such as, e.g., [Jaynes et al., 2005; King et al., 2005] or [Brock et al., 2005] (cp. Chapter 4), are based on the assumption that the spatial relationships in yield for different site-years are related in a certain pattern and thus base their management zoning decisions on yield data.

The EC_{25} variable is an indirect and integrated measurement of a number of soil properties showing indirectly in the apparent electrical conductivity. It is expected to have an influence on yield and thus shows consistently throughout the data sets as having a high

spatial variable importance. This is consistent with existing articles on different crops, e.g. [Sears et al., 2005; Pettersson et al., 2006; Tremblay et al., 2010].

As outlined in Section 3.6 for the F440 site, the *SORTE* variable had an outstanding influence on yield. A similar effect has been described in a study by [Miao et al., 2006] for corn yield prediction using artificial neural networks, where the hybrid planted determined yield to a great extent, but where yield was also related with cation exchange capacity and relative elevation.

Potentially Novel Results

The topography attributes derived from the DEM present a rather mixed picture. There is no single of those variables to be of importance throughout the models. One of the more general results is that the terrain attributes tend to be important once fertilizer is not applied in a site-specific, but in a uniform way, i.e. in the “constant”, “low, constant” or “mapping” strategies. An early correlation-based study on the importance of topographic and soil attributes given by [Kravchenko and Bullock, 2000] concluded that, apart from soil organic matter, topographic attributes explained around 40 per cent of yield variation. Terrain attributes such as curvature and slope specifically explained yield variation for extreme topographical locations such as undrained depressions or eroded hilltops.

A number of studies related yield to elevation: [Pena-Barragan et al., 2010] (sunflowers), [Miao et al., 2006] (corn), [Tremblay et al., 2010] (corn). However, these studies employed additional variables such as weed infestation, cation exchange capacity or N sufficiency indices, which were unavailable in this study.

SLOPE is one of the important variables in the “constant” strategy subsets of F611 and F631. In [Reuter et al., 2005], similar findings were reported for shoulder and foot-slope positions of German rye and barley fields, where yield differed significantly between years and the slope largely helped in explaining the variation. Furthermore, a study by [Tremblay et al., 2010] showed *SLOPE* in conjunction with *EC25* and *ELEVATION* to be successful in deriving economic management rules for corn yield, similar as in the study of [Kravchenko and Bullock, 2000] for corn/soybean yield.

The *WETNESS INDEX* seems to play an important role in the “constant” subsets, such as those for F440, F550, F611 and F631, as well as the “mapping” subsets of F550. The prevalence of this variable in the “constant” subsets may be caused by the absence of variable fertilization, such that yield is then mainly influenced by the *WETNESS INDEX*. A similar suggestion holds true for the *CATCHMENT* variables, which exhibit a certain *SVI* in the “constant” and “constant, low” strategies of F440 and in the “mapping” strategy of F550. Related studies on the significance of hydrologic terrain attributes concluded that those attributes explain a large percentage of yield variation (cp. [Kumhálová et al., 2011] (cereals), [Iqbal et al., 2005] (cotton)). Lower importance results were obtained for potatoes in [Persson et al., 2005] for drainage area, gradient and elevation variables.

CURVATURE only seems to be of importance for the F631 site. The remaining digital elevation model variables are not exposed as having an outstanding spatial variable importance. In a study by [Kaspar et al., 2003] that included curvature as a variable to quantify the spatial variation of yield, negative correlation between relative elevation, slope and curvature was determined during less than normal growing season precipitation, while

this effect was reversed in years with greater than normal precipitation. The particular influence of precipitation on variable importance has also been determined by [Sears et al., 2005], where yield variation could be explained by EC_a and terrain attributes only for a particularly wet year.

3.8 Summary and Conclusion

This chapter revolved around the particular problem of identifying potentially important variables for yield prediction. Based on the assumption that yield prediction is essentially a regression problem, a setting was devised which is able to handle the kinds of spatial data sets encountered in this study. It basically consisted of changing the standard cross-validation approach into a spatial cross-validation approach by introducing a sampling based on spatial clustering. As detailed in [Ruß and Brenning, 2010b], neglecting to account for the spatial nature of the data sets leads to a systematic underestimation of the regression models' prediction error. Based on this spatial cross-validation approach, a permutation-based spatial variable importance approach was devised [Ruß and Brenning, 2010a]. It consists of measuring the increase in a trained regression model's error associated with the (repeated) permutation of a single predictor variable in the test set. Using the five PA data sets presented in Chapter 2, the main question posed at the beginning of this chapter could be answered. It could be shown that the developed SVI approach delivered comprehensive results, i.e. it clearly identified variables as important that were known from expert knowledge to be important. Furthermore, additional variables, especially terrain attributes, could be identified to be potentially useful with regard to yield prediction and were put into context using applicable literature results. This emphasized the importance of using digital elevation model data in precision agriculture.

As a desired side effect, a comprehensive study on the predictive performance of the different regression models was performed. This resulted in the recommendation of certain regression models for similar yield prediction tasks. Numerous suitable models can be used, depending on what type of relationship between variables is expected and should thus be modeled. Choosing a linear model and support vector regression as well as bagging seems to be a reasonable choice from a data mining point of view. The neural network approach previously used in [Weigert, 2006] yielded inferior results compared to the other regression models.

Henceforth, the developed SVI approach can readily be used in a practical setup, e.g. to assess the usefulness of novel sensor data or to analyze commercial and proprietary fertilization strategies which would otherwise be black-box systems.

Chapter 4

Exploratory Spatial Clustering for Management Zone Delineation

4.1 Introduction

Similar to the yield prediction task encountered in Chapter 3, *management zone delineation* is a classical task in agriculture. It usually revolves around a question such as *how the site should be split up into homogeneous subfields or zones for a particular purpose*. Without any particular small-scale data at hand, the answer to this question has often been based purely on experts' long-term experience. However, with the advent of high-resolution sensors and the resulting data sets such as those presented in Chapter 2, this task is likely to be handled more precisely and less labor-intensively by data mining techniques tailored to the particular purpose. From a data mining point of view, the task is likely to be handled best as a spatial clustering problem. The problem and a possible result on one-dimensional spatial data sets in this thesis are illustrated in Figure 4.1.

This chapter provides insights into management zone delineation (MZD) in precision agriculture (PA) and the associated computational tasks. Having formalized the management zone delineation as a spatial clustering problem in Section 4.2, this chapter proceeds with a detailed review of existing work in this area in Section 4.3, introducing the current state-of-the-art. Based on the shortcomings and recommendations in existing work in PA, requirements for an improved novel delineation approach are presented in Section 4.4. Based on these requirements, existing work in spatial clustering and key algorithms are reviewed to outline the major shortcomings when applying them to the kind of data sets encountered here (Section 4.5). The novel approach itself is laid out in detail in Section 4.6. Results of this approach and the effects of different parameter settings are given in Section 4.7, before the chapter is concluded.

The main ideas of the algorithm HACC-SPATIAL to be developed during the course of this chapter were published in [Ruß et al., 2010a; Ruß and Schneider, 2010]. Selected results were published in [Ruß and Kruse, 2011b], along with a further algorithmic formalization. A shortened version of this chapter featuring preliminary results was published in [Ruß and Kruse, 2011a].

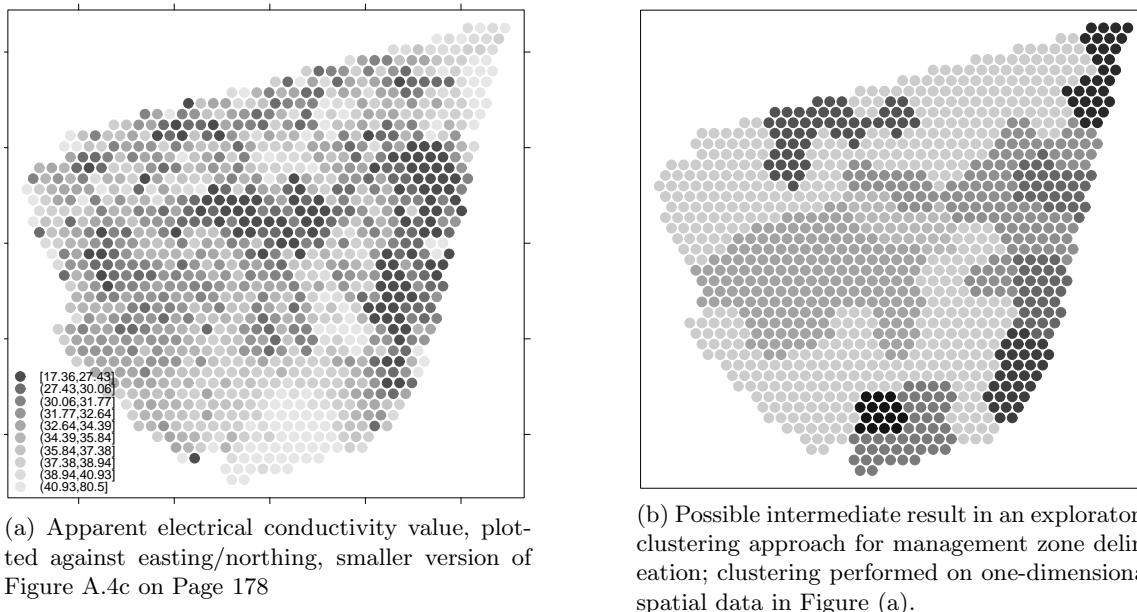


Figure 4.1: Exploratory spatial clustering for management zone delineation, performed on one-dimensional spatial data (EC25 on F550). Left: input data, right: possible clustering result

4.2 Management Zone Delineation as an Exploratory Spatial Clustering Problem

A task commonly occurring in agriculture is the so-called *management zone delineation* (MZD). There are a number of cases where this task is carried out. One example is basic fertilization: based on the biologically valid assumption that certain soil minerals are necessary for healthy plant growth, these minerals must be made available to the plants. These minerals often exist in sufficient quantities in the soil, but are not in a chemical state which allows the plants to easily tap into the mineral reservoirs. Furthermore, they may not be available at all. Therefore, basic fertilization is applied, which aims to make the minerals available. However, since the fields are usually heterogeneous, different parts of the field may require different amounts of basic fertilization. Zones can also be desired for delineating yield potential or vegetation zones. Determining these so-called management zones is therefore an important task. However, the first step before determining zones is to try to understand the data sets which is mostly an exploratory task.

From the data mining and algorithmic perspective, the task is as follows: given a set of data points (or data vectors) which are spatially sampled from some data collection of an agricultural field, turn these spatial data points into clusters (management zones) for a particular purpose. There are two requirements for the clusters. First, the cluster property should hold, i.e. for this type of data sets, the intra-cluster variability in feature space should be low, while the inter-cluster variability in feature space should be high. Second,

these clusters should be mostly contiguous to enhance the understanding and exploration of the underlying (soil and other) processes. This latter requirement is not motivated by the available field technology which is nowadays able to handle heterogeneity quite well but rather by the need for human understanding, which is the key idea of data mining.

Hence, the most appropriate definition of a spatial cluster with regard to geographical information systems (GIS) and the respective data sets is provided by [Jacquez, 2008]:

A spatial cluster might then be defined as an excess of [...] values (for field-based data, such as a grouping of excessively high concentrations of cadmium in soils) in geographic space. [...] For now, it is useful to think of a “cluster” as a spatial pattern that differs in important respects from the geographic variation expected in the absence of the spatial processes that are being investigated.

This may be regarded as an extension to the general clustering principle, where one wants to form groups in such a way that objects in the same group are similar to each other, whereas objects in different groups are as dissimilar as possible [Kaufman and Rousseeuw, 1990]. From a GIS-based point of view, the result of management zone delineation should be a choropleth map (cp. Section 2.7.5).

Historically, the development of management zones has been spawned by the need to characterize spatial variation with little to no soil sampling. According to [Khosla et al., 2008],

[...] these zones should be regions within a field that have similar yield-limiting factors. Although there are several techniques to delineate management zones, most rely on little to no soil sampling, and so they have the potential to be more feasible economically than sampling on a grid. Regardless of the technique used, once a field has been divided into management zones agricultural inputs are applied variably to meet the yield-limiting factors inherent to each zone.

In this work, the underlying purpose is, e.g., basic fertilization, as laid out above. If soil sampling data are available, such as in the F550 data set, those are likely to be most suited and are used here. If those data are not available, which is usually the case due to the cost of soil sampling, other data such as the electrical conductivity and additional digital elevation model data may be usable. Therefore, although motivated by the specific task of basic fertilization, the focus is rather to develop a generic spatial clustering approach for relatively high-resolution soil sampling data taking into account the specifics of the PA data sets.

4.3 Literature Review in Precision Agriculture

This section serves as a reference that captures the most recent and the most relevant work towards the goal of management zone delineation as a data-driven approach. The literature review on management zone delineation is presented in chronological order. The delineation of management zones has been used as a method of subdividing fields into parts with different properties for a long time. However, this has usually been done using expert

and long-term knowledge of the respective field. The advent of modern GPS technology in the early 1990s has not had a significant impact until 2000, when the *selective availability* restriction was decommissioned, allowing for higher precision of GPS data. Due to further recent advances in technology, the delineation of management zones has turned into a data-driven approach for subdividing the fields. Therefore, the approaches below represent the cutting edge in this topic, spawned by the public availability of GPS from 2000 onwards.

The following literature review is purely focused on the technical part of management zone delineation. A short recent comparison of management zone delineation approaches has been presented by [Guastaferrero et al., 2010], which, however, neglects the exploratory angle to MZD. There are, of course, further topics and different angles to this question. One of those is to consider the zoning problem from the point of economic feasibility: depending on the crop, the field topology, the soil parameters, the available equipment and further parameters, the logistic effort in using pre-determined zones might or might not be worth the net return. In [Tozer and Isbister, 2007], a recent economic study for wheat fields in Western Australia is conducted which concludes that the usage of management zones is profitable in most of the approaches where heterogeneous soil is tilled. For cotton, another recent study is given in [Velandia et al., 2006], also presenting positive net returns. Furthermore, a large portion of the approaches reviewed below also consider the economics of the zoning approach. This is mentioned where appropriate. Based on the recommendations and shortcomings identified in the investigated articles, the (PA-based) requirements for a novel approach towards management zone delineation and variable importance is derived in Section 4.4.

In the following sections, the section titles roughly conform to the original articles' titles. Each section features a description of the authors' work, followed by a short evaluation.

4.3.1 Forming spatially coherent regions through classification

In the study of [Lark, 1998], spatially coherent regions are defined by fuzzy classification, smoothing the fuzzy memberships, then defuzzifying the smoothed memberships to allocate each individual data record to one of the classes. Each of these three consecutive steps is well reasoned for by the author. The overall goal is to find spatially coherent regions by classification of multi-variate data. The author uses three years' worth of yield maps from a 6 ha field in Bedfordshire, East England, UK. The data density is roughly 250 data records per ha.

The first step, fuzzy classification, is done in order to avoid the crisp classification performed by other algorithms. It is reasoned that the class membership information returned by fuzzy clustering is otherwise lost in other algorithms. The class membership is used to determine how to treat spatially neighboring points in the next step. Consider a situation where the class of maximum membership to which an individual belongs is different to that of its neighbors. If its membership in the class of its neighbors is not much less than in the class of maximum membership, then it may be reallocated to the same class as its neighbors. On the other hand, if the individual has a very high membership in its maximum membership class, then it may be decided to treat it differently from its neighbors. These distinctions would not be possible in case of a crisp classification.

The second step consists of spatially smoothing the membership values in each class for each individual. The values are replaced by a spatially weighted average of the membership values in the neighborhood R which is a local subset of all the individuals. This smoothing step, with a few weighting parameters which determine the handling of the different membership values, is subject to empirical settings by the user. It clears the path for the third step, which is the defuzzification of the averaged membership values. The class of maximum membership for each individual depends, as a result, on the original membership values both at the individual and its neighbors. The results of this process are depicted in Figure 4.2, comparing the standard defuzzification approach with defuzzification after spatial smoothing.

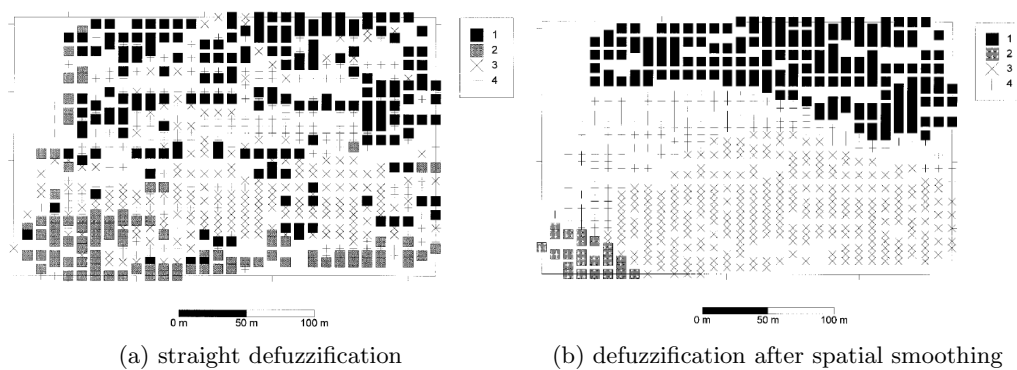


Figure 4.2: Spatially coherent regions through classification, figures from [Lark, 1998]. Figure 4.2a shows straight defuzzification of fuzzy-c-means classes, while Figure 4.2b shows the result when an additional smoothing step is included.

The main issue of existing management zone approaches is clearly identified in [Lark, 1998]: often simple (fuzzy) clustering is applied, without considering spatial relationships between data points. This usually leads to non-coherent (spatially non-contiguous) zones which the three-stage approach above aims to overcome. Though the approach itself is rather successful in the technical stage, there is no meaning in the sense of additional information carried by the resulting zones. Furthermore, additional data variables such as vegetation and soil sampling result in different zones when applying the same procedure as above, which is an expected result. Essentially, due to the weighting and smoothing parameters to be set, this approach is exploratory. A user is bound to experiment with the possible parameters and examine the large solution space. This requires an analysis of the generated management zones. It is unlikely that this approach generalizes well to data sets which contain more than a yield data variable. There would certainly have to be a tradeoff between spatial contiguity and the number of management zones. It is therefore unlikely that this approach can be applied as-is to the multi-variate data sets available in this thesis. Nevertheless, the basic concept of creating spatially coherent regions should be extended to creating spatially coherent *meaningful* regions using additional data variables.

4.3.2 Management Zones through Data Layering

The approaches of [Franzen and Nanna, 2003, 2006] represent a basic data analysis approach to zone delineation for nitrogen management. In [Franzen and Nanna, 2003] different data variables are evaluated towards their potential for zone delineation. First, single variables are used and their correlation coefficients with base nitrate values are determined. Second, combinations of variables and their correlation coefficients are determined. The six variables are topography, yield, order 1 soil survey¹, aerial photography, satellite imagery and EC_a . A subset of the variables consisting of topography, satellite imagery and yield mapping is determined as yielding the highest and most consistent correlation. Those three variables are used in a simple summation of those layers, with a few weights attached. The results are shown in Figure 4.3. The main result of this work is that multi-year data should not be used since the zones differ considerably between years. Although the result in Figure 4.3 is essentially a contour map, which may be valid as the result of a MZD approach, a simple weighted data variable summation depends fully on the weight settings. Furthermore, [Franzen and Nanna, 2006] assume explicitly that the underlying data variables are positively correlated with each other, which is not the general case.

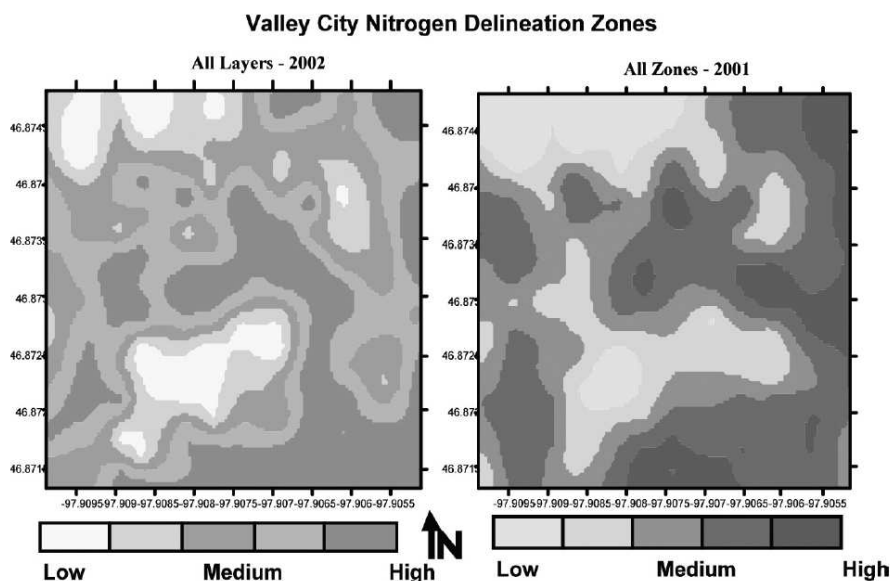


Figure 4.3: Nitrogen delineation zones from [Franzen and Nanna, 2006], those are similar to the contour maps introduced in Section 2.7.4. The zones differ between years.

In a subsequent article, [Derby et al., 2007], two similar management zone delineation approaches are presented, among others. The data were obtained on a field in North Dakota, USA. First, based on the three variables N (soil test nitrate), deep EC_a and yield,

¹A soil survey of order 1 is the most fine-scale, expert-based survey which could traditionally be achieved. It allows to deduce information about field subareas down to 1 ha. Higher order soil surveys are less fine-scale, up the size of districts and counties.

a clustering is performed with a fixed number of four to-be-determined management zones. In addition, a data layer weighted overlay method is described. This essentially consists of using elevation, EC_a and newly created yield rank data for a summation. The yield rank data are generated from the previous five years of yield data. For each data point it is determined whether it was above, below, or approximately at the average yield for this site-year, making the approach more robust and considering intrinsic properties of the field which are not reflected in the data. The five-year data are then averaged per data point, which returns the yield rank data. From those three variables, the management zones are determined via a simple weighted summation method, resulting in the desired four different management zones.

Both zone delineation methods, of which the second is shown in Figure 4.4, return similar zones. However, there is no clear justification of the results when comparing them to the yield potential maps. It is also pointed out that using the management zones is economically inferior in those years when the weather conditions are the main impact factor determining yield. This indicates that additional variables are necessary which cover the weather impact during the growing season. Similar to [Franzen and Nanna, 2006] the choice of summation weights determines the final management zone delineation. Figure 4.4 also shows that the spatial contiguity of the resulting zones is neglected or would have to be tackled in an additional step rather than in the actual zoning approach.

4.3.3 Cluster Analysis of Yield Patterns

In [Jaynes et al., 2003] a multi-year approach for determining management zones on a field in Iowa, USA, is presented. The authors aim to explain interrelations and correlations between variables and between different years. Sampled data from six years are used, with a total of seven variables in addition to the yield variable. The data are sampled along eight transects with 28 sampling points each, resulting in 224 data points for this 16 ha field. The variables are as follows:

Yield: the corn yield in the respective year

EL: the elevation of the field

SL: the slope of the field at the data point

PL: the plan curvature of the surface (perpendicular to the direction of SL)

PR: the profile curvature of the surface (in the direction of SL)

AS: angle of slope, in degrees, calculated as $north - 180$

EC_a : apparent electrical conductivity, measured with EM-38 sensor

SC: soil color, from airplane imaging, after image processing

Further details on the data and the field properties can be found in [Kaspar et al., 2003], as well as in a revised version of [Jaynes et al., 2003] in *CompAg* ([Jaynes et al., 2005]).

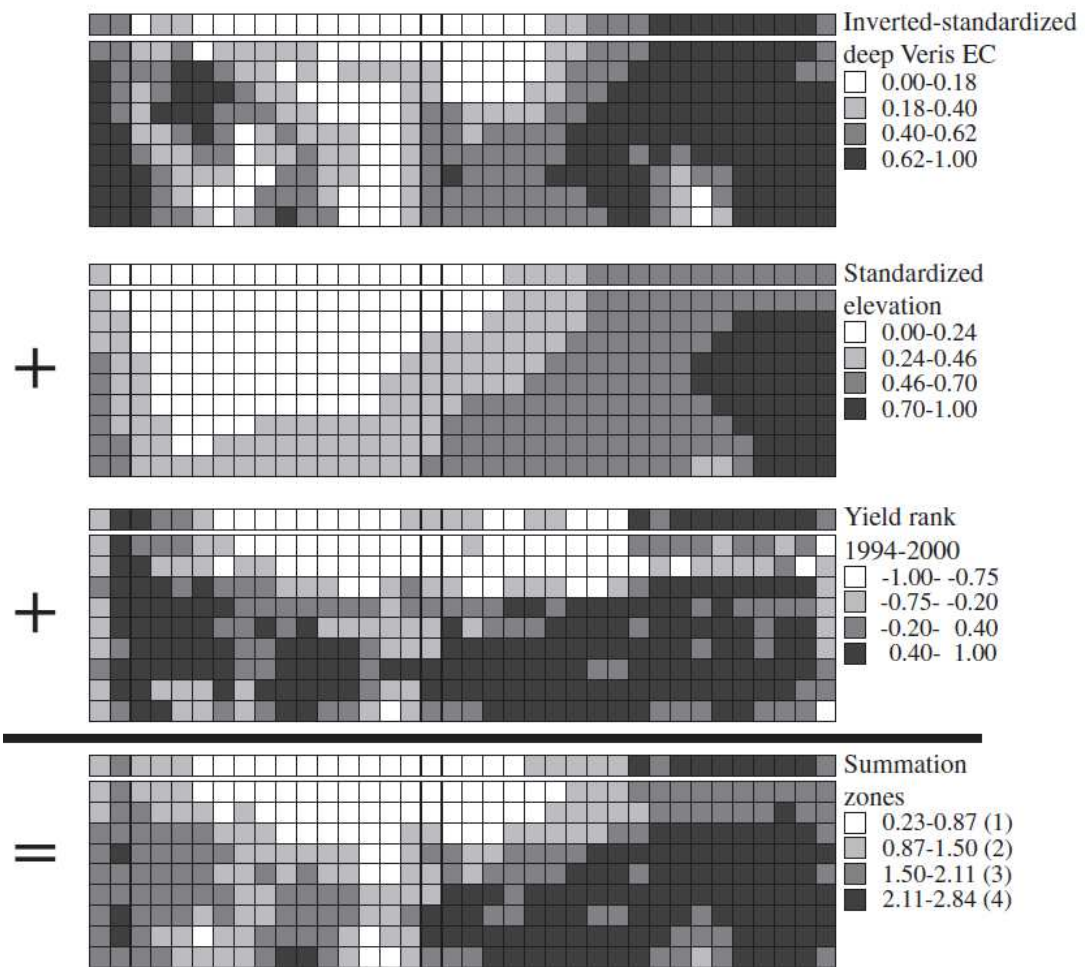


Figure 4.4: Grid-based management zones using data summation from [Derby et al., 2007].

Based on the above variables, the authors propose a three-step process of *partitioning, interpretation and profiling*. A (non-spatial) k -means cluster analysis is performed, mainly for the purpose of data reduction to a manageable number of types. It is reasoned that the clustering algorithm groups each yield plot such that it belongs to one and only one cluster. The number k of clusters is varied from 2 to 8 to determine the optimal number of clusters. The spatial autocorrelation of the resulting clusters is determined using Moran's I statistic. Optimal results are obtained with cluster numbers of three and five. The clustering process succeeds in dividing temporal yield patterns into spatially contiguous areas. A detailed explanation using expert knowledge of the field follows, confirming the method's success under the given circumstances.

Once the clusters are obtained, a canonical multiple discriminant analysis is performed to reveal which field variables contributed significantly towards classifying yield plots into clusters. The above-mentioned variables are used in a stepwise discriminant process which is analogous to stepwise regression. In the first step, the field variable which contributed most to discriminating between the clusters is brought into the discriminant model. In the second step, the variable that contributed most to the discriminatory power of the model, and is not already in the model and above a certain significance level, is entered into the model. Before a new variable is added, it is also checked whether the weakest variable in the model exceeds a preset significance level. The stepwise process continues until all variables in the model meet the criterion to stay in and none of the remaining variables meet the criterion to enter the model. It is reported that EC_a , PL, SL and EL (in this order) contribute significantly to summarizing the differences between the clusters. Due to the correlation between some variables, especially in relation to EC_a , this variable contributes a major part to the clustering.

The above approach targets two important aspects related to yield data. First, for the determination of management zones it is typically unclear what the precise number of zones should be. This signifies the usage of an exploratory approach. Second, the authors point out that the clustering algorithm does not consider any spatial information and would deliver spatially non-contiguous zones if the underlying correlations and spatial point patterns were less pronounced than with the provided data sets.

4.3.4 Characterizing Spatial Variability Using Management Zones

The authors of [Schepers et al., 2004] follow the idea of assessing spatial and temporal variability on irrigated fields with management zones. The data used are taken from irrigated corn fields in Minnesota, USA. The available data are a five-year series of corn yields, bare soil brightness images, elevation and EC_a , at a data density of around 5,700 points for this 51 ha field, giving a data density of around 110 points per hectare. Based on the assumption that not all of the available data have an effect on the spatial variability of yield, a principal components analysis is performed. Out of the five principal components, the first two explain 85% of the total variability and are therefore retained for the management zone delineation approach. The authors do not elaborate in more detail how the delineation is performed, except for the fact that an unsupervised classification is applied. This is likely to be a clustering approach. The resulting four management zones are quite distinct

in their soil properties and also in their yield variabilities. However, it is noted that the temporal variability of yield (five site-years) is much greater than the spatial variability between the management zones. It is assumed that the temporal variability is due to different precipitation amounts, even though the field is irrigated. The authors emphasize that the management zones approach is valid, but should be generated from one site-year only since temporal data changes can not be captured by static management zones.

The research presented in [Schepers et al., 2004] considers the circumstances of temporal and spatial variability in between and within management zones. The generation of these zones is quite straightforward: the authors use the first two components of a principal components analysis and present these to an unsupervised classification algorithm, which is likely to be a variant of clustering. Using the principal components blurs the single variable's effect on the management zone and, much worse, leads to zones which are hard to explain. The resulting four management zones in the study capture a significant amount of the spatial variability of the field, but fail to account for the temporal differences in yield over the five site-years. This leads to the assumption that multi-year data should not be used for a management zone delineation approach since external influences such as weather and especially precipitation lead to errors in this modeling stage. Furthermore the authors point out that the use of crop-based in-season remote sensing data could have a large impact on this management zone approach, which has also been shown by [Raun et al., 2002]. These data can cheaply be acquired and allow for the assessment of the crop's nutrient status, especially in terms of nitrogen.

4.3.5 Fuzzy C-Means Clustering on EC_a , Yield and Elevation Data

An advanced approach for management zone delineation is described in [Kitchen et al., 2005]. The authors propose an idea for semi-automatic management zone delineation using fuzzy c-means clustering. They use yield, elevation and EC_a data from ten consecutive years on claypan soil fields in north-central Missouri, USA, growing corn and soybeans. They split the data into two sets, one containing the yield information and the other one carrying the EC_a and elevation information. Both data sets are then clustered using fuzzy c-means clustering as built into the management zone analyst software [Fridgen et al., 2004]. For univariate clustering, the Euclidean distance is used, whereas for multivariate clustering the Mahalanobis distance measure is used. They aim to extract three management zones, since it is argued that more than three or four management zones are of little use in practical terms. The clustering process on both data sets returns two maps with (likely) different management zones, which are then compared. The overall agreement between the different clusterings is determined via the overall accuracy and a coefficient calculation. Different combinations of parts of the EC_a data set are also generated and the obtained clusterings are compared. Finally, according to the authors, the approach is successful in generating appropriate management zones. In addition, it can be confirmed that EC_a data can be used for this purpose. A depiction of the zoning approach is given in Figure 4.5.

The approach by [Kitchen et al., 2005] fails to explicitly account for the spatial relationships of the data points by using a non-spatial clustering algorithm. This leads to spatially scattered and non-contiguous zones, as shown in Figure 4.5. Moreover, judging from the

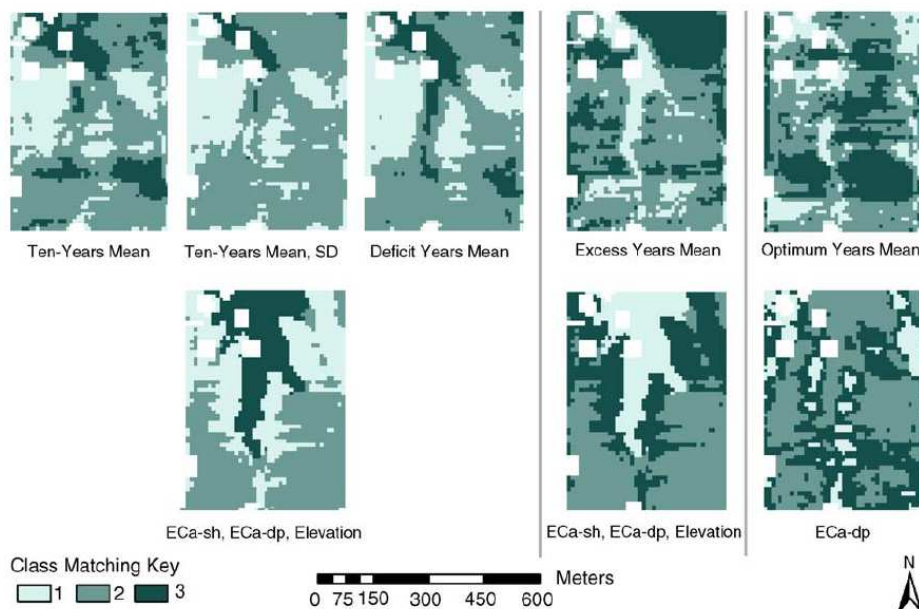


Figure 4.5: Figure from [Kitchen et al., 2005], reference yield productivity zone maps (top) compared against the best performing productivity zone maps derived from unsupervised clustering of EC_a and elevation (bottom).

figure, the zones are only marginally related to the underlying data and would be hard to explain in a simple way. However, when soil sampling data are not available, it can be confirmed that the EC_a value may be used, although cautiously and without enabling conclusions to specific underlying soil properties.

4.3.6 EC_a and Yield Maps for Management Zones

The authors of [King et al., 2005] investigate the relationship between electromagnetic sensing (EC_a) and yield mapping for delineation of management zones. The data are obtained from four fields in the UK, varying in size from 6 to 18 ha and having contrasting soil types. The crops grown are winter wheat, spring beans and winter barley. Soil samples are taken at a density of roughly 1 sample per hectare on an approximately regular grid; the samples are then analyzed for particle size distribution, organic carbon content and bulk density. A measurement for the available water is determined. Furthermore, EC_a readings (using the EM38 sensor) are taken at the study sites and the sequences of yield maps (multi-year maps) for these fields are included into the data sets.

On these data, fuzzy clustering is applied to determine management zones, where the optimal number of zones is determined by using the normalized classification entropy value, resulting in four management zones. The zones based on the yield maps seem to be closely related to some of the soil properties, which is determined by regression modeling. In particular, subsoil clay content and subsoil sand content seem to be a major factor of influence for yield. On the other hand, the EC_a readings correlate strongly with the type

of soil. Furthermore, this correlation pattern holds true independently of whether the EC_a measurements are taken in moist or dry conditions, or during summer and winter. Again, regression modeling suggests that EC_a is strongly related to topsoil clay and sand, with some minor correlations to further physical properties. Clustering of the EC_a readings therefore results in stable and reliable management zones which are clearly related to soil properties and eventually to yield.

The process of clustering yield, soil and EC_a data undertaken in [King et al., 2005] clearly shows the advantages of using apparent electromagnetic conductivity measurements for determining management zones. The readings of the EC_a sensor are stable, regardless of the weather conditions that the field is subjected to. Furthermore, the management zones generated from EC_a readings are closely related to soil properties and therefore also closely related to the available yield maps. Hence, expensive soil sampling techniques may (to a certain part) be replaced by the inexpensive and non-invasive technique of measuring the apparent electrical conductivity. The application of fuzzy c-means clustering is similar to the one presented in Sections 4.3.5, 4.3.7, 4.3.8 and 4.3.12, and may lead to spatially non-contiguous zones since spatial relationships are not considered in the clustering.

4.3.7 Yield-Based Zones for Crop Rotations

The authors of [Brock et al., 2005] investigate the delineation of management zones for a corn-soybean rotation in east-central Indiana, USA. Special attention is paid to evaluate the usage of fuzzy c-means clustering for delineating zones and to compare different management zones for different crops with each other. The data sets consist of multi-year high-resolution yield data, collected at 1 second-intervals using yield monitoring systems, coupled with a differential GPS. After preprocessing, these data are available to the management zone delineation approach.

Both objectives mentioned above are achieved by running a fuzzy clustering algorithm on the yield data. Based on whether the clustering is performed on the complete data set or a crop-based subset, different management zones result. The yield-based management zones (YB) use the complete data set, whereas the corn-yield-based (CYB) and the soybean-yield-based (SYB) zones use the respective subsets. For determining the optimal number of zones, the fuzzy performance index and the normalized classification entropy are used. In addition, the total reduction of within-MZ variance is determined. It is investigated whether there are significant differences between the SYB and CYB zones as well as the YB zones. The results are that, depending on the four fields under study, management zones are quite different for different crops – therefore, the combination of consecutive years of different yields is not recommended. This effect is likely to be due to the different soil requirements for each crop. The different clustering performance indices return different numbers of optimal clusters, between two and six, with a recommendation of three management zones. It is further mentioned that the land area of the smallest possible management zone (equipment restrictions) should be considered in the determination of the optimum clustering.

The first task presented in [Brock et al., 2005], using fuzzy c-means clustering for zone delineation, is quite similar to Sections 4.3.5, 4.3.6, 4.3.8 and 4.3.12. The creation of management zones is successful, at least based on purely yield data. However, using multi-year

and multi-crop yield data from the same site is not recommended, since the management zones for different crops are relatively different. The size of the management zones must not fall below a certain level, which is determined by the available machinery and further economic considerations. As with the remaining fuzzy c-means clustering approaches, spatial contiguity is not assured and different numbers of management zones can only be generated by repeating the clustering with a different parameter value.

4.3.8 Site-specific Management Zones via Fuzzy Clustering

[Li et al., 2007] present an approach that uses remote sensing and sampled soil data to delineate management zones in coastal saline land in northern China. The authors work on a 10.5 ha field growing cotton. The available data for this field are the normalized differenced vegetation index (NDVI) from satellite imagery, 139 soil samples on a regular grid and EC_a measurements. The soil samples are processed into the following seven variables: OM, TN, AN, AP, AK and CEC. Furthermore, the end-of-current-season yield is available. Following a conventional statistical analysis, a principal components analysis is performed, resulting in two principal components which together explain 88% of the total variability in the data set. Based on these principal components maps, a fuzzy c-means algorithm is used to delineate three management zones. The number of management zones is determined via the fuzzy performance index and the normalized classification entropy, both yielding an optimal number of three components. The found management zones reveal distinctly different soil chemical properties, characterized by the reduction in in-zone variance of each variable. Clear management zones can be found, where the properties of one of these zones seem to be more optimal for crop growth than in the other two zones.

The approach of [Li et al., 2007] is similar to the ones presented in Sections 4.3.5 and 4.3.12. Fuzzy clustering is used after reducing the data to a low number of its principal components. However, the effect of using the NDVI image is unclear. Since this is one of the standardized vegetation indices it should be investigated whether or not it should be used in determining management zones. Furthermore, there might be correlations between the NDVI image and certain crop variables which may be interesting from a data cleaning perspective. Some of the variables may be redundant. Again, spatial contiguity can not be ensured. Furthermore, having only the principal components in the final zones leads to problems with an explanation of the zones. The study is one of the few to actually use high-resolution soil sampling data.

4.3.9 Management Zones based on Soil Fertility

In [Ortega and Santibáñez, 2007], a cluster-based approach to management zone delineation is compared with two additional ad-hoc approaches fulfilling the same purpose. The data are sampled from 13 fields in Central Chile with an average density for 8.1 and 11.9 soil samples per hectare, for the 2002/2003 and 2003/2004 corn-growing seasons, respectively. The soil samples which are taken are analyzed for six chemical properties: pH, EC_a , OM (organic matter), AN (available nitrogen), AP (available phosphorus) and AK (available potassium).

First, a *k*-means clustering is computed for the field, based on the standardized variable values. A fixed number of four management zones is chosen because (according to the authors) this is the maximum number of management zones that can be handled by conventional farming equipment. Second, based on the available six variables plus the dry-matter yield variable, a SI (soil index) variable is computed. SI is a simple linear combination of the variables' values for each data point. To determine the weights for this linear combination, two methods are proposed, which are described as follows:

SIPC – principal components analysis The weights for the linear combination of the variables are taken from a principal components analysis. Generally, the first few components explain most of the total variance in the data set.

SICV – coefficient of variation It is assumed that those variables showing more variability in the field should have a greater weight in a linear model. The variables are thus standardized to the $[0, 1]$ -interval. This maintains the value of the coefficient of variation (CV). The relative weight for each variable is obtained as $w_i = \frac{CV_i}{\sum_j CV_j}$, where the i subscript refers to the respective variable's value.

The SIPC and SICV values are calculated, mapped and classified into four management zones. The classification criterion for the four classes is described as the mean soil index plus or minus one standard deviation, which yields four ranges for the SIPC and SICV values, respectively.

The effectiveness of the approach is judged by the relative variance (RV). It reflects the proportion of the variance of any variable explained by the zoning algorithm. Its interpretation is similar to that of the coefficient of determination in linear regression (R^2). RV is calculated as $RV = 1 - \frac{S_w^2}{S_T^2}$, where S_w^2 is the variance of variable w within the management zone and S_T^2 is the total variance of this variable within the whole field [Dobermann et al., 2003]. The results of the three zoning methods (clustering, SIPC, SICV) are rather mixed. The degree of correlation between yield, soil chemical properties and the zones is weak or non-existent. It is argued that the measured soil properties are not the limiting factor for yield and that there are likely other soil properties determining crop yields. The effect could also be due to a generally high level of applied fertilizer which may mask otherwise significant relationships. It is also suggested to use topographical and remote sensing data which have been found to be valuable for management zone delineation.

The above approach by [Ortega and Santibáñez, 2007] consists of delineating management zones through three different methods: clustering, SIPC and SICV. The methods themselves are justified, however, the sampled soil data are likely to be insufficient for the task of zone delineation. Therefore, as the authors suggest, topographical and remote sensing data should be used, rather than or in addition to the soil data. The authors also suggest a maximum of four management zones. The clustering approach seems to be the most straightforward idea – however, the details are unclear, such as the distance/similarity measures or the actual input data. The SIPC and SICV approaches are rather complicated when compared to the simple clustering approach and may hamper understandability of the management zones. It is also unclear whether a linear combination of the soil variables

is sufficient to be used as an input to a zone delineation approach. Again, the spatial contiguity of the zones is not ensured and spatial data aspects are not taken into account explicitly.

4.3.10 Management Zone Delineation by Image Analysis

[Roudier et al., 2008] propose to use an improved watershed algorithm for management zone delineation, based on image analysis of different soil and crop characteristics. The standard watershed algorithm works as follows: given a grayscale image of a field, acquired from remote sensing facilities, construct the gradient image. This should have a minimum where the original image has a valley or a ridge and it should have a maximum where the original image has a steep slope. A flooding algorithm is applied on the gradient image, starting from the (local) minima. The flood level rises; whenever two different floods meet, a “dam” is built up. Those “dams” are the resulting contours of the segmentation. However, this standard watershed algorithm results in a large number of segments, which is often not useful and would require extensive post-processing.

Therefore, the authors propose the introduction of a “flooding lag”. This basically means that a dam is not built up each time two floods meet, but rather when there is a significant difference in height. This parameter has to be determined experimentally and a method for estimating it using a variogram is proposed. Finally, the number of zones is reduced by merging specific neighboring regions in a regularization step, where a *fit parameter* is used to determine which regions to merge. This *fit parameter* is essentially an energy function determined as a linear combination of compactness, regularity and radiometry (see [Baatz and Schäpe, 2000] for a description of this last parameter).

The new version of the algorithm is then tested on four study sites in Burgundy, France. The input variable is termed *biomass* and is acquired from an airborne sensor. 11 spectral bands are used, with a spatial resolution of 5 metres. The standard algorithm results in over-segmentation, splitting the fields into 100–350 management zones. The improved algorithm reduces the number of zones to 8–40. Despite this still being a high number of zones, the results are satisfactory to the authors.

It is noted by the authors that the work successfully tackles an issue in the watershed algorithm but raises a new issue in the regularization step – determining the optimal number of zones. Furthermore, their solution can not be automated since manual labor is required when estimating the variograms. Nevertheless, a suggestion is pointed out: merging spatially adjacent zones to reduce the number of management zones. From a practical point of view, a small number (3–5) of management zones is suitable. Therefore, a simple and computationally inexpensive approach to field segmentation could be used initially. After this step, appropriate criteria should be developed for merging neighboring zones. It would clearly be desirable for these zones to be contiguous, although this is not necessary from the point of modern farming technology. The spatial resolution of the data in this study is relatively high, yet there is only one variable being used in the determination of zones. Further field variables should therefore be taken into account. The spatial contiguity of the management zones has been explicitly considered.

4.3.11 Site-Specific vs. Yield-Based Management Zones

In [Khosla et al., 2008] a long-term study of the differences between site-specific management zones (SSMZ) and yield-based management zones (YBMZ) is conducted. The study is carried out on five grain fields in northeastern Colorado, USA. The SSMZ approach uses three data layers: a) panchromatic aerial imagery of each field following conventional tillage operations; b) producer's knowledge of the field's topography and c) the producer's past crop management experience in the field. The YBMZ is intended to improve the accuracy of the SSMZ technique by including yield monitor data. Five data layers are available here: a) color infra-red bare-soil aerial imagery, b) soil organic matter content, c) soil cation exchange capacity, d) soil texture and e) the yield map of the previous growing season. Both techniques then continue to apply a k -means clustering approach to the data layers, with a desired number of three management zones.

Henceforth, the resulting zones from both techniques are compared to each other using three criteria: a) crop productivity, b) an assessment of accuracy and c) experts' subjective classification. For the **crop productivity**, the three resulting management zones of each approach are compared based on the grain yield. The SSMZ's yield in the high and medium zones is significantly higher than that of the YBMZ approach in its respective zones. For the low-yielding zones, the yield results are reversed. This clearly favors the SSMZ over the YBMZ approach in terms of identifying productivity potential. In terms of the second criterion, **accuracy assessment**, both approaches' management zones are compared to an additional yield clustering into high, medium and low yield. Finally, the method returns a percentage of agreement between the YBMZ and yield and between SSMZ and yield. It is found that the SSMZ approach relates more strongly to the spatial patterns of grain yield. The **subjective evaluation** returns the same result: SSMZ seems to better characterize the field's spatial patterns than YBMZ. However, both methods fall short of fully capturing the fields' entire spatial variation.

Since SSMZ seem to be the more promising approach, further research into the relations between zones and soil properties is conducted. It is pointed out that special soil properties, such as soil bulk density and soil texture are statistically significantly different between the zones of SSMZ. Since an aerial image was used as one of the data layers for SSMZ, this is unsurprising. However, it is concluded that further features of the soil help to better delineate management zones. Further details on this study can also be found in [Hornung et al., 2006].

[Khosla et al., 2008] presents an effective comparison between the creation of yield-based management zones vs. the creation of site-specific management zones. Both zoning approaches differ in that they use different data layers: yield-based zones are generated based on the site's heterogeneity in yield, while site-specific zones are based on the variability of other and/or additional variables. It can be shown that purely yield-based approaches should be replaced by or at least augmented with approaches that evaluate additional available data layers. However, it can clearly be seen that "more data layers" does not automatically mean "better management zones". Furthermore, it is unclear how much the expert knowledge in the SSMZ approach contributes to the accuracy of the zoning. This is especially important considering that expert knowledge is usually expensive and must

be collected manually and (often) tediously. An additional drawback is the relatively low sampling density of 2.5 points/ha, due to the manual sampling approach. Again, a non-spatial clustering algorithm has been used which does not explicitly account for the spatial contiguity of the clusters.

4.3.12 Management Zones based on Soil Chemical Properties

In [Xin-Zhong et al., 2009], a rather recent approach to the delineation of management zones is presented. The authors study a tobacco field in Central China. Their data are based on manually taken soil samples on a 100 m grid, resulting in 81 points for this 87 ha field. Their first research objective is to quantify the spatial variability of soil fertility variables and the second objective is to delineate the management zones by the combined usage of principal components analysis (PCA) and a fuzzy clustering algorithm. The available variables are values for pH, organic matter, total nitrogen, alkalytic nitrogen, available phosphorus, available potassium and cation exchange capacity. After a thorough descriptive statistical and geostatistical analysis, the core components of PCA and fuzzy clustering are explained. Fuzzy clustering is applied on the first and on the combination of the first and second principal components (PCs) of the data set. The first PC explained roughly 50% of the total variance and the second an additional 21%. A depiction of the principal components can be found in Figure 4.6.

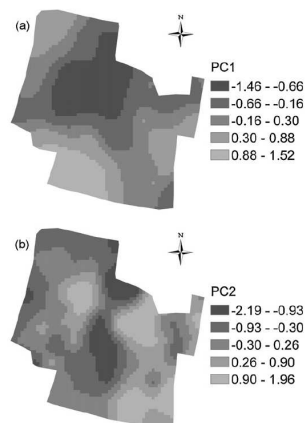


Fig. 3 – Contour maps for (a) first and (b) second principal component (PC).

(a) Principal Components

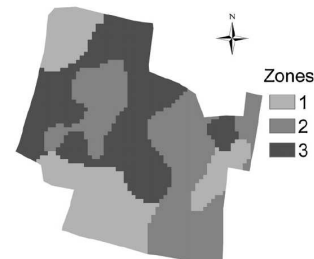


Fig. 5 – Management zones (MZs) map for optimum clusters in study area.

(b) Resulting Zones

Figure 4.6: Management zones from [Xin-Zhong et al., 2009], showing principal components and resulting management zones. Although the resulting zones are in principle what should be expected and the zones are mostly contiguous, the principal components do not allow for an easy interpretation of the zones.

The optimal number of clusters is determined via the FPI (fuzzy performance index) and the NCE (normalized classification entropy). The results of running the clustering algorithm with different numbers of clusters suggest that a number of three management

zones is an optimal choice. A comparison of the three management zones returns mixed results. On the one hand, the in-cluster variability of some chemical properties is low and the intra-cluster variability is high, as should be expected. This is true for pH, OM, AP, AK and CEC. On the other hand, there are no significant differences in variability for AN and TN. The productivity (yield) levels on the field are not taken into account. The results are inconclusive, due to temporal, in-season changes the authors also suggest that taking crop-based remote sensing variables into account would be useful. This could lead to more efficient application of nitrogen fertilizer.

The above approach of [Xin-Zhong et al., 2009] represents an interesting application of principal components analysis and fuzzy clustering towards the delineation of management zones on a tobacco field. The key drawbacks of this article are as follows: First, the spatial resolution of the sampling points is relatively low at 0.93 points/ha. At this granularity, the borders of the management zones are likely to be quite fuzzy. Second, the results of the management zone delineation are inconclusive. There are some interdependencies in the data set which could be explained, but there are many others left unexplained. There is no indication whether some of the variables are therefore more or less useful for the purpose of zone delineation. Third, although the resulting map is an example of a rather successful approach to delineating zones, the zones can hardly be explained due to the underlying principal components analysis. As the authors note, due to the in-season temporal changes on the field, a combination of soil sampling with in-season crop-based remote sensing might be useful in capturing changes within the zones.

4.3.13 High-resolution Remotely Sensed Data for Zone Delineation

A very recent approach to management zone delineation has been undertaken by [Song et al., 2009]. Based on soil sampling, yield and remote sensing data, the authors investigate using different sets of data layers for generating management zones. The data were obtained from a field in the Changping district of Beijing, at an experimental station for precision agriculture. Winter wheat was grown with two dressings of nitrogen fertilizer. The soil sampling area which is used in the experiments has a size of $90 \times 90 m^2$. 81 soil samples are taken on a regular grid of $100 m^2$ per grid cell, resulting in 100 points per hectare. The wheat yield was recorded such that each grid cell was $3 m^2$ in size, resulting in a spatial resolution of 3,300 points per hectare. In addition, commercial imagery from the Quickbird satellite is available, at a spatial resolution of 0.6 m, resulting in 22,500 data points in four spectral bands. The optimized soil-adjusted vegetation index (OSAVI), which is closely related to vegetation properties such as leaf area index, vegetation cover, vegetation biomass and crop growth, is computed. The yield data are those obtained at the end of the considered growing season and not those of the previous season.

Based on the above data, three subsets are generated: a) soil OM, AP, EK (extractable potassium) and wheat yield, b) remote sensing data (OSAVI), c) the union of the sets of a) and b). Fuzzy c-means clustering is used on each data set to create management zone maps. The optimal number of clusters is determined in the same way as in Section 4.3.3: by varying the number of clusters from 2 to 8 and considering the fuzzy performance index. The optimal number of zones is, again, determined to be *three*. The resulting management

zone maps are compared via the coefficient of variation of the different variables. The results are, however, mixed. Due to the experimental structure, it can easily be compared whether the addition of remote sensing data into a zoning approach lowers the spatial variability in terms of yield – this is clearly not the case, since the coefficients of variation for yield in each management zone are only marginally better once remote sensing data are added.

The authors of [Song et al., 2009] conclude that Quickbird OSAVI data can reflect the spatial variation in wheat growth during the early growing stage and can also show the spatial variation in soil properties and yield. This result, however, can certainly not be concluded from the shown experiments. On the contrary, the presented data clearly state that using remote sensing data in addition to soil and yield data does not improve the coefficient of variation in terms of yield for the management zones. Furthermore, from the results it is also clear that management zones based on remote sensing data alone can only marginally improve the coefficient of variation and should therefore only be used in conjunction with further data from the considered field. Nevertheless, the authors' work is one of the few that works on high-resolution sampling and remote sensing data. While it is in principle desirable to have very high-resolution data, the area of study should also be statistically representative, which may be an issue with this work. Again, the spatial contiguity is not ensured due to using fuzzy c-means clustering without considering the spatial nature of the data.

4.3.14 Recent Survey on Site-Specific Management Zones

Finally, [Khosla et al., 2010] presents the most recent survey on the usage of management zone delineation approaches in the past two decades. Most importantly, it gives an overview about the different site properties that have been used to delineate zones. Figure 4.7 supplies a graphical representation of the numbers presented by the authors. While most approaches were based on chemical and physical properties as well as on sensing data, less work has been done using landscape variables and crop properties. It is assumed that a comprehensive exploratory algorithm for management zone delineation may fall into most of the categories presented in [Khosla et al., 2010] by being able to use most of the available data sources. Moreover, Figure 4.8 shows that different management zones are to be expected when using single data variables as an input for management zone delineation. This also confirms the expectation that the management zones are bound to be different depending on the purpose they are developed for.

4.3.15 Summary of Management Zone Delineation Approaches in Precision Agriculture

In the previous Sections 4.3.1 through 4.3.14, the major relevant work of the past decade regarding management zone delineation approaches was presented and evaluated. The existing approaches were selected to be data-driven rather than manual. The fields and the used data sets are necessarily quite diverse, covering a large spectrum of available approaches. Each of the previously reviewed advances were rather successful in their particular intended purposes. Nevertheless, a few recurring issues can be identified, which boil down to three

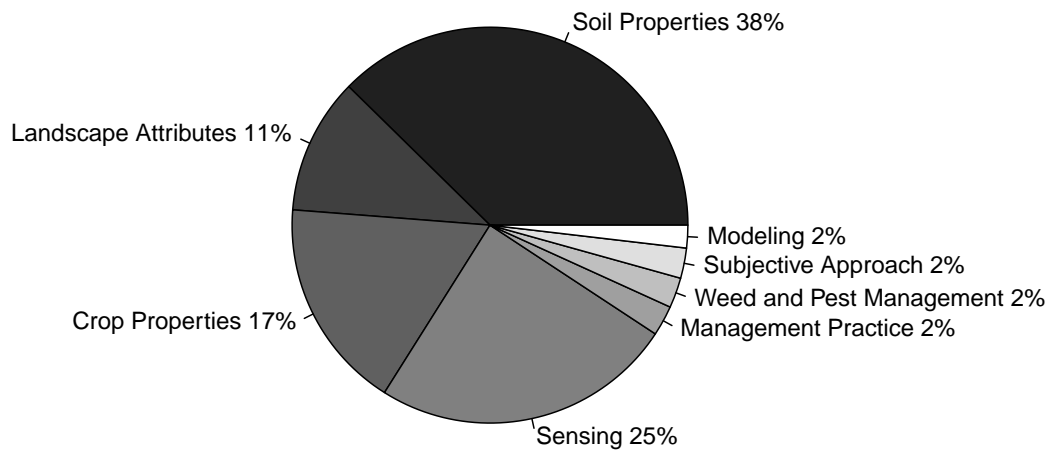


Figure 4.7: Proportions of data and field properties underlying different management zone delineation approaches, from [Khosla et al., 2010].

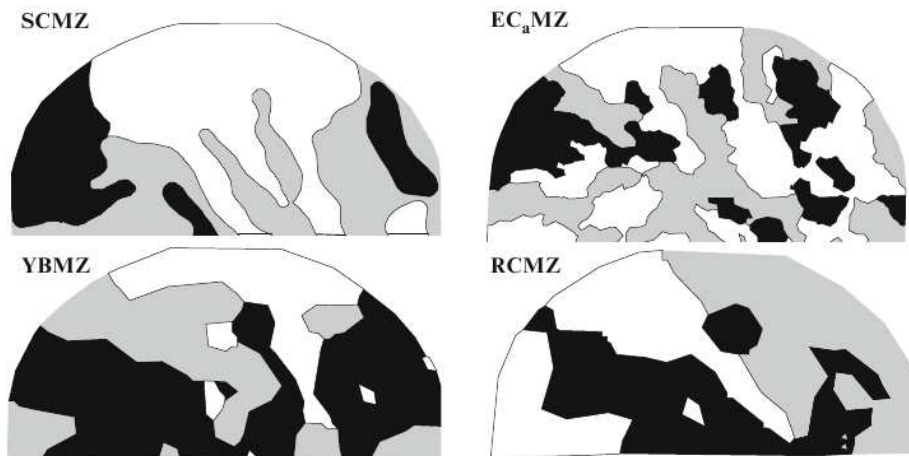


Figure 4.8: Comparison of four management zone delineation approaches based on different data sources. The three gray levels depicted represent low (white), medium (gray) and high (black) productivity zones. Although the authors of [Khosla et al., 2010] conclude that only the medium zone lacks correspondence in the four plots, this seems to also be the case for the low and high productivity zones, at least visually.

major points summarized in the following. These are cast into specific requirements in the following Section 4.4.

Subjectivity of Management Zones

The management zones that are delineated are always tailored to the particular application. This typically means that each approach is inherently subjective and that there is no generic approach that could fit each application equally well. On the other hand, this also means that management zone delineation is never done against benchmark zones in a way similar to a clustering result being compared to a reference clustering. Hence, there is no absolute quality measure for the result of an MZD approach. Although the result of an MZD is sometimes compared to a reference map (e.g. yield distribution or similar), MZD itself is a rather exploratory data mining task. A user aims to enhance his understanding of the study area and would like to gather insights into the available data sets, which is one of the goals pursued in clustering approaches. So the result of MZD should rather be a structure that supports exploration of the spatial data (as a tool) instead of providing a fixed clustering.

Understandability of Management Zones

Along similar lines as above, the understandability of the derived management zones should be ensured. Since the zones are subjective and the user aims to gather insights, the original data variables should be retained as far as possible. Finding zones after applying principal components analysis to the data, using linear combinations of the data, or applying further transformations might lead to more crisp zones and more straightforward computation in the end, but the understandability of those newly generated variables is typically lost.

Non-Spatial Clustering Applied to Spatial Data

An issue which is at the core of this thesis is the usage of spatial methods for spatial data sets. In most of the related work, non-spatial clustering, such as k -Means or fuzzy c -Means, has been applied to spatial data. This often lead to discontinuities in the resulting zones which are in some of the cases smoothed or otherwise “fixed” in the aftermath. Moreover, algorithms like k -Means or fuzzy c -Means are sensitive to their initialization, which may lead to unstable results. An exploration of the possible solution space is mainly achieved by changing the k or the c parameters, leading to different numbers of management zones but with typically highly different spatial layouts: setting k to four and five should from a user’s point of view result in similar solutions (i.e. local changes), rather than totally different ones.

4.4 Requirements

Assuming that management zone delineation (MZD) is a spatial clustering task, as described in Section 4.2, the requirements for a spatial clustering algorithm, such as those laid out in [Estivill-Castro and Lee, 2002], should be obeyed. Furthermore, additional requirements resulting from the specific MZD task should also be complied with. Therefore, this section is subdivided into these two requirements groups. The groups overlap partly. There are

hard requirements which must be considered (H) and soft requirements which are desirable (S).

4.4.1 Spatial Clustering Requirements

The generic spatial clustering requirements obtained from [Estivill-Castro and Lee, 2002] and tailored to management zone delineation are presented in the following.

H1: multi-level In the final clustering or in intermediate steps (if available), the user should be able to examine single clusters to obtain a deeper understanding of the data. Although eventually a small number of management zones is to be delivered, clearly structured results are likely to reveal further, previously unknown information.

H2: exploratory nature This is related to the previous requirement. Since a user is ultimately interested in examining the particular zones and understanding the data, the clustering algorithm should allow for exploration of the data set rather than just being confirmatory.

H3: incorporate spatial proximity This requirement is at the core of spatial clustering. While in non-spatial settings there is no natural neighborhood of data records, there certainly is one in spatial settings. Adjacent points are also likely to be more similar than more distant points due to spatial autocorrelation in the available PA data.

H4: efficiency and effectiveness The efficiency is measured in terms of computation time (speed) for a clustering, while the effectiveness is determined via the quality of the clustering. In an exploratory setup with PA data, the speed is rather important: a user is likely to try different algorithm setups and quickly wants to check the results. This includes clustering with the whole data set rather than using samples. The quality ultimately is user-dependent and, as such, not considered here further.

S1: only few parameters / understandability In order to minimize preconditions and parameters, the spatial clustering algorithm should have as few of those as possible. It is added that the parameters which *do* exist should have a rather easily understandable meaning to the user, such as the final cluster number.

4.4.2 Management Zone Delineation Requirements

Similar to the general requirements towards spatial clustering algorithms presented in the previous section, MZD poses a few further requirements mainly due to the specialties of the data sets which are outlined below.

H5: no density differences in data Since the PA data in this thesis are sampled on a regular grid (F550) or of uniform spatial density, there are no density differences in the spatial coordinates which may be used as a clustering input. The algorithm has to handle those data accordingly. Nevertheless, as stated with requirement **H3**, the spatial neighborhood of a data point must be considered, along with the actual point's variables.

H6: spatial contiguity of clusters While in traditional clustering the clusters are by definition contiguous, in (geo-)spatial clustering this aspect must be considered explicitly. For the purpose of management zone delineation the resulting clusters should be *mostly* contiguous. This requirement is rather fuzzy due to a tradeoff: on the one hand, strictly enforcing contiguity might lead to management zones which conceal important facts about the underlying data. On the other hand, having a few management zones scattered in numerous pieces over a field does not contribute to a better understanding of the underlying processes either. This requirement extends **H3**.

S2: flexible number of management zones A zone might intrinsically consist of sub-zones which may be worth examining, leading to a hierarchical, tree-like structure to be explored by the user. Exploring this structure eliminates the need for a fixed number of management zones and is closely related to **H1**, **H2** and **H4**.

4.5 Spatial Clustering Algorithms

Since traditional non-spatial variants of clustering algorithms typically exhibit a number of drawbacks when used in a spatial setup (cp. Section 4.3), algorithms which aim to deal accordingly with spatial data sets have been proposed. Clustering approaches have been roughly subdivided into four groups: *hierarchical*, *partitioning*, *density-based* and *grid-based* approaches. Most of the spatial clustering algorithms are hybrids and therefore elude this categorization by borrowing ideas from more than one category. The major representatives of those algorithms are presented in the following. They are sorted descendingly according to the number of requirements which are violated. Although a few of the algorithms may be ruled out quickly due to their basic assumptions, they are nevertheless presented to illustrate the main drawbacks and the progress in the area of spatial clustering. The main and recurring aspect of the available data sets is that the data records are located in geographic space. Therefore, a clustering algorithm must be able to incorporate both geographic space and feature space.

4.5.1 Naive Approach: Include Geocoordinates in Data Vectors

A simple approach to include the spatial relationships in the data sets would be to merge the spatial and the non-spatial part of the data. Assuming that in the non-spatial part every data record can be written as a vector of variables, this would add two more variables to each vector: the spatial coordinates (x, y) of the respective data point.

Applying a prototype-based clustering algorithm such as k -Means to those data may work if the expected management zones are remotely similar to a Voronoi tessellation of the field with convex zones. However, this is not the general case. For example, assume a field which is cone-shaped in terms of elevation, with the highest elevation in the center of the field. It is likely that the management zones are best laid out in rings around the center of the field, provided that elevation and other relief variables may play an important role in management zone delineation (cp. Chapter 3). To obtain these zones, the distance (or similarity) calculation would have to be changed according to the user's hypothesis of the zones. For other forms of management zones, the distance calculation would also have

to be adapted throughout the course of the algorithm to account for different shapes of discovered zones. Furthermore, the underlying objective function would be hard to adapt to cater for requirement **H6**, the spatial contiguity of clusters. For those reasons, this naive approach is not considered further.

4.5.2 DBSCAN/OPTICS

DBSCAN, developed by [Ester et al., 1996; Xu et al., 1998], is a density-based clustering algorithm which may be used for spatial data. The basic idea for the algorithm is that for each point of a cluster the neighborhood of a given radius (ε) has to contain at least a minimum number of points (**MinPts**) where ε and **MinPts** are input parameters.

More precisely, DBSCAN proceeds to identify clusters via the difference in data density for in-cluster and out-of-cluster points. A distinction is made between a cluster's core points and border points with respect to the data density. At border points the data density is expected to be significantly lower than around core points. The authors of DBSCAN further elaborate upon concepts such as density-reachability which are required to formalize the assumption that clusters are dense while noise is not. The algorithm starts by randomly selecting a point from the data set and either growing a cluster from it or marking it as a border point. Noise points are determined as not belonging to any cluster.

DBSCAN requires only a few parameters to be set, which are not discussed here further in detail. The main reason that DBSCAN is not applicable as-is to precision agriculture data for management zone delineation purposes is that its underlying assumption is violated. The available spatial data points in this work do not exhibit dense and sparse regions in space. The points are sampled on a regular grid through geostatistical methods – density-based clustering approaches can not be used on this type of spatial data sets. The basic idea of dense clusters and sparse noise is, however, related to the image segmentation ideas behind region-growing approaches in Section 4.5.7.

Similar to the partitioning algorithm CLARANS presented in Section 4.5.5, the showcase for DBSCAN is depicted in Figure 4.9. If the data are spatially distributed with different densities, DBSCAN finds those clusters. This is not the case in the data sets which this thesis wraps around.

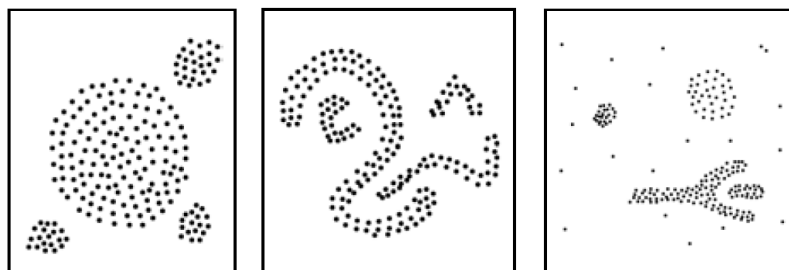


Figure 4.9: Spatial objects to be clustered, figure from [Ester et al., 1996].

Along similar lines, the OPTICS algorithm by [Ankerst et al., 1999] also intends to reveal hierarchical clustering structures from complex data sets. It does not create a multi-

level clustering directly, but rather computes an enhanced cluster ordering which can be used for setting a range of parameters to detect multi-level clusters. Nevertheless, the underlying assumption that clusters are spatially dense areas is also at the core of OPTICS which can therefore not be used here.

4.5.3 CLIQUE

Similar to DBSCAN, the CLIQUE algorithm devised by [Agrawal et al., 1998] is a hierarchical bottom-up clustering algorithm based on the assumption that dense regions in the data space are clusters. The authors approximate the density of the data points by partitioning the data space into subspaces of equal per-axis dimensions, i.e. equal volume. The density of each subspace can then be approximated easily by the number of data records in each volume. Afterwards, clusters of high density can be found by separating the regions according to the valleys of the density function. In this function, clusters are unions of connected high-density units within a subspace. Since CLIQUE is also based on the assumption that clusters are regions of higher data density than in the area around those regions, it can not be applied here.

4.5.4 STING

A further density-based approach is presented by [Wang et al., 1997], called STING (statistical information grid-based approach). The spatial area is divided into rectangular cells, where different sizes of such rectangular cells correspond to different resolutions and form a hierarchical structure. Each cell at a high level is partitioned to form a number of cells at the next lower level. For each of these levels and cells, statistical information is calculated and stored such that later clusterings can be constructed quickly. Figure 4.10 illustrates the process. Although STING originated in the area of spatial databases and its basic idea is rather intuitive, it can not easily be extended to cases where more than one variable, in addition to the spatial coordinates, is to be used. Furthermore, STING, like DBSCAN and CLIQUE, is density-based, rendering it unusable for spatial data which are spatially distributed at equal density.

4.5.5 CLARANS

A partitioning clustering algorithm which has been explicitly developed for clustering spatial data is CLARANS [Ng and Han, 2002]. It is based on randomized search in a graph-theoretic environment. Starting with n spatial objects, it aims to find k medoids which best describe the clustering. Here, a *medoid* O_m is a representative object for each cluster, meant to be the most centrally located object within the cluster. The randomized search is performed on a graph-based abstraction where the graph's nodes are represented by a set of k objects $\{O_{m_1}, \dots, O_{m_k}\}$. This intuitively indicates that the O_{m_1}, \dots, O_{m_k} are the selected medoids. The set of graph nodes is $\{\{O_{m_1}, \dots, O_{m_k}\} | O_{m_1}, \dots, O_{m_k} \text{ are objects in the data set}\}$. However, this graph is large, therefore a reduction of the search space is necessary. CLARANS achieves this by only examining a sample of a node's neighbors while proceeding through the graph. The algorithm is then further developed into clustering spatial objects other than spatial points, i.e. spatial polygons.

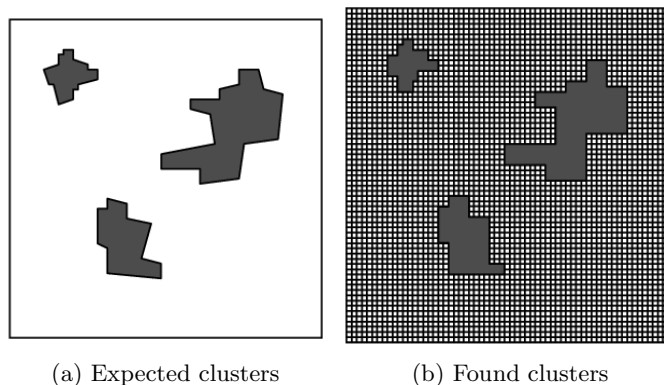


Figure 4.10: STING results for two-dimensional data: the left figure shows the expected result, while the right figure shows the grid and the result that STING supplies. The correct clusters are found, though they are slightly deformed due to the rectangular grid. Figure from [Wang et al., 1997].

Although CLARANS explicitly aims to deal with spatial data sets, the underlying assumption is that the structure to be discovered is hidden exclusively in the spatial information rather than or in addition to the non-spatial information. An example data set which is tackled successfully in an earlier version of CLARANS [Ng and Han, 1994] is shown in Figure 4.11. This very assumption renders CLARANS useless for management zone delineation: the spatial structure of the data is usually a regular grid with the points being equally distributed in the area under study. Therefore, CLARANS is unable to discover a meaningful structure using only the spatial part of the data sets.

4.5.6 AMOEBA

A multi-level clustering algorithm explicitly designed for exploratory analysis of geographical data has been described by [Estivill-Castro and Lee, 2000], called AMOEBA. It works hierarchically to find sets and subsets of spatial clusters for 2D points. The authors assume that point proximity in space is best captured using a Delaunay triangulation, rather than raster-based or vector-based proximity. While in raster-based proximity the neighborhood of a point in space is determined by the raster size, in vector-based proximity this neighborhood is based on the distance of two points (cp. Figure 4.12). The Voronoi tessellation on the other hand, as a dual of the Delaunay triangulation, represents spatial adjacency explicitly. If two points share a Voronoi edge, they are considered neighbors. The Delaunay triangulation can be constructed efficiently and is at the core of AMOEBA's clustering.

Having constructed the Delaunay triangulation, AMOEBA proceeds to look for short edges in the triangulation, since those determine nearby points which are likely to fall into a cluster. Different cutoff criteria for the edges' length account for different levels of the clustering and essentially work towards generating a hierarchy of clusters. This is depicted in Figure 4.13.

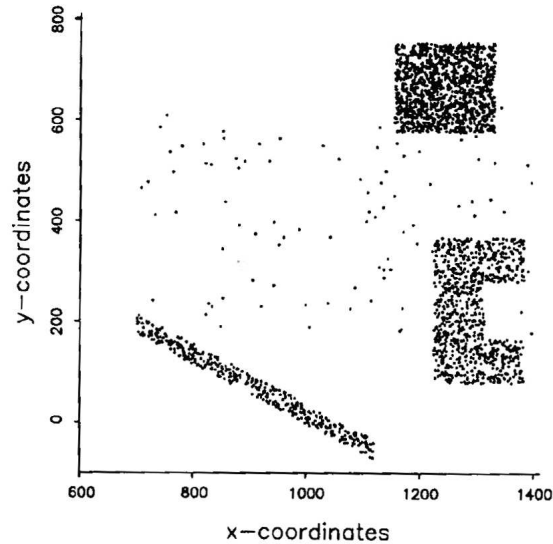


Figure 4.11: Spatial objects to be clustered, figure from [Ng and Han, 1994].

For the purpose of MZD, AMOEBA is not directly suitable. First, there are no variables attached to the 2D data records in AMOEBA and second, the underlying assumption is, as with the density-based algorithms in general, that the 2D points are non-uniformly distributed in space. Nevertheless, AMOEBA fulfills the requirement **S2** and would allow for an exploration of the solution space due to its hierarchical structure consisting of clusters and subclusters.

4.5.7 Region-Growing Approaches

Region Growing is a general image segmentation approach towards subdividing an image into regions which exhibit certain characteristics. It takes the input image and a set of seeds, which are randomly chosen or selected by the user. These seeds determine the objects which are to be segmented. Starting from the seeds, the regions are grown by comparing a seed's unallocated adjacent image pixels to the regions. The pixel with the smallest difference, according to some similarity measure, is allocated to the respective region. This continues until all pixels are allocated to a region. The similarity is typically determined between a pixel's intensity value and the region's mean intensity value. In the case of PA data sets, the pixels would be the spatial data records, though with more than one dimension to be considered by the algorithm.

While this basic segmentation method seems to be applicable to the management zone delineation problem encountered in this thesis, there are a few drawbacks. First, the segmentation results depend on the choice of the seeds. Different runs of this algorithm on precision agriculture data sets are likely to deliver zones which are inconsistent with each other, due to different seeds. An example is shown in Figure 4.14, where neighboring seeds lead to different region outcomes. Second, image segments are typically not joined, but

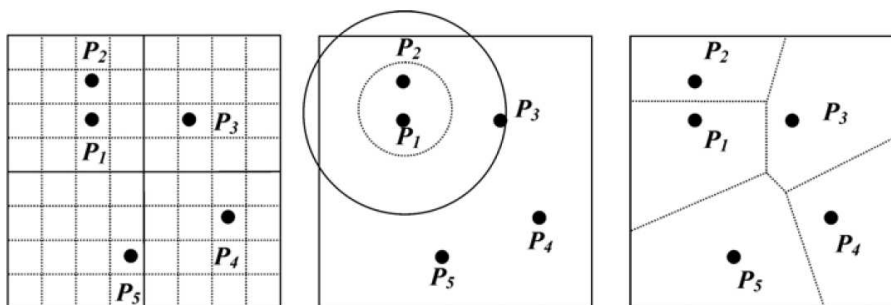


Figure 4.12: Spatial proximity basics for AMOEBA, figure from [Estivill-Castro and Lee, 2002]. The figure shows different types of proximity: raster-based (left), vector-based (middle) and Voronoi-based (right).

remain distinct different regions. Figure 4.14 illustrates this point: the two found segments are not very different from each other and may in practice actually describe one segment with similar properties in a precision agriculture application. Third, a management zone may consist of sub-zones which are *not* spatially adjacent (requirement **H6**) – this practically rules out the usage of standard region-growing approaches. Fourth, region growing is typically based on the assumption that there are edges in an image where a characteristic pixel property is very different on each side of an edge. This assumption does not hold true in general for PA data, since those are typically spatially autocorrelated and the assumed hard edges typically do not exist. For these reasons, region growing approaches are not considered further here.

4.5.8 Sweep-line Approach

In [Zalik and Zalik, 2009], a sweep-line-based approach to spatial clustering is presented. Clusters are generated by connecting those (spatial) points which are close enough, i.e. whose distance is below a threshold. This may lead to non-convex clusters which would be a desired effect in the MZD task. Finding these clusters is achieved by two consecutive sweep-lines which move into a fixed direction over the data points. The space between these sweep-lines is the actual working area where clusters are generated, while the space in front of the first sweep line is unclustered and the space behind the second sweep line is clustered.

Judging on the basis of [Zalik and Zalik, 2009], the experiments seem to be rather successful in clustering spatially distributed data points, although the details of the algorithm are rather unclear. However, the authors do not explicitly target data sets which are spatially uniformly distributed, although this is claimed in connection with new sensors. Satellite and aerial image-based sensors return data which is typically on a regular grid, such as the data for MZD. The authors claim that these data are “typically in the form of discrete points associated with additional scalar values” which is true but misses the important aspect about the spatial distribution of these points which their algorithm clearly relies upon. Therefore, although aimed at spatial analysis, this sweep-line approach

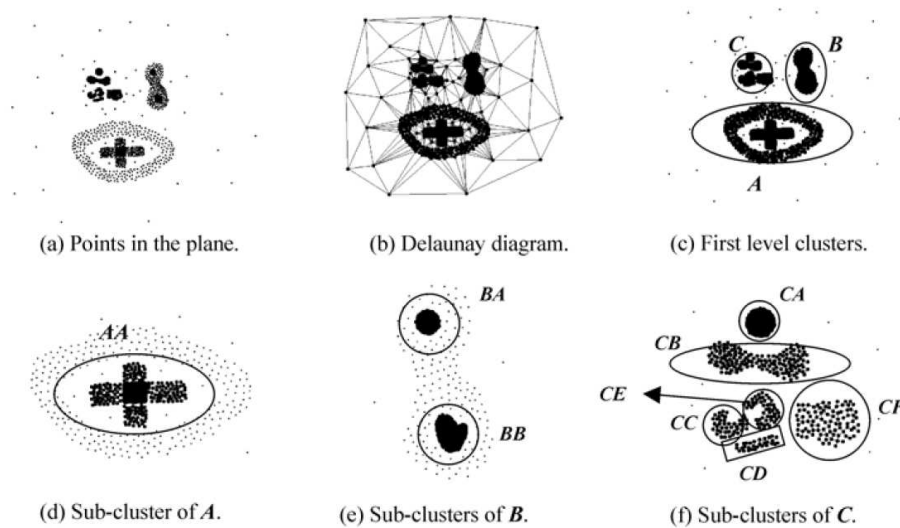


Figure 4.13: Spatial clustering with AMOEBA, figure from [Estivill-Castro and Lee, 2002]. The clustering algorithm works on spatial points in the plane (a), constructs a Delaunay diagram (b) and generates a hierarchy of different levels of clusters (c-f).

is not applicable for MZD. Nevertheless, the graph-based proximity of spatial points is an approach worth consideration in a MZD algorithm.

4.5.9 MOSAIC

MOSAIC, as presented in [Choo et al., 2007], has been explicitly developed to overcome the limitation of existing representative-based clustering algorithms to only report convex clusters. A further underlying assumption for MOSAIC is that, in theory, hierarchical clustering is capable of detecting clusters of arbitrary shape. These two approaches to clustering are joined to yield a hybrid representative-based hierarchical clustering algorithm. The resulting algorithm greedily merges neighboring clusters maximizing a given fitness function.

In the algorithm, the first step is to run a representative-based clustering algorithm to create a large number of clusters. As a second step, a proximity graph (a Gabriel Graph [Gabriel and Sokal, 1969]) is used to determine which clusters to merge in an iterative hierarchical approach. This graph is a more general variant of a Delaunay/Voronoi graph (shown e.g. in Figure 4.12) to determine point or cluster proximity. In this way, non-convex clusters are approximated as the union of small convex clusters which were obtained in the first step, hence the algorithm's name.

Without adaptations, MOSAIC is unfit for the particular spatial data sets encountered in the MZD task: it does not explicitly consider the difference between geographic space and feature space. However, in principle the hybrid approach reveals a few ideas which might be incorporated into a novel management zone delineation approach. First, using a representative-based algorithm (such as k -Means) to obtain an initial set of small clusters is

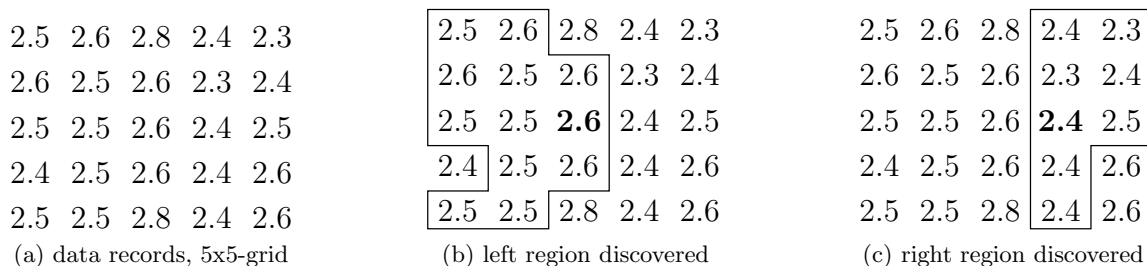


Figure 4.14: Region growing on PA data, illustrating two drawbacks: depending on the seeds (bold font in (b) and (c)), the algorithm finds different regions. Overall, the whole region, except for the high values at (1,3) and (5,3), is likely to be one practically relevant management zone.

applicable to the spatial part of the PA data sets. Since those data are uniformly distributed in geographic space, this would lead (for k -Means) to an initial Voronoi tessellation. Due to spatial autocorrelation, the points contained in such a Voronoi cell are likely to be rather similar in feature space. Second, subsequent merging could yield non-convex clusters in a hierarchy which would be explorable to a user.

4.5.10 ICEAGE

Further away from traditional spatial clustering and closer towards spatial clustering tailored to georeferenced data is ICEAGE (Interactive Clustering and Exploration of Large and High-Dimensional Geodata [Guo et al., 2003]). As an input to the algorithm, the spatial coordinates of the actual data points are used, and a Delaunay triangulation is constructed, similar to AMOEBA in Section 4.5.6. From this triangulation, the minimum spanning tree (MST) is constructed. A cluster in the MST is then viewed as a chain of points and can be hierarchically constructed: at the lowest level, each cluster/chain contains a single point. Each chain has two end points. When two clusters are merged into one with a new edge, this edge is positioned between the closest two ends of the two chains. This behavior can be seen in Figure 4.15. The found clusters are then explored further with density- and grid-based subspace clustering similar to CLIQUE (Section 4.5.3) to reveal sub-clusters and intrinsic structure in the spatial clusters. In an earlier publication ([Guo et al., 2002]) the clustering is also revealed to be exploratory and interactive.

Essentially, ICEAGE exploits the available spatial data sets in two steps: first, the geographic space is used for clustering, whereas in the second step, the feature space is further clustered to reveal interesting details. This is in principle applicable to the PA data sets for MZD since it uses both parts of the data (geographic space and feature space), although not simultaneously. However, the underlying assumption for the first clustering step is violated by the PA data sets: the data are spatially uniformly distributed which may lead to inconsistencies with the Delaunay triangulation (cocircularities) and, more importantly, does not reveal meaningful spatial clusters in the first place to explore further

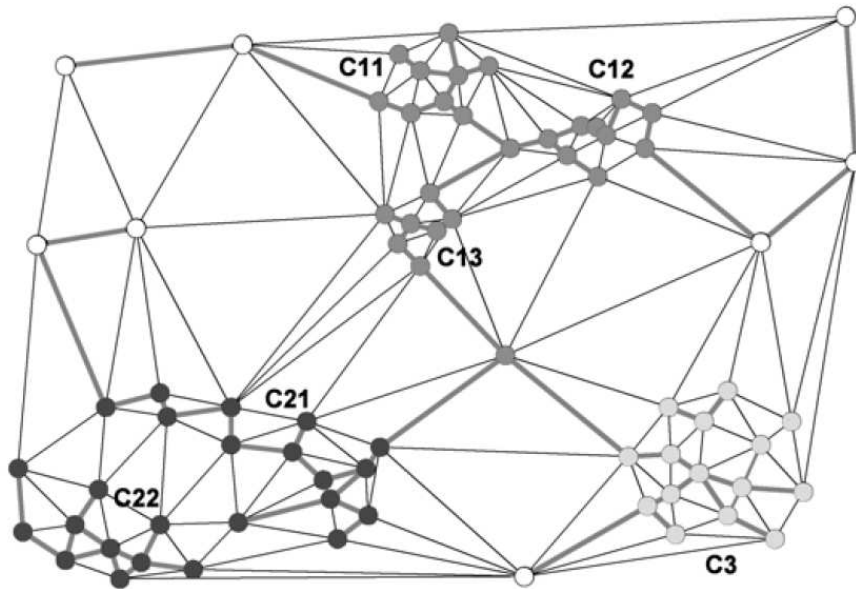


Figure 4.15: Clustering with ICEAGE, figure from [Guo et al., 2003]. The figure shows the Delaunay triangulation, the minimum spanning tree (thick edges) and hierarchical clusters. Boundary points (white) are removed by the algorithm and not included in any of the clusters.

during the course of the algorithm. However, the exploratory and interactive nature of the spatial clustering is pointed out.

4.5.11 SKATER

Spatial *K*luster Analysis by Tree Edge Removal (SKATER) was proposed by [Assuncao et al., 2006] to incorporate spatial contiguity constraints into regionalization². SKATER first builds a spatially contiguous graph from the data objects and removes edges which do not connect geographic neighbors. It then constructs a minimum spanning tree from this neighborhood graph. This tree is recursively heuristically partitioned to create a provided number of regions. Graph edges with high dissimilarity are usually pruned first to minimize an objective function.

The first problem with this approach is that it uses contiguity constraints in a static way: the contiguity matrix is not dynamically updated during the clustering process. Hence, spatial objects that are not spatial neighbors at the beginning may fail to be captured as being spatial neighbors during some later step, when both end up in adjacent clusters. Second, a minimum spanning tree may lead to *chaining* problems, which can be derived from Figure 4.16. Third, this approach is again based on the assumption that the spatial

²*Regionalization* in geographic information systems conforms conceptually to *management zone delineation* in precision agriculture and to *spatial clustering* in computer science.

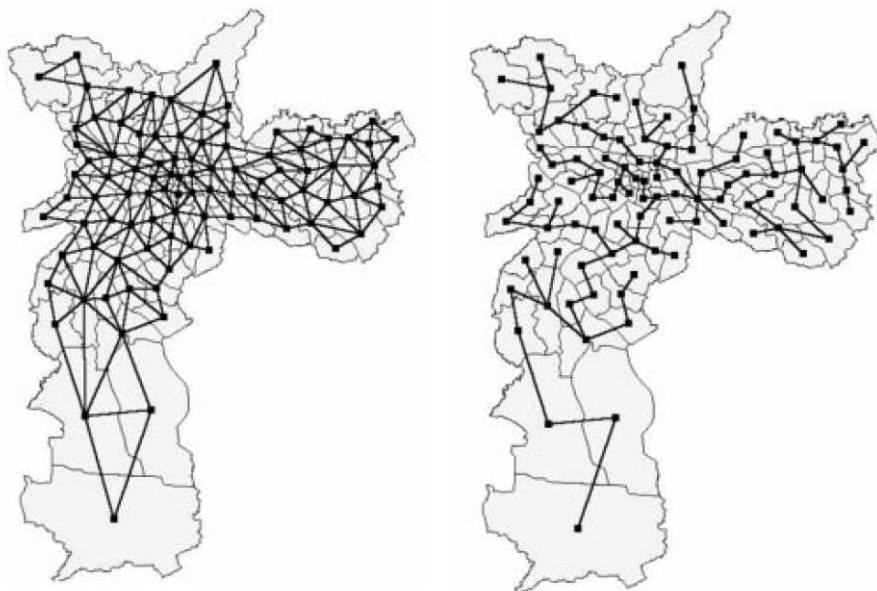


Figure 4.16: Regionalization with SKATER, figure from [Assuncao et al., 2006]. In the left figure, the connectivity graph is shown, while the right figure shows the minimum spanning tree (MST) based on the connectivity graph. This MST is later partitioned into spatially contiguous clusters by certain removing edges.

distribution of the data points shows dense and less dense regions. This is not the case in this study.

4.5.12 REDCAP

In [Guo, 2008] a regionalization approach based on agglomerative clustering with contiguity constraints is presented. The author aims to extend three commonly used hierarchical clustering methods: single linkage (SL), average linkage (AL) and complete linkage (CL) clustering. In addition, two different contiguity constraining strategies are introduced. Those are shown in Figure 4.17.

The combination of these three methods with the two constraints (full-order and first-order) leads to six contiguity-constrained agglomerative clustering methods. It is clearly stated that contiguity-constrained agglomerative clustering requires essentially that two clusters cannot be merged if their union is not spatially contiguous. Each of the available six methods works agglomeratively from single spatial data points and ends up with a spatially contiguous tree. This tree is then partitioned according to an objective function, which is to minimize the total heterogeneity value of all regions. The heterogeneity of a region is a measure of variable similarity among the spatial objects inside a region. Finally, the six methods are evaluated on a data set from the US presidential election in 2004, showing the different outcomes.

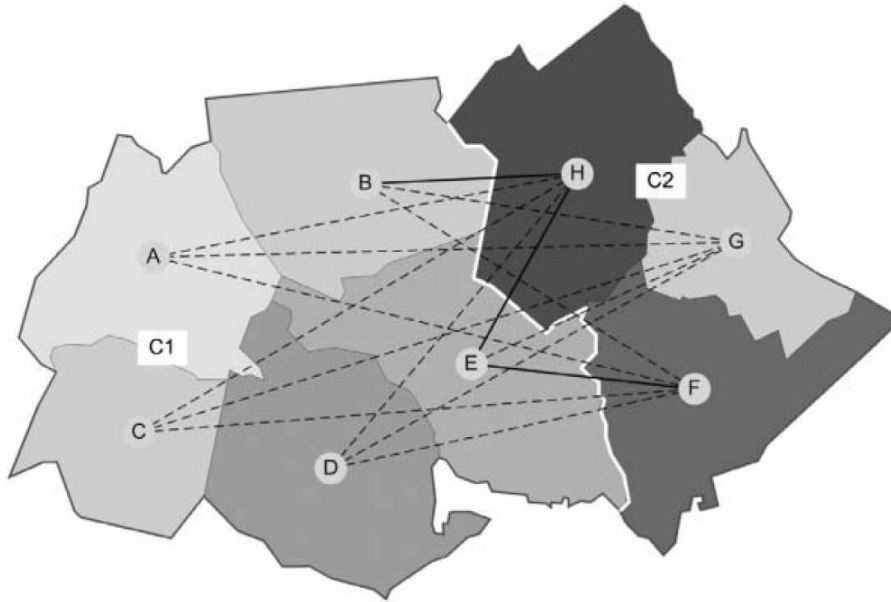


Figure 4.17: Regionalization with REDCAP, figure from [Guo, 2008]. The difference between first-order and full-order constraining strategies is shown. The eight counties are grouped into two clusters $C_1 = \{A, B, C, D, E\}$ and $C_2 = \{F, G, H\}$ according to their variable values (*not* according to spatial distances), signified by greyscale values. There are 15 edges connecting C_1 and C_2 . BH, EH, EF are first-order edges (connecting spatially adjacent clusters), while the full-order strategy would make use of all 15 edges.

Although this algorithm satisfies seven of the eight MZD requirements, it violates **H6**, which is special for management zone delineation. While REDCAP explicitly ensures the spatial contiguity of the resulting clusters, this is insufficient for MZD where one cluster may be split into a small number of sub-clusters which are spatially non-adjacent. This is not necessarily the case, but the algorithm should account for this occasion, which is not possible with REDCAP. Nevertheless, the clustering results of [Guo, 2008] clearly show that the outcome highly depends on the used method. At the same time, this illustrates that in spatial clustering the *perceived* quality of the result is user-dependent, rather than comparable to an optimal result. This should therefore lead to exploratory clustering approaches rather than confirmatory ones for MZD.

4.5.13 Summary of Spatial Clustering Algorithms

This section introduced a number of existing spatial clustering algorithms and described their specific drawbacks which make them unsuitable for the management zone delineation task encountered here. Although most of the algorithms are meant to work on spatial data, those are not typically the kind of spatial data encountered here: a set of uniformly distributed spatial data points with variable vectors attached to them. The algorithms are not usually able to handle this type of data consisting of a spatial and a non-spatial

part. However, even SKATER (Section 4.5.11) and REDCAP (Section 4.5.12), although in principle able and designed to handle this type of data, exhibit some drawbacks which should be addressed accordingly in order to design an exploratory management zone delineation approach. While the collection of high-resolution spatial data has seen an enormous increase in the past ten years, the methods to cope with this amount and type of data have only started to be developed in the past few years, judging by the presented relevant precision agriculture work (Section 4.3) and the spatial clustering work in this section.

Based on the requirements presented in Section 4.4 and the identified drawbacks of existing spatial clustering approaches in Section 4.5, a novel approach to exploratory spatial clustering designed for management zone delineation is presented in the following section.

4.6 Hierarchical Agglomerative Clustering with a Spatial Constraint (HACC-spatial)

This section shortly examines the main ideas from spatial clustering and provides the reasoning for developing a novel approach. Hierarchical clustering is briefly introduced before being extended to constraint-based clustering. The novel algorithm HACC-SPATIAL (Hierarchical Agglomerative Constrained Clustering with Spatial Contiguity) is described in detail before results are presented in Section 4.7.

The clustering approaches in Section 4.5 typically borrow ideas from more than one of the mentioned four groups: *hierarchical*, *partitioning*, *density-based* and *grid-based*. Nevertheless, their underlying assumptions fall into only one of those categories. In the case of PA spatial data sets, three of the four categories can be ruled out:

density-based Since density-based algorithms (DBSCAN, OPTICS, etc.) assume that clusters are dense regions whereas noise is not dense, they can not be applied to the PA data since the data are spatially uniformly distributed, without density differences. Ruling out those approaches satisfies requirement **H5**.

grid-based Those approaches define a grid on the data set to subdivide it into small cells. Those small cells are then analyzed for data density and dense regions are expected to be clusters. As above, the spatial data density in the PA data is mostly constant. A grid overlaid on the spatial data would also have to be adapted for each data set and likely irregularities, like field borders or (real) holes in the data set due to power poles, sheds etc.

partitioning The main representatives of this category are *k*-Means and fuzzy *c*-Means, which have been shown to be not applicable to the PA data sets. A naive extension, including the spatial points' coordinates as variables (Section 4.5.1), also fails to grasp the structure in the data sets. Furthermore, partitioning approaches typically find convex clusters, which is not necessarily the desired result (in geographic space) here.

Therefore, the remaining category contains hierarchical approaches. The basic choice then is whether to choose an agglomerative (bottom-up) or divisive (top-down) approach.

While an agglomerative approach starts with each spatial data point as its own cluster and consecutively merges points/clusters into larger clusters, a divisive approach considers the complete data set as one cluster and consecutively splits it into subclusters. In both cases, a tree of clusters and subclusters, a *dendrogram*, results which can be explored by the user, satisfying requirements **H1** and **H2**. Furthermore, obtaining different numbers of clusters does not require re-running the algorithm, satisfying requirement **H4**.

Similar to the above exclusion of partitioning approaches, the problem with divisive approaches is whether the data should be split according to geographic or feature space or both. Splits in feature space are not typically the same as in geographic space. On the other hand, agglomerative approaches can in principle consider merging in feature space, but only merge points or clusters which are adjacent in geographic space. Therefore, the focus is on an agglomerative approach here. The only requirements which have to be fulfilled are those pertaining to the spatial proximity in the data (**H3**) and the spatial contiguity of the clusters (**H6**). Hierarchical clustering would further fulfill requirement **S2**: explaining a hierarchical algorithm to an average precision agriculture expert is assumed to be easily understandable and therefore more acceptable in practice. ³

With the growing amounts of georeferenced and personalized data, standard clustering algorithms from the above groups have recently been extended by so-called constraints to cater for the demand. For spatial clustering, these constraints are briefly mentioned in [Han et al., 2001] and in two more recent surveys [Zeitouni, 2002; Kotsiantis and Pintelas, 2004]. If the spatial proximity and the spatial contiguity were considered as constraints during the clustering process, hierarchical agglomerative clustering with those constraints would possibly fulfill the owing requirements above. Therefore, the following section further explores the concept of introducing spatial constraints into hierarchical clustering.

4.6.1 Literature on Hierarchical Constrained Clustering

Clustering with constraints dates back to at least 2001 ([Wagstaff et al., 2001; Yang et al., 2001]), and if the related topic of *regionalization* is considered, the first approaches were developed in the 1970s. Especially the contiguity constraint (spatial or not) has seen a few efforts, as surveyed by [Gordon, 1996]. For the purpose of regionalization of British counties, [Spence, 1968] proposes an algorithm for the delimitation of regions and regards this as a multivariate classification problem. The principal components of the underlying data sets are computed and the data are grouped by Ward's grouping procedure. However, the data in this study are not distributed with uniform spatial density, which rules out this approach. More towards the agricultural area, soil-mapping has been handled as a constrained classification problem by [Webster and Burrough, 1972b,a], also first using an ordination step, which is basically a principal components analysis. A similar study on Norwegian administrative regions is performed by [Byfuglien and Nordgård, 1973], and can be ruled out for the same reason. The comparative study for regional taxonomy presented by [Fischer, 1980] identifies the main linkage criteria for hierarchical agglomerative regionalization strategies and also identifies the spatial contiguity constraint to be important for the regionaliza-

³personal experience: HACC-SPATIAL could be explained in an average conference talk at the *International Conference on Precision Agriculture, Denver, July 2010* and subsequent short invited talks at a similar audience. The feedback from the audience confirmed that the main ideas had been understood.

tion. An early version of *constrained agglomerative hierarchical classification* applied to soil sampling data is presented in [Perruchet, 1983]. In the work of [Margules et al., 1985], the contiguity constraint is termed an adjacency constraint. More recent work mainly reiterates these basic concepts, such as [Davidson and Ravi, 2005; Morales and Mendizabal, 2010], albeit typically with a few improvements. However, the idea of using contiguity constraints in clustering is typically reiterated.

Each of [Webster and Burrough, 1972a; Perruchet, 1983; Margules et al., 1985] contains a similar definition or notion of contiguity for the zones, regions or classes which are to be generated. [Webster and Burrough, 1972a] define it as follows:

[...] Although the resulting classes will usually form compact regions, there can still be fragmentary inliers and outliers. To avoid fragmentation completely, location can be used as a ‘contiguity constraint’. Similarities between pairs of individuals are calculated as usual without regard to location. Agglomeration then proceeds by fusing geographically contiguous similar individuals or groups. The resulting regions will often be more compact than if no contiguity constraint were applied.

However, the above idea has been incorporated by using the geographic location of the data records as another variable in the feature space. This is not desirable here for reasons given in Section 4.5.1. In the light of his algorithm CAHC, [Perruchet, 1983] describes the contiguity and the constraint as follows:

... The Constrained Agglomerative Hierarchical Classification [CAHC] is applicable to all the data sets represented in two distinct spaces. The first one, called descriptor space, is the one where the usual analysis is done. The second one, called constraint space, is specified by the introduction of a relation of contiguity used as a constraint during the classification. The basic idea is to favour the pair of clusters which are structurally close (i.e. contiguous) during the aggregations. In practice, the hierarchy depends on the choice of a contiguity threshold, defined as the maximal distance beyond which two points are not contiguous and hence cannot be aggregated.

However, the author uses the constraint as a local threshold rather than a global one, which is fine for the data sets under study, but can not be easily carried over to the spatially uniform and dense data sets in precision agriculture. Furthermore, CAHC does not allow for relaxing the contiguity constraint, hence the generated zones are strictly contiguous. Building on the above ideas, [Margules et al., 1985] reformulate the contiguity as an adjacency constraint as follows:

[...] a method is described for agglomerative hierarchical numerical classifications of geographic data, which includes location as an extrinsic character. The idea is to apply an adjacency constraint to a classification, allowing only those objects (in this case, land units of some kind) adjacent to other objects or groups of objects to join them. Adjacency may be defined in any appropriate way. [...] If the level of heterogeneity in groups that result from a classification

constrained by adjacency is too high, a compromise may be required, trading off complete contiguity for greater homogeneity. [...] one way to get a compromise classification might be to assign a threshold level of stress below which the adjacency constraint would not apply.

Hence, the author notes the shortcoming of the strict contiguity and proposes to relax it using a tradeoff parameter based on a classification stress value. However, this work does not seem to have been continued.

For constraints-based clustering in general, the work of [Wagstaff, 2002] comprises a thorough review of existing algorithms, parts of which might be applicable here. The author explicitly describes the *spatial contiguity* constraint for spatial data as a type of global clustering constraint using neighborhood information, albeit for image segmentation and later for GPS trace mining [Schroedl et al., 2004]. The constraints are presented as being *hard* or *soft*, meaning that the final clustering outcome *must* or *can* consider these constraints. An additional feature of constrained clustering algorithms is the existence of *must-link* and *cannot-link* pairwise constraints for data records. Furthermore, the work of [Wagstaff et al., 2005] encounters a similar agricultural problem to the one in this thesis, but the focus is shifted towards yield prediction on a county scale with low-resolution data, rather than using high-resolution data for management zone delineation.

[Klein et al., 2002] can be seen in close conjunction with the work of [Wagstaff, 2002]: while the latter introduces constraints into the k-Means and the COBWEB incremental conceptual clustering algorithm [Fisher, 1987], the former sets the focus on constrained hierarchical complete-linkage clustering. The three recent algorithms have been compared in [Klein et al., 2002], with the result that it depends on the data set and the clustering target whether one algorithm outperforms the other. However, when it comes to explaining an algorithm to an average agriculture expert, a hierarchical agglomerative approach has the certain advantage of being much more straightforward to understand and therefore more acceptable in practice.

4.6.2 HACC-spatial

Based on the three similar definitions of contiguity-constrained clustering above, this section develops a novel algorithm which follows the main ideas from those definitions. Therefore, the task encountered in this chapter, namely generating *mostly contiguous* clusters for management zone delineation, is tackled by using a spatial contiguity constraint in a hierarchical clustering approach.

In Section 4.3, one of the requirements that existing management zone delineation approaches aim to fulfill is the contiguity of the zones. This can not be guaranteed by classical clustering algorithms which are based on the data records' variables only while neglecting the spatial component. Furthermore, one zone may still be spatially split into two or more parts, but should not be scattered too much further among the field. In the context of [Wagstaff, 2002], this can be treated as a user-dependent global constraint in this exploratory clustering task. Given a hierarchical agglomerative clustering setup, the constraint is of the "cannot-link" type: clusters which are not spatially adjacent can not be merged. It is furthermore likely to be a hard constraint at the beginning of the algorithm,

while switching off the constraint at a later stage should serve the “mostly contiguous” requirement. This issue can be resolved by introducing an additional tradeoff parameter (as mentioned in [Margules et al., 1985]), which is described later.

The proposed approach HACC-SPATIAL embraces the main idea from hierarchical agglomerative clustering, starting with single data records and subsequently merging these according to the data records’ similarity. However, not every cluster pair which is similar can be merged: the constraint which assures spatial contiguity of the clusters is taken into account in the selection of the pair to merge. To account for the fact that zones need not be strictly contiguous, this constraint is relaxed after a certain tradeoff threshold has been reached.

HACC-SPATIAL consists of a main phase which takes care of the hierarchical agglomerative clustering including the constraint. Due to the specialties of the precision agriculture data sets, an optional pre-tessellation phase can be performed before the actual clustering. Since the data are typically spatially autocorrelated, adjacent data records may initially be grouped purely according to their location rather than their variables’ values. This latter phase is described first.

Optional Field Tessellation

Having decided for a hierarchical agglomerative clustering approach for spatially autocorrelated data, the possibility of using an initial tessellation of the data according to their spatial coordinates exists. Due to spatial autocorrelation, spatially adjacent data records are likely to be very similar in their variables. Therefore, by tessellating the field into a fixed number of spatial clusters $n \leq N$ (where N is the number of data records), the clusters are still very likely to contain similar (adjacent) data records while a lot of the ensuing computational effort of the hierarchical agglomerative merging step can be saved.

With the above prerequisites, at least two simple tessellation approaches fulfill the requirements. The first is a simple grid-based approach, i.e. cutting the site into square or hexagonal areas of an appropriate size. Since the sites are not necessarily rectangular and often have irregularities, this is likely to generate a few artifacts, but since those initial clusters are then merged in the main phase of HACC-SPATIAL, this issue is unlikely to lead to further problems. The second approach is to perform a k -means clustering on the data records’ spatial coordinates (cp. Section 3.2.4 on Page 35). This creates a basic tessellation, while explicitly assuming that, due to spatial autocorrelation, the resulting spatial clusters contain similar data records. Furthermore, the k -means tessellation returns a voronoi diagram of the data records’ coordinates, of which the dual representation is the Delaunay triangulation. This allows for easy computation of the list of neighbors for each cluster [Gold and Remmele, 1997]. A depiction of a tessellation step for the F550 data set is given in Figure 4.18.

Main Phase: Hierarchical Agglomerative Clustering with Constraint

The main phase of HACC-SPATIAL starts with clusters, which either consist of one data record each or, alternatively, are groups of similar adjacent records which are optionally generated on the assumption of spatial autocorrelation. The task now is to merge these

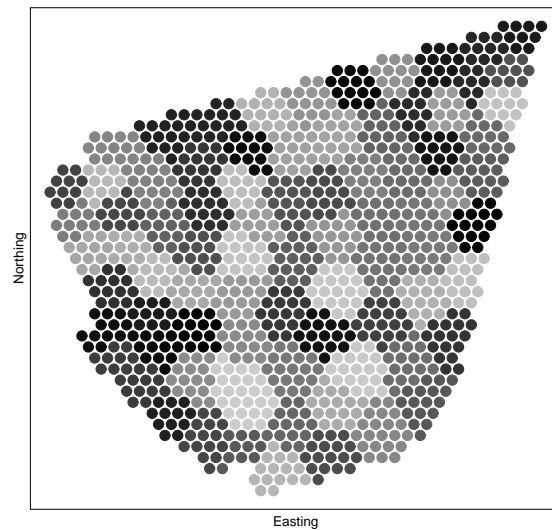


Figure 4.18: F550, tessellation of field into 80 zones. Grey shades are for showing the clusters only, they do not pertain to any particular variable.

clusters consecutively into larger clusters. However, in addition to any chosen similarity or distance measure, a spatial constraint must be taken into account. Since the final result of the clustering is assumed to be a set of spatially mostly contiguous clusters, only those clusters should be merged which are a) similar (with regard to their variables' values) and b) spatial neighbors (adjacent).

Cluster Similarity / Cluster Distance In classical hierarchical clustering, the standard measures for cluster similarity are single linkage, complete linkage and average linkage [Jain et al., 1999]. However, when extending the clustering to the spatial data encountered here, these criteria merit some explanation:

single linkage determines cluster similarity based on the smallest pairwise distance between all objects from the clusters. For adjacent clusters, due to spatial autocorrelation, it is likely that there are always a few points at the borders of the clusters which are very similar, for each neighbor (cp. Figure 4.19). This would lead to a chaining effect, since clusters adjacent in geographic space would always be considered similar in feature space. Therefore, single linkage does not provide an appropriate measure for which neighbor to choose for merging.

complete linkage determines the similarity of clusters based on the pairwise distance of those objects which are farthest away from each other in feature space. For adjacent clusters, due to spatial autocorrelation, these objects are also likely to be rather far away from each other in geospace and not representative for a cluster. This would again lead to a chaining effect and less meaningful clusters when considering a spatial constraint.

average linkage determines the similarity of clusters based on the average of the pairwise distances between all objects in the clusters. For geographically adjacent clusters, this measure is expected to provide a sufficient similarity criterion. According to [Manning et al., 2008], the computational effort of average linkage is quadratic ($\mathcal{O}(n^2 \cdot \log n \cdot V)$, where V is the number of variables), since a distance matrix containing pairwise distances is typically used and updated throughout the algorithm.

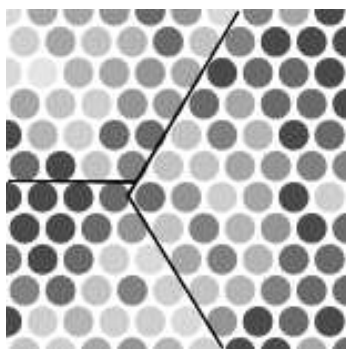


Figure 4.19: For the single linkage criterion, the right and the bottom left cluster would be considered very similar since the points with minimum distance in feature space (similar grey shades) would determine the overall cluster distance. Due to spatial autocorrelation, geographically adjacent data points, separated by a cluster border, are likely to be very similar, represented by similar colors. The same would hold for the other possible cluster pairs. This figure is a selected area from Figure A.4c on page 178 (EC25 variable) to illustrate the single-linkage issue.

Spatial Contiguity Constraint Since it is not required that one zone is strictly contiguous, i.e. consists of just one spatially contiguous area on the field, it is clearly valid if one zone comprises those data records which are similar but is made up of two or more larger areas on the field. This would still be considered immensely useful in practice, since the focus of this clustering approach is on exploratory clustering rather than providing a fixed clustering.

At the beginning of the hierarchical agglomerative clustering the spatial contiguity constraint ensures that only spatially adjacent clusters are merged. During the course of the algorithm, however, a pair of clusters which are not spatially adjacent may become more similar in feature space than any other pair of adjacent clusters. Therefore, the constraint should be switched off from that point onwards. Naturally, an arbitrary step during the clustering can be set for switching off the spatial constraint, e.g. after half of the overall merging steps. Nevertheless, the aforementioned effect of the similarity ratio between adjacent and non-adjacent clusters can be used to decide heuristically when to turn off the spatial contiguity constraint. This concept of spatial contiguity is related to the *spatial diversity* provided by [Li and Claramunt, 2006], notwithstanding the authors' focus on classification in GIS rather than spatial clustering.

It is expected that the feature space distances between adjacent clusters as well as between spatially non-adjacent clusters increase during the course of the algorithm. To make the aforementioned decision (switching off the constraint), a stable criterion is needed. For average linkage, maximum and minimum distances typically vary greatly before and after a cluster merging step. For average linkage, the sequence of average distances between the clusters throughout the clustering process forms a rather stable criterion (cp. Figures 4.20 and 4.21) upon which to base the decision. The *contiguity ratio* is introduced in Equation 4.1 as the ratio between the mean distances of spatially adjacent and spatially non-adjacent clusters. The comparison between the two extreme cases is provided in Figures 4.20 and 4.21.

$$contiguity_{spatial} = \frac{meanDistance_{adjacent}}{meanDistance_{non-adjacent}} \quad (4.1)$$

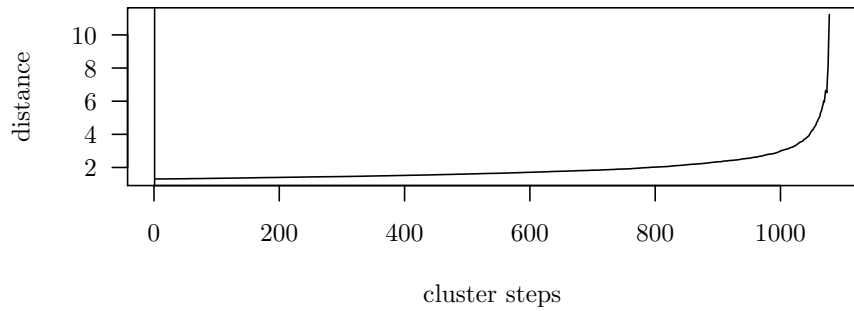
Allowing the user to influence the spatial contiguity can now be accommodated for by a *contiguity ratio threshold*. As soon as the contiguity ratio reaches this threshold, the spatial contiguity constraint is switched off and HACC-SPATIAL proceeds in the same way as an unconstrained, traditional hierarchical agglomerative clustering algorithm.

Algorithmic Description The algorithm's formalization is provided in Algorithm 5. A few annotations regarding the implementation are listed below. The implementation has been carried out in R [R Development Core Team, 2009], therefore the algorithmic notation is kept close to R's syntax.

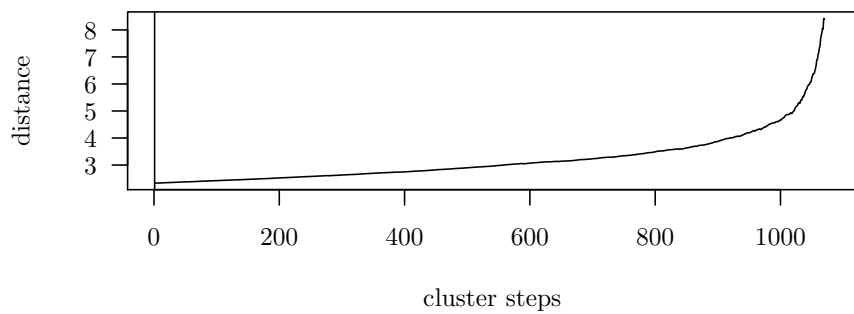
distance measure For distance calculations ($dist(c_i, c_j)$), currently Euclidean distance is used. Since the number of variables in the clustering is rather small, this is sufficient. For higher numbers of variables, the Cosine or other distance measures may be employed. For further discussion on choosing an appropriate distance measure, see, e.g. [Weihs and Szepannek, 2009]. Based on the chosen distance measure, a weighting of variables may optionally be applied.

distance matrices Since the chosen distance measure is symmetric, i.e. $dist(c_i, c_j) = dist(c_j, c_i)$, the $dist_a$ and $dist_{\bar{a}}$ matrices which store the calculated distances between the cluster representatives of adjacent and non-adjacent clusters are diagonal matrices where the cells are by default set to NA.

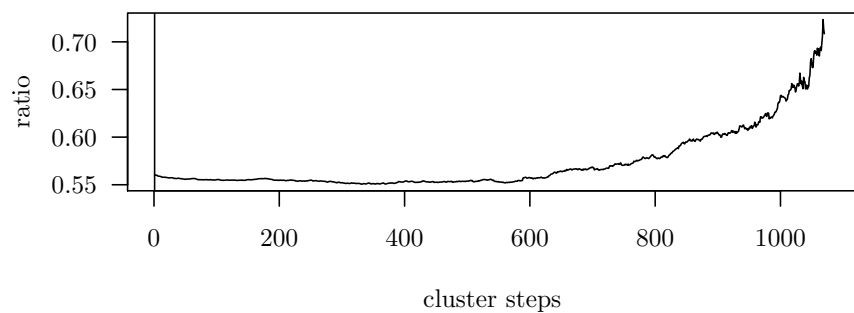
storing merging information During the merging process, each newly merged cluster retains the information from which clusters it has been created. Therefore each merged cluster retains a dendrogram of its sub-clusters.



(a) Mean Distance of Spatially Non-Adjacent Clusters

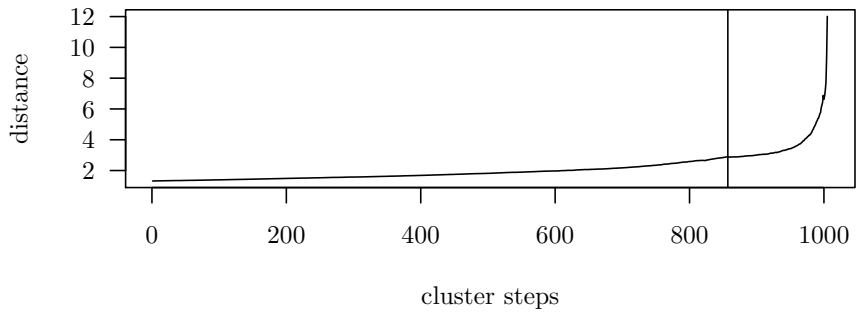


(b) Mean Distance of Spatially Adjacent Clusters

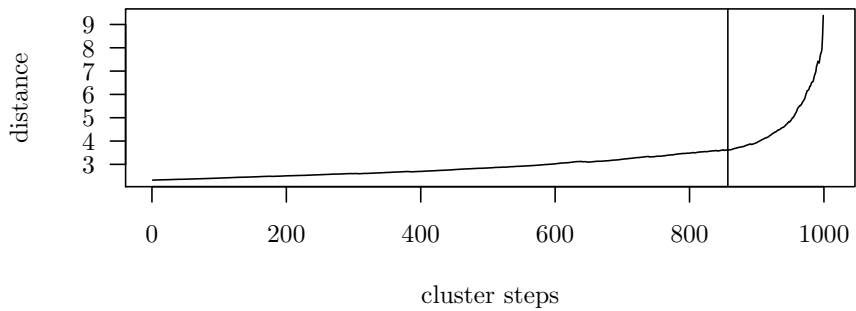


(c) Contiguity Ratio

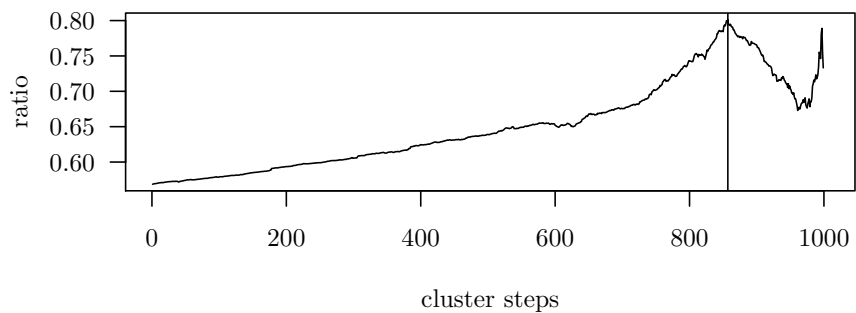
Figure 4.20: Mean distances during clustering, data set F550 (P, K, MG, PH), contiguity threshold exceeded at first step, therefore the clustering proceeds without any spatial contiguity assurance and according to standard hierarchical agglomerative clustering. Two figures from this clustering are shown in Figures 4.27a and 4.27c.



(a) Mean Distance of Spatially Non-Adjacent Clusters



(b) Mean Distance of Spatially Adjacent Clusters



(c) Contiguity Ratio

Figure 4.21: Mean distances during clustering, data set F550 (P, K, MG, PH), contiguity threshold exceeded at step 860. Until this step, only the most similar geographically adjacent clusters are merged, while after this step the merging decision is based on the feature space distance only. Two figures from this clustering are shown in Figures 4.27b and 4.27d.

Algorithm 5 HACC-SPATIAL

```
# input:
#   V ... set of  $i$  georeferenced data vectors
#    $k$  – tessellation resolution,  $k \leq i$ 
#   cp – contiguity constraint parameter
5: #    $\text{dist}_a, \text{dist}_{\bar{a}}$  – distance matrices holding average distances
#   between adjacent/non-adjacent clusters
# output: a dendrogram of the hierarchical clustering

# split phase, run  $k$ -means clustering on spatial locations of data vectors
10:  $C \leftarrow k\text{-means}(V, k)$ 
return spatial clustering  $C$ 
# merging phase, iteratively merge clusters according to cp
spatialconstraint  $\leftarrow$  TRUE
repeat
15: # determine and store cluster distances
for each spatially adjacent cluster pair  $(c_i, c_j) \in C$  do
     $\text{dist}_a[\text{i,j}] \leftarrow \text{dist}(c_i, c_j)$ 
end for
for each spatially non-adjacent cluster pair  $(c_i, c_j) \in C$  do
20:     $\text{dist}_{\bar{a}}[\text{i,j}] \leftarrow \text{dist}(c_i, c_j)$ 
end for
# determine minimum/median distances and contiguity
 $\text{mindist}_a \leftarrow \min(\text{dist}_a)$ ,  $\text{mindist}_{\bar{a}} \leftarrow \min(\text{dist}_{\bar{a}})$ 
 $\text{contiguity} \leftarrow \frac{\text{mean}(\text{dist}_a)}{\text{mean}(\text{dist}_{\bar{a}})}$ 
25: # switch off constraint when cp is reached
if  $\text{contiguity} \geq \text{cp}$  and spatialconstraint then
    spatialconstraint  $\leftarrow$  FALSE
end if
if spatialconstraint then
30:    clusterpair  $\leftarrow$  which( $\text{dist}_a == \text{mindist}_a$ , arr.ind=TRUE)
else
    if  $\text{mindist}_a \leq \text{mindist}_{\bar{a}}$  then
        clusterpair  $\leftarrow$  which( $\text{dist}_a == \text{mindist}_a$ , arr.ind=TRUE)
    else
35:        clusterpair  $\leftarrow$  which( $\text{dist}_{\bar{a}} == \text{mindist}_{\bar{a}}$ , arr.ind=TRUE)
    end if
end if
     $i \leftarrow \text{clusterpair}[1]$ ,  $j \leftarrow \text{clusterpair}[2]$ 
     $C \leftarrow C \setminus (c_i, c_j)$  # remove most similar cluster pair
40:  $C \leftarrow C \cup (c_i \cup c_j)$  # add newly merged cluster
    update:  $\text{dist}_a, \text{dist}_{\bar{a}}$ 
until number of clusters = 1
return dendrogram of management zones  $C$ 
```

4.7 Evaluation of HACCSpatial on PA Data Sets

The algorithm presented in Section 4.6 aims to incorporate a spatial contiguity constraint into hierarchical agglomerative clustering. Since it has primarily been developed for the purpose of management zone delineation, this purpose is further elaborated upon in this section. Results from the data sets under study are presented and underline the algorithm’s exploratory clustering nature. In particular, the focus is on the user-influencable spatial contiguity threshold and the initial tessellation.

4.7.1 Experimental Setup

For practical tasks, the EC25 variable is typically used when delineating basic fertilization zones. If soil-sampling data (P, K, MG, PH) are available, such as in the F550 data set, those may be used as well. For other purposes, vegetation-based zones can be generated using the vegetation indices REIP32, REIP49. Yield-based management zones can be generated using past site-years of YIELD. To demonstrate the influence of the spatial contiguity threshold and the initial tessellation, the experimental setting is laid out in Table 4.1.

data set	variables	contiguity threshold	tessellation	Section	Figure	Page
F611	EC25	{0.5, 1.0}	–	4.7.2	4.22	116
F440	REIP49	{0.3, 0.5, 1.0, 2.0}	–	4.7.2	4.23	117
F631	EC25	{0.3, 0.5, 0.9, 1.0}	–	4.7.2	4.24	118
F440	YIELD07	{0.5, 1.0}	–	4.7.2	4.25	119
F611	REIP32,REIP49	{0.5, 1.0}	–	4.7.3	4.26	121
F550	P, MG, K, PH	{0.5, 0.8}	–	4.7.3	4.27	122
F610	EC25	{0.5, 1.0}	–	4.7.4	4.28	123
F611	EC25	{0.5, 1.0}	200	4.7.5	4.29	125
F440	REIP49	{0.5, 1.0}	200	4.7.5	4.30	126
F631	EC25	{0.5, 1.0}	250	4.7.5	4.31	127
F440	REIP32	{0.5, 0.7, 0.9, 1.0}	120	4.7.5	4.32	128
F440	N1	1.0	{100, 640}	4.7.5	4.33	130
F611	REIP32,REIP49	1.0	200	4.7.6	4.34	131
F550	P, MG, K, PH	1.0	200	4.7.6	4.36	133
F610	EC25	1.0	100	4.7.7	4.37	135

Table 4.1: Overview of HACCSpatial experiments, each line describing the data set under study and the parameters (contiguity threshold and tessellation) varied. The upper part features experiments without the initial tessellation, while the lower part examines the effect of the initial tessellation. Each part starts with experiments for single variables and proceeds towards using multiple variables later. In the end of each part, HACCSpatial is examined for incomplete data sets such as those from the F610 site.

4.7.2 Zones without Initial Tessellation, One Clustering Variable

Starting with F611 and the variable EC25 in Figure 4.22, the effect of the contiguity threshold ct can be shown. While at a low setting of $ct = 0.5$ the spatial constraint is switched off after 3994 of 4969 steps, at a high setting of $ct = 1.0$ it is switched off at 4870 of 4969 steps and clearly generating much more contiguous zones.

The second comparison, using F440 and the REIP49 variable in Figure 4.23, has a larger spread of $ct \in \{0.3, 0.5, 1.0, 2.0\}$, thereby making HACC-SPATIAL act like an unconstrained hierarchical agglomerative clustering for $ct = 0.3$ and, on the other end of the range, having it run with the spatial contiguity constraint throughout the course of the algorithm. While the extreme cases show the algorithm's ability to run in those settings, the outcome is not immediately useful in the sense that it enhances the knowledge about the F440 site. The results for $ct \in \{0.5, 1.0\}$ exhibit much clearer zones that correspond to the actual REIP49 variable for a vegetation-based management zone.

Again generating an EC25-based management zone for the F631 site in Figure 4.24, the different settings of $ct \in \{0.3, 0.5, 0.9, 1.0\}$ create slightly different clusterings while still revealing underlying spatial structure. Again, the extreme cases for the contiguity threshold are not the most immediately useful in terms of helping in understanding the site.

For yield-based management zones, Figure 4.25 provides an example for F440 using YIELD07 with $ct \in \{0.5, 1.0\}$. Both results expose the site's yield circumstances, however, the low and high spatial contiguity settings clearly have an effect on the zones. While $ct = 0.5$ produces roughly two large visible zones, the setting of $ct = 1.0$ shows much more scattered zones. On this site, both results visually correspond rather well to the YIELD07 variable. There is a clear boundary between the northern and the southern part of the site which has been discovered by both versions.

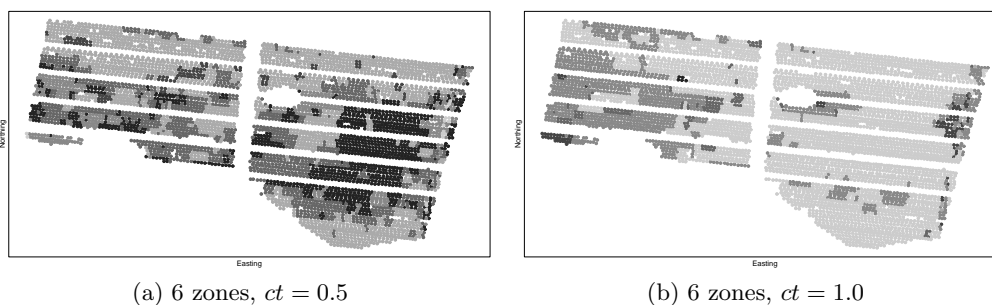


Figure 4.22: HACC-SPATIAL on F611, using the variable EC25 (cp. Figure A.10a on Page 186). The figures illustrate the influence of the contiguity parameter. While the left figure ($ct = 0.5$) shows six rather scattered zones, the right figure ($ct = 1.0$) exhibits a clear spatial structure among those six zones and enhances the understanding of the site.

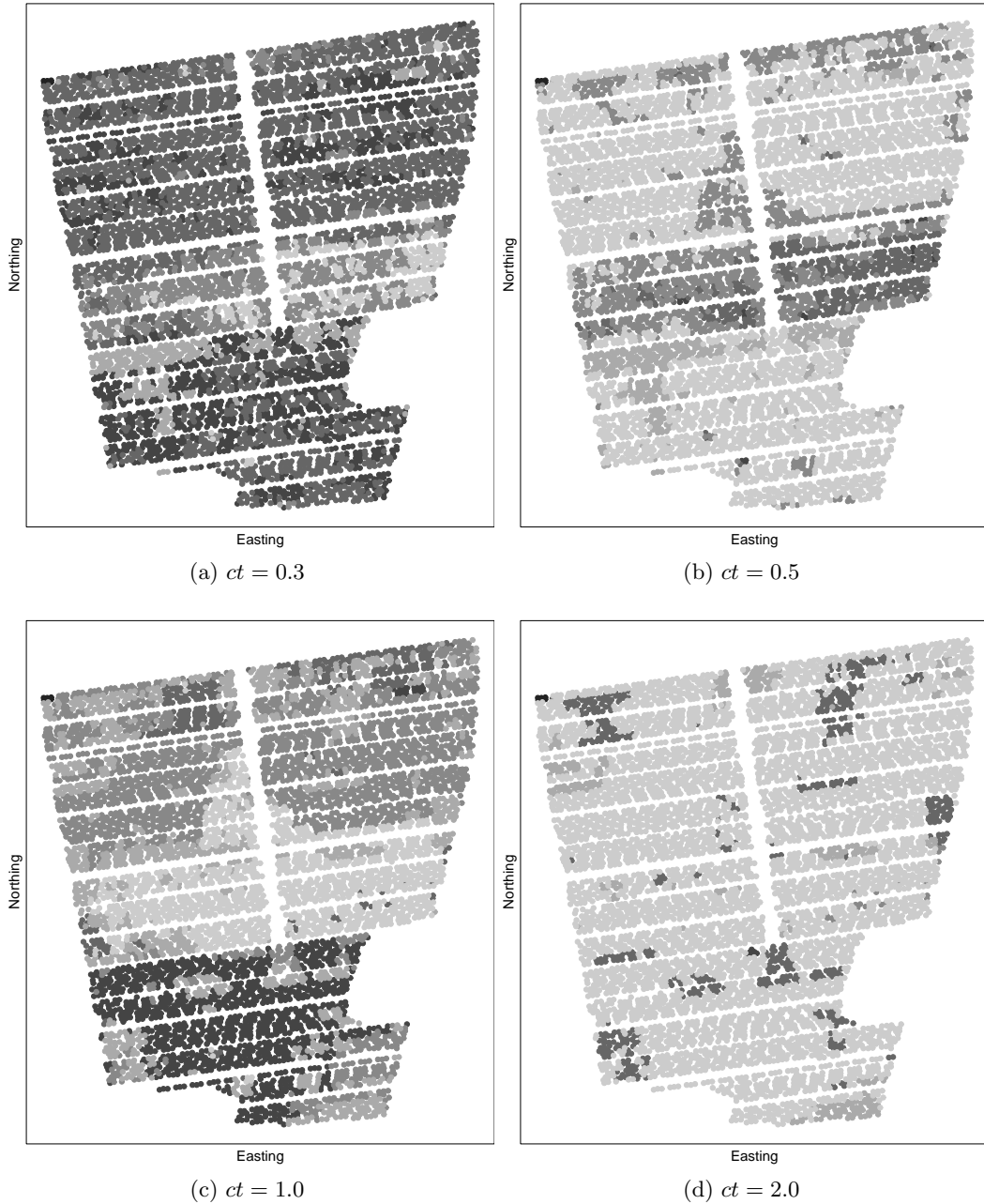


Figure 4.23: HACC-SPATIAL on F440, using the variable REIP49 (cp. Figure A.1a on Page 174), comparing the outcomes of different contiguity thresholds ($ct \in \{0.3, 0.5, 1.0, 2.0\}$), each with 6 zones left. With a rising threshold, the generated zones become more spatially contiguous, which has traditionally been accomplished by an additional smoothing step. However, neglecting (a) or enforcing spatial contiguity (d) does not lead to immediately meaningful zones, while setting ct in between these bounds creates much more pronounced zones.

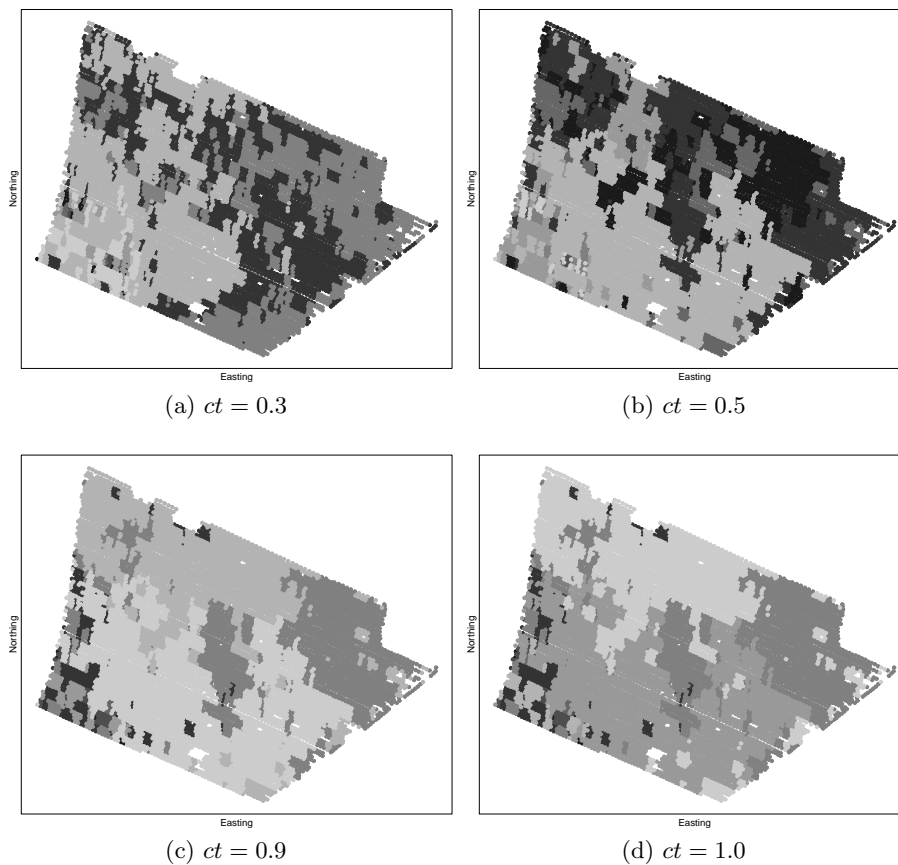


Figure 4.24: HACC-SPATIAL on F631, using the variable EC25 (cp. Figure A.12b on Page 189), comparing different contiguity thresholds $ct \in \{0.3, 0.5, 0.9, 1.0\}$. The higher ct is set, as designed, the more contiguous the zones tend to be. Figures (c) and (d) are quite similar since the deactivation of the spatial constraint occurred only 8 steps apart (after 7642 and 7648 steps, respectively, out of 7847 steps in total). The settings for $ct \in \{0.5, 1.0\}$ can be compared to Figure 4.31, where the only difference is the initial tessellation, leading to much smoother zones.

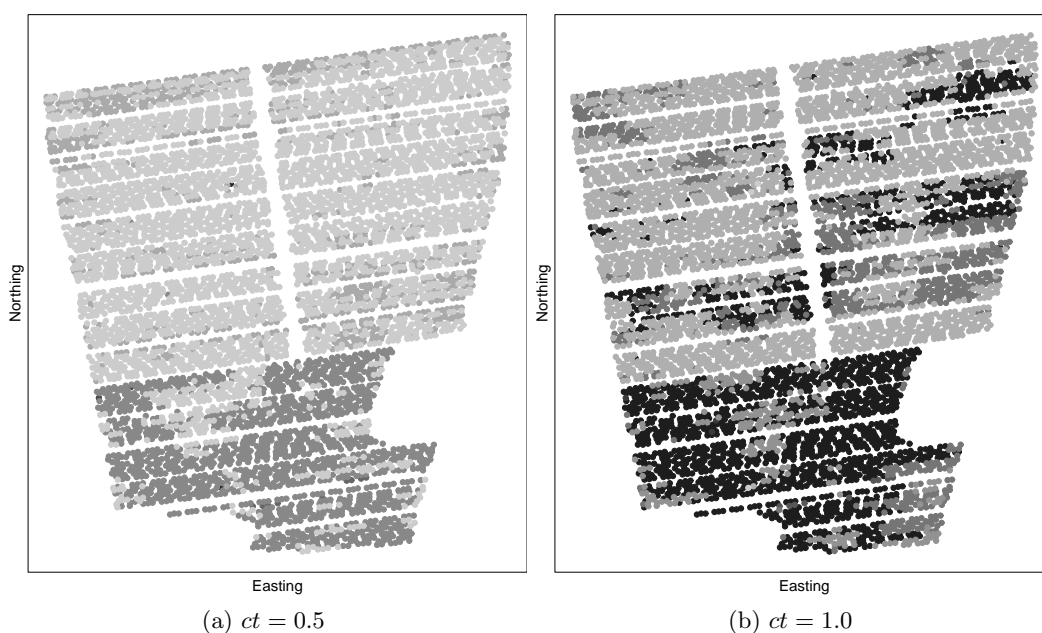


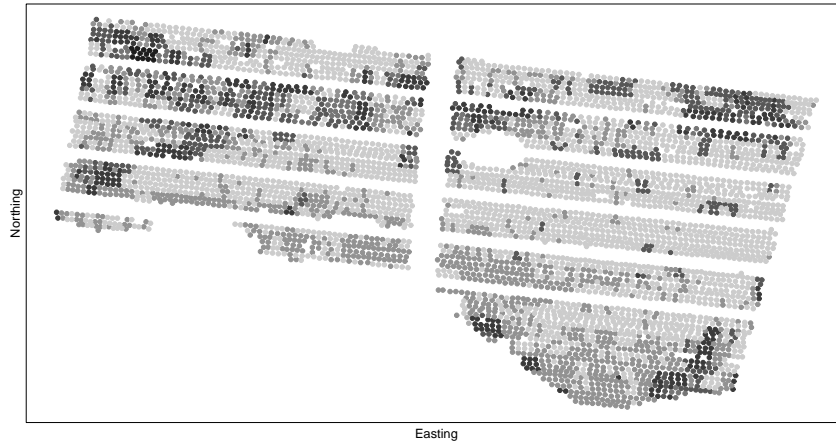
Figure 4.25: HACC-SPATIAL on F440, using the variable YIELD07 (cp. Figure A.2c on Page 175), as an example of generating yield-based management zones from one year's YIELD data. The contiguity threshold has been set to $ct = 0.5$ for the left figure and to $ct = 1.0$ for the right figure. While the left figure is much more contiguous, in essence having two large visible zones which correspond to the YIELD07 variable, the high contiguity setting shows more scattered zones distributed over the site. Incorporating multiple site-years of YIELD variables would simply involve the addition of those variables to the data.

4.7.3 Zones without Initial Tessellation, Multiple Clustering Variables

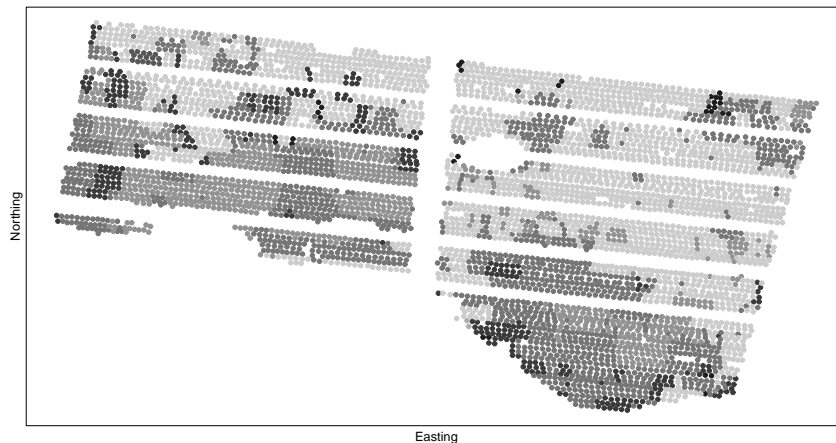
For using multiple clustering variables, two examples are provided in Figures 4.26 and 4.27. The original variables upon which the clustering is based are highly correlated and therefore allow for a visual comparison with the clustering results. While the first example would be a vegetation-based management zone using REIP32, REIP49 of the F611 site, the second would be a soil-sampling-based management zone generated on the P, MG, K, PH variables of the F550 site. Figure 4.26 shows the results of setting $ct \in \{0.5, 1.0\}$ and using both the REIP32 and REIP49 vegetation indicators. Those are correlated and should therefore exhibit rather distinct zones. However, the low contiguity threshold setting reveals a lot of scattered zones of which two or three are distinctly recognizable visually. In the high contiguity setting, these zones are shifted and their outline has changed. Therefore, the spatial contiguity threshold clearly has an effect on the contiguity of the clusters. Furthermore, for this data set, an additional initial tessellation reveals much clearer results, which is shown in Figure 4.34. For the soil sampling data on F550, the threshold is set to $ct \in \{0.5, 1.0\}$ and the clustering is shown for 12 and 3 clusters, respectively. At the low contiguity setting, as expected, the resulting zones are rather scattered, but still expose the main areas on the field which (for all four variables) show similar behaviour. The three zones can be characterized as follows:

- 1: largest zone, dark grey:** low PH, low P, low MG, low K
- 2: field borders, southern part, black:** high PH, high P, high MG, high K
- 3: east to west across the field, light grey:** high PH, high P, low MG, high K

However, the high spatial contiguity setting $ct = 1.0$ tends to expose a certain border running from north through south along the site, which exists in the data. Due to the high spatial contiguity threshold, HACC-SPATIAL is unable to merge clusters on both sides of this border. Finally, setting ct at a value too high also, by design, tends to prefer spatial contiguity rather than cluster similarity, which leads to meaningless clusters at the end of the clustering procedure.



(a) $ct = 0.5$, 6 zones



(b) $ct = 1.0$, 6 zones

Figure 4.26: F611, HACC-SPATIAL on REIP32 and REIP49 variables, low ($ct = 0.5$) and high ($ct = 1.0$) spatial contiguity setting, showing six zones. At the low spatial contiguity setting, the zones are rather scattered. Nevertheless, the southern border of the site seems to constitute one zone (dark gray), while a large zone (light gray) covers most of the site. In the high contiguity setting below, this large zone is changed and the zone at the southern site border is larger. However, for clearer results an initial tessellation may be used, which is shown in Figure 4.34.

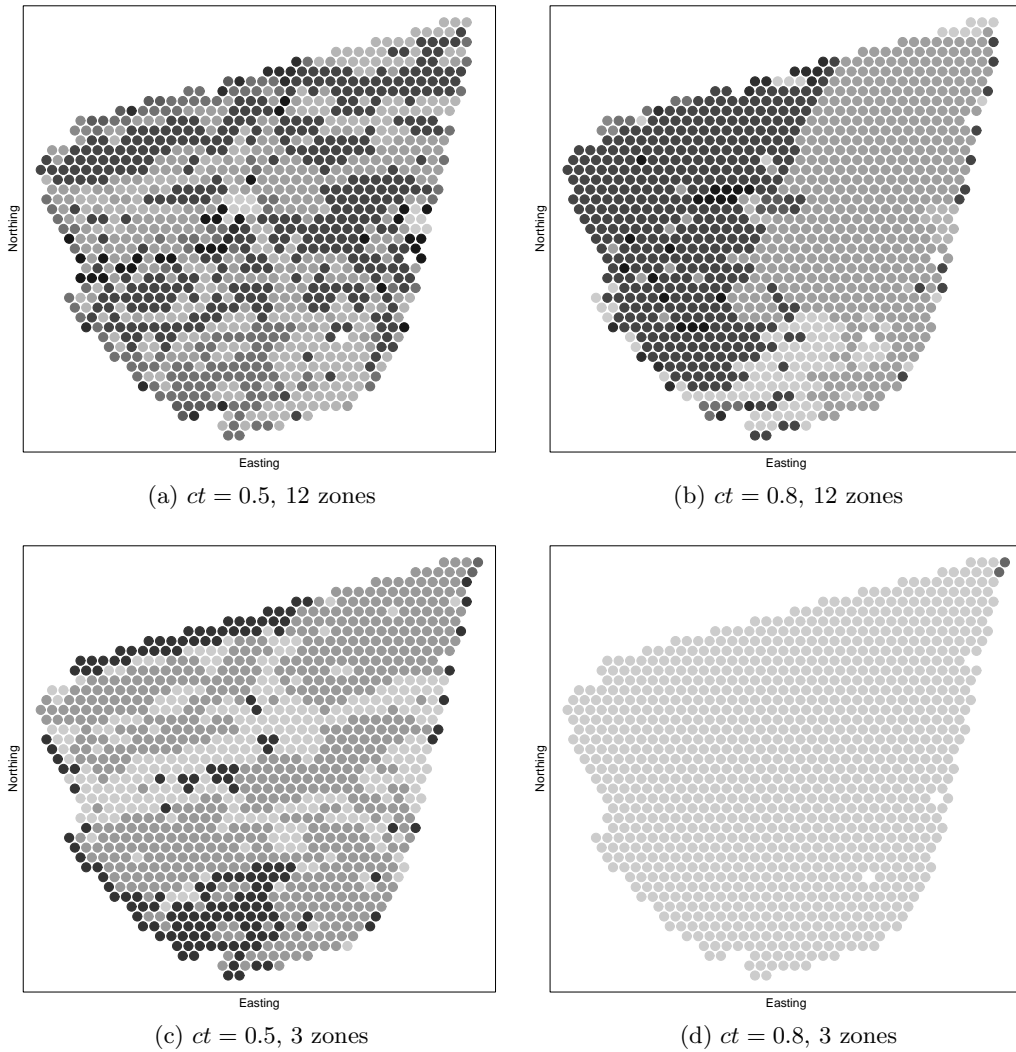


Figure 4.27: F550, effect of spatial contiguity threshold (cp. Figures 4.20 and 4.21). The clustering was performed on the four soil sampling variables from the F550 site (cp. Figure 4.35 on Page 132).

Without spatial contiguity constraint (low threshold 0.5, left), HACC-SPATIAL tends to produce zones which are scattered throughout the field. Nevertheless, a certain structure is emergent. On the right side, the clustering proceeds with a spatial contiguity threshold (0.8) which leads to much more contiguous zones which still exhibit spatial structures in the underlying data sets. However, towards the very end of the clustering, the zones tend to become meaningless since only a few outliers at the border remain to be merged.

4.7.4 Zones without Initial Tessellation, Incomplete Data Set

As a test case for HACC-SPATIAL when dealing with typical, erroneous data sets resulting from practical operations, the EC25 variable from the F610 data set is used. Spatial adjacency or spatial neighborhood of the data records is quite different from the data sets handled so far, which were rather uniformly distributed on the site. As can be obtained from Figure 4.28, HACC-SPATIAL without an initial tessellation step fails to account for irregularities like missing data strips and harvesting lanes. Nevertheless, with an initial tessellation, this can be accommodated for, as shown in Figure 4.37.

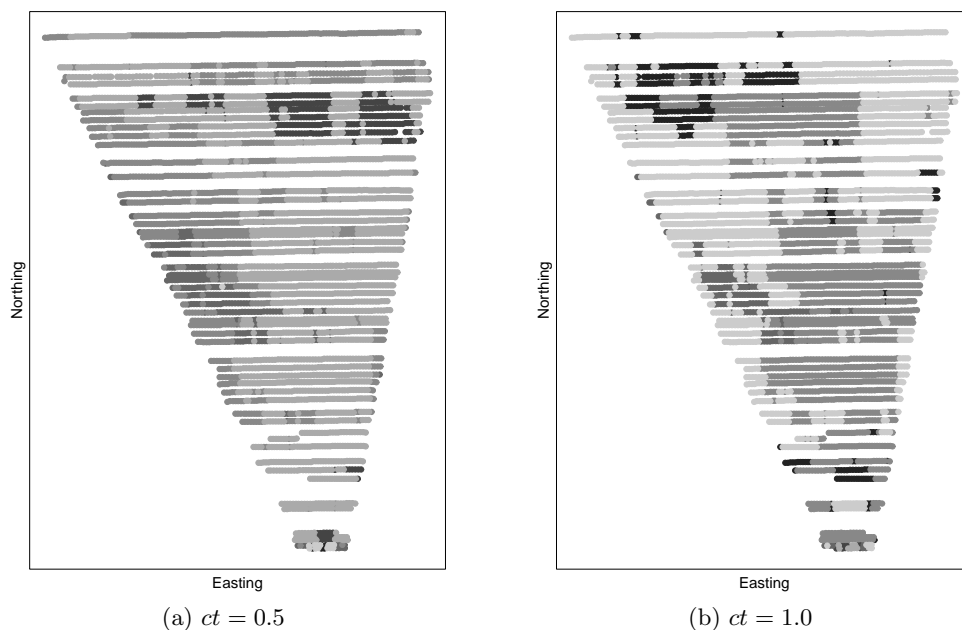


Figure 4.28: HACC-SPATIAL on F610, $ct \in \{0.5, 1.0\}$, using the variable EC25 (cp. Figure A.7b on Page 182). In the standard version (without an initial tessellation but with a spatial contiguity threshold HACC-SPATIAL creates zones which are visually rather inconsistent with the original data. The issue here is the geographical distance of points which is skewed due to the missing lines. Data records which would normally be spatially adjacent are not adjacent and therefore not considered for merging. For this type of data sets, the initial tessellation proves worthwhile, as demonstrated in Figure 4.37.

4.7.5 Zones with Initial Tessellation, One Clustering Variable

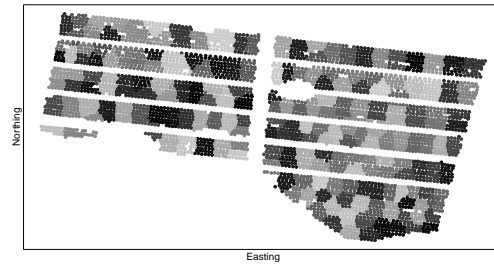
This section shows the effects that choosing in initial tessellation has on the outcome of HACC-SPATIAL. As with the examples without the initial tessellation, it starts with the clustering based on one variable and proceeds towards multiple variables and an erroneous data set.

Figure 4.29 depicts the different stages of HACC-SPATIAL for the EC25 variable of the F611 field. HACC-SPATIAL is run with $ct \in \{0.5, 1.0\}$ and the effects of this setting are displayed. While the beginning (Figure 4.29a) of the clustering is the same, the course of the clustering is different. At low spatial contiguity (left figures), even towards the end of the clustering, no meaningful clusters are emergent. At high spatial contiguity (right figures), those clusters start emerging after around half of the clustering (Figure 4.29c) and are clearly visible after 180 of 200 merging steps (Figure 4.29e). In the latter figure, the resulting clusters already highly correspond to the actual EC25 variable and can possibly contribute to subdividing the field.

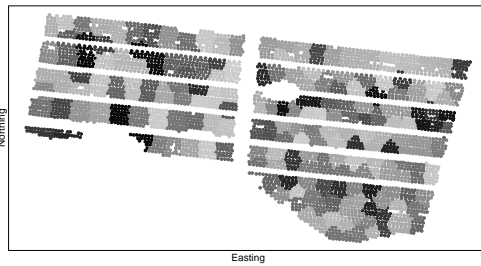
Figure 4.30 presents the clustering of the REIP49 variable on the F440 field, using 200 clusters initially. The left figures again show a low contiguity setting ($ct = 0.5$), while the right figures show a high spatial contiguity setting ($ct = 1.0$). As with Figure 4.29, the spatial contiguity structure already starts emerging earlier during the course of the algorithm (Figure 4.30b) and keeps getting clearer towards the end of the clustering at a high contiguity setting. On the other hand, it remains largely undiscovered at a low contiguity setting, even towards the end of the clustering (Figure 4.30e). Furthermore, the result in Figure 4.30f may be too coarse for a practical application. By traversing the dendrogram of the clustering backwards, one could examine stages at which certain clusters have been merged and compare those clusters, leading to a deeper understanding of the field circumstances. E.g. if the northern half of the field in Figure 4.30f appears too uniform in the clustering, i.e. is assumed to be more heterogeneous in reality, the clustering can easily be checked for this issue. In this case, Figure 4.30d would show a few sub-clusters (sub-zones) which could be taken into account in a practical setup. Naturally, HACC-SPATIAL can be run without the initial tessellation to compare the results.

A further result is presented in Figure 4.31, where the variable EC25 of the F631 field is used for management zone delineation. The field is initially tessellated into 250 clusters and the clustering is run with low and high contiguity settings ($ct \in \{0.5, 1.0\}$) to compare the results. As in the preceding results, clustering with low spatial contiguity yields mostly non-contiguous clusters (as expected) until spatially contiguous clusters start emerging towards the very end of the clustering (Figure 4.31e). On the other side, clustering with high spatial contiguity starts showing emergent clusters after around 200 merging steps (Figure 4.31b) and subsequent clusters clearly correspond to the actual variable value (Figure A.12b). The clusters are not limited to convex shapes and account for the irregular shape of the field (missing data, irregular borders, “holes”). If the clustering in Figure 4.31f is deemed to coarse, the hierarchically structured clustering easily allows for subdividing single clusters by traversing the dendrogram.

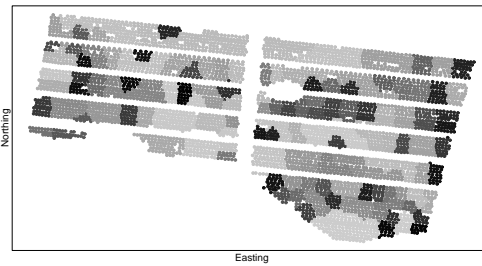
A direct comparison of the results of HACC-SPATIAL when applied to the same input data with an initial tessellation and varying contiguity thresholds is provided in the following. Figure 4.32 shows the REIP32 variable on the F440 field, clustered by HACC-SPATIAL,



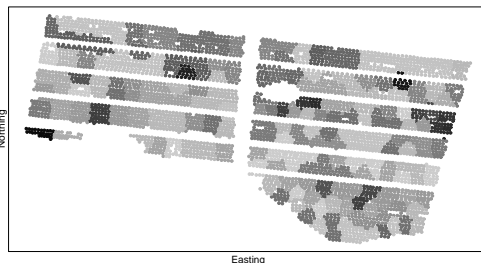
(a) 200 initial clusters



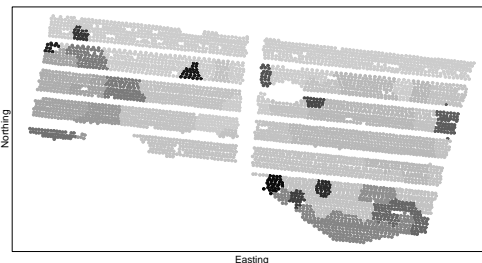
(b) after 100 steps, $ct = 0.5$



(c) after 100 steps, $ct = 1.0$



(d) after 180 steps, $ct = 0.5$



(e) after 180 steps, $ct = 1.0$

Figure 4.29: HACC-SPATIAL on F611, using the variable EC25 (cp. Figure A.10a on Page 186). The figures illustrate the influence of the contiguity parameter. The beginning (a) of the clustering is identical, with 200 initial clusters. Figures (b) and (d) show the clustering with low contiguity ($ct = 0.5$), while (c) and (e) show the clustering with high contiguity ($ct = 2$). The difference in spatial contiguity is distinct: while (b) and (d) exhibit barely any spatially contiguous structures throughout the clustering, (c) and (e) present visible spatial clusters which are quite congruent with the EC25 variable right from the beginning.

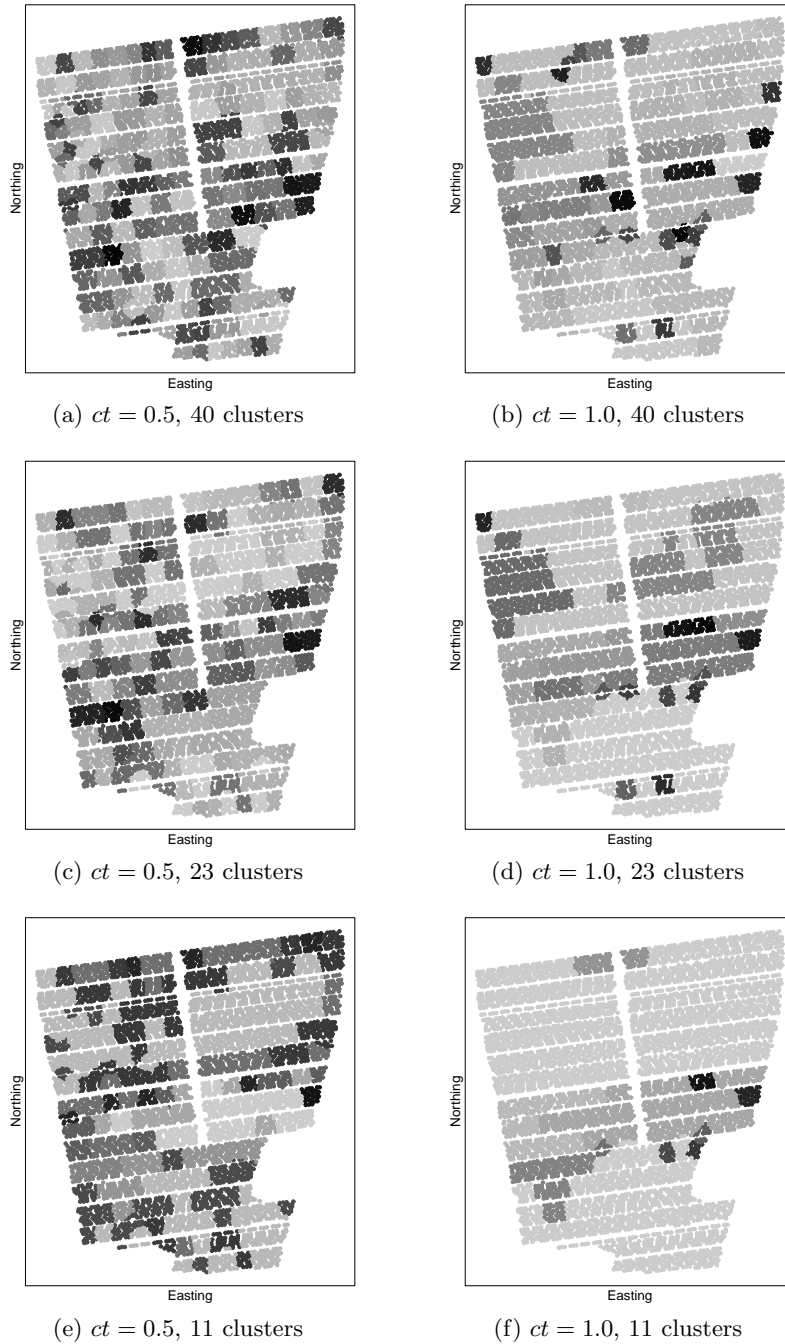


Figure 4.30: HACC-SPATIAL on F440, using the variable REIP49 (cp. Figure A.2b on Page 175), comparing low contiguity (left) and high contiguity (right). While (a), (c) and (e) show scattered clusters, even towards the end of the clustering, Figures (b), (d) and (f) exhibit clear contiguous clusters at different levels of the dendrogram. The spatially contiguous clusters correspond to the actual variable values and the different clustering stages provide a deeper understanding of the “behaviour” of REIP49 on this field.

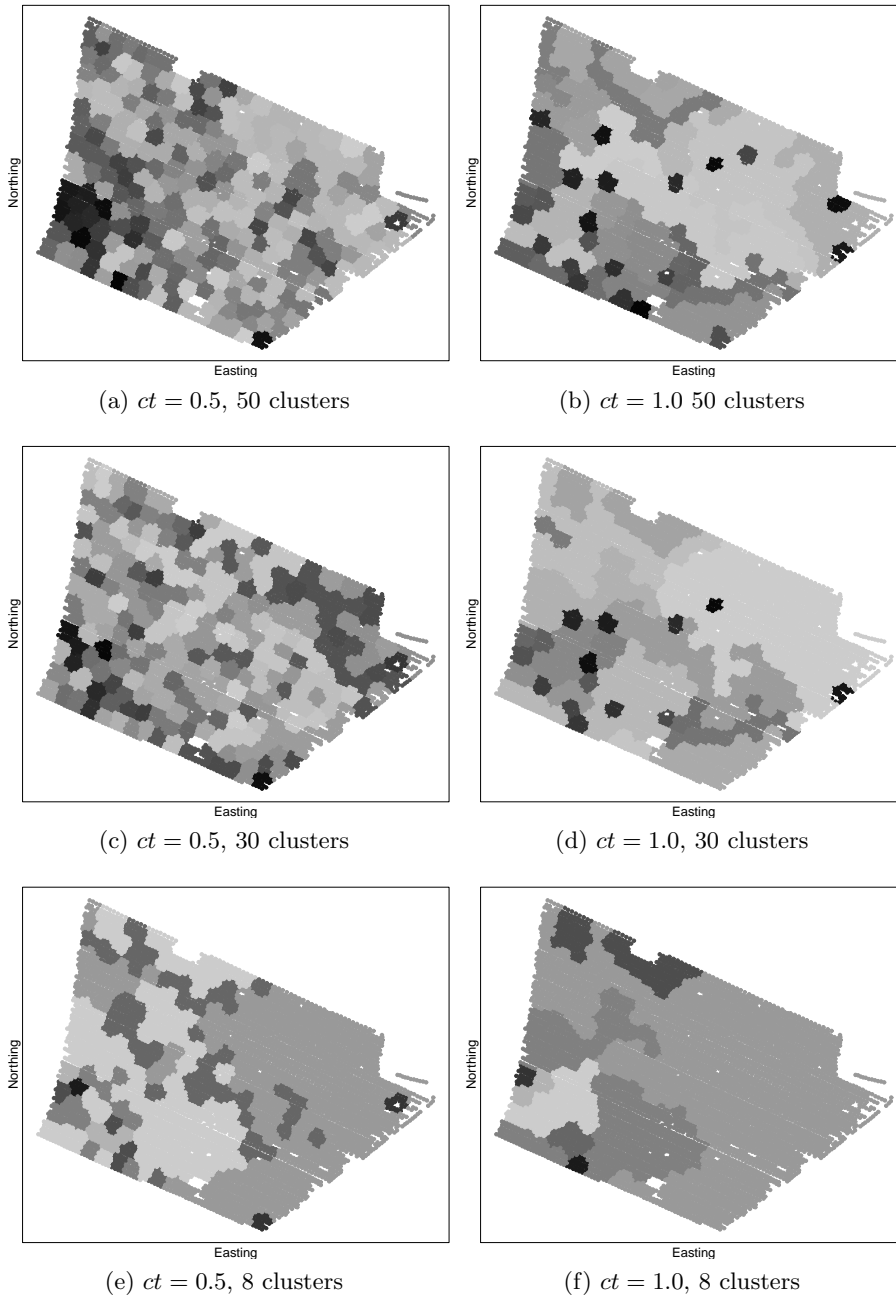


Figure 4.31: HACC-SPATIAL on F631, using the variable EC25 (cp. Figure A.12b on Page 189), starting with 250 clusters. As in Figures 4.29 and 4.30 clustering with low (left figures) and high (right figures) spatial contiguity shows considerable differences in the spatial structure of the resulting clusters. At low spatial contiguity the algorithm starts producing visible spatially contiguous clusters only towards the end of the clustering (e), while spatially contiguous clusters start emerging much earlier when clustering with high spatial contiguity (b).

showing the stage at which 15 clusters are left. While Figure 4.32a shows almost no visible spatial contiguity, this changes gradually towards Figure 4.32d where the clusters are mostly spatially contiguous.

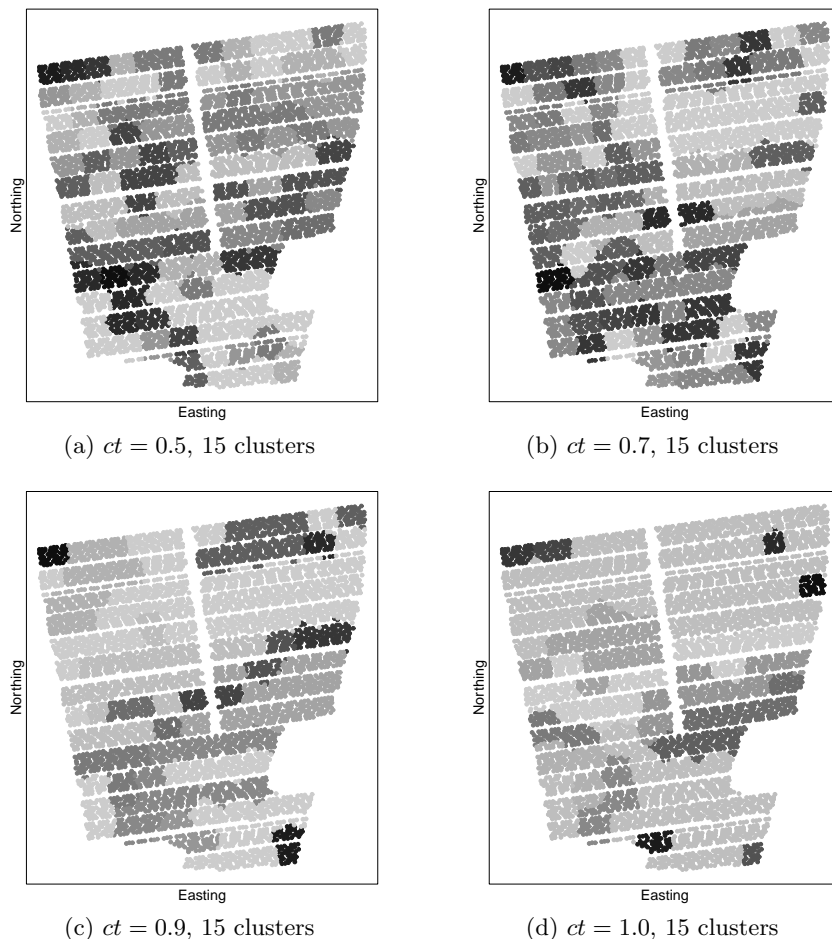


Figure 4.32: HACC-SPATIAL on F440, 120 initial clusters, using the REIP32 variable and demonstrating the effect of different spatial contiguity settings. The contiguity threshold as defined in Equation 4.1 is varied from 0.5 via 0.7 and 0.9 to 1.0 (left to right, top to bottom). While (a) shows spatially rather scattered clusters, the change in the designed contiguity ratio threshold varies the spatial contiguity of the clusters until the spatial contiguity is enforced in Figure (d).

Figure 4.33 explores the effects of assuming spatial autocorrelation when the variable under study is differently spatially autocorrelated. This is the case for the N1 variable on the field F440 (cp. Figure A.1a on Page 174). It exhibits clear strips along which different fertilization strategies were carried out. HACC-SPATIAL optionally exploits spatial autocorrelation in its tessellation phase: a k -Means clustering is performed to generate roughly equal-size initial clusters. Those are likely to contain similar data records, due to spatial autocorrelation. A coarse initial tessellation produces clusters containing data records which are rather dissimilar and also spanning large areas – too large for the purpose

of the ensuing hierarchical agglomerative clustering. The initial tessellations are shown in Figures 4.33a and 4.33b, while for the latter clustering an initial cluster size of ten data records was used, leading to 640 initial clusters. At larger cluster sizes, HACCC-SPATIAL is bound to miss the borders of the actual zones, simply because the initial clustering is too coarse (Figure 4.33c). At smaller cluster sizes, towards the end of the algorithm, the clusters are much more congruent to the actual zones (Figure 4.33d).

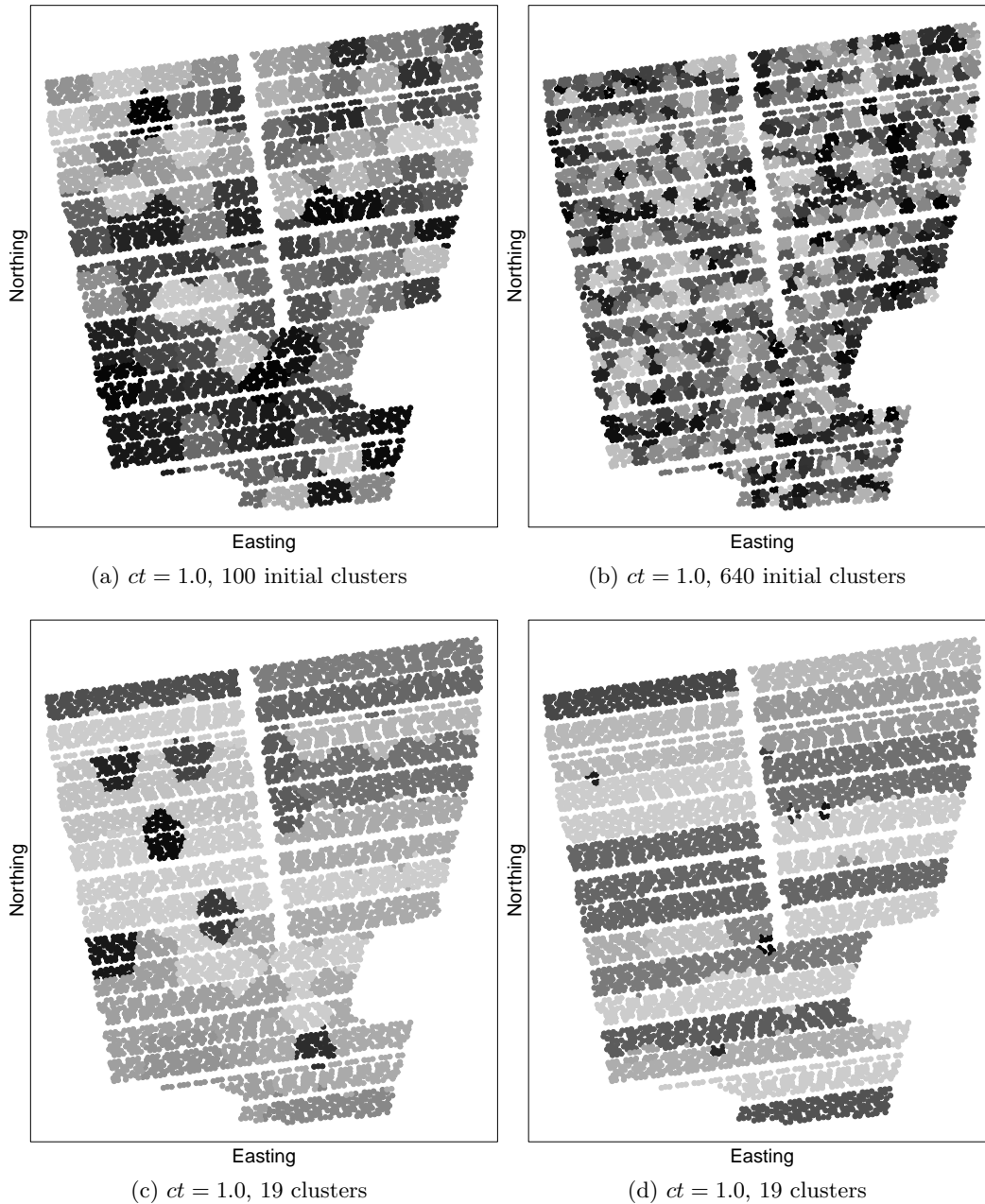


Figure 4.33: HACC-SPATIAL on F440, using the variable N1 (cp. Figure A.1a on Page 174), starting with 100 clusters (left) and 640 clusters ($\frac{1}{10}$ of data set size, right). Since the N1 variable is distributed in clearly visible strips in an experimental fertilization layout, it does not exhibit the same amount of spatial autocorrelation as the natural variables EC25 or REIP. Both trials are set to run at high spatial contiguity ($ct = 1.0$), while the initial tessellation is varied. HACC-SPATIAL succeeds at delineating the different field strips (Fig. (d)), but must be set to a rather high number of initial clusters, otherwise the artificially generated field strips may be missed (Fig. (c)).

4.7.6 Zones with Initial Tessellation, Multiple Clustering Variables

The REIP32 and REIP49 values are often highly correlated. Consider the F611 field (cp. Figures A.9a and A.9b). The question of how to determine vegetation zones based on the REIP value can be answered by HACC-SPATIAL. Running the algorithm in a high spatial contiguity setting returns distinct areas which align approximately with the vegetation indicators' values (cp. Figure 4.34). While this is not a task that is typically handled by clustering algorithms in precision agriculture, the clustering outcome serves as a straightforward example of what HACC-SPATIAL can accomplish and what the algorithm can contribute to better understanding the field.

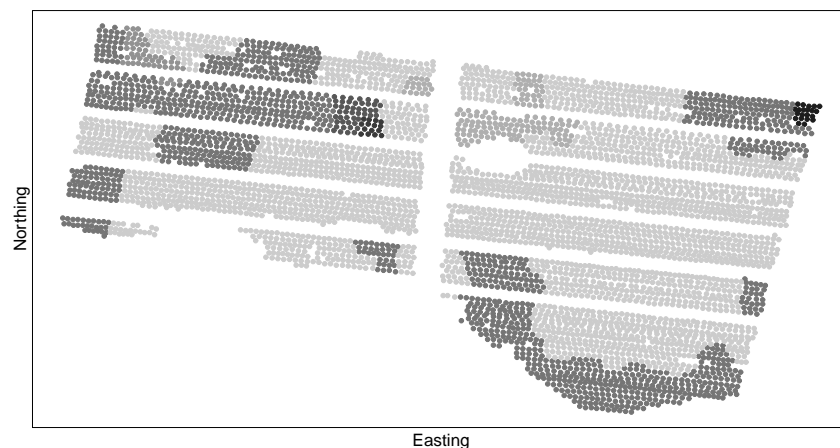
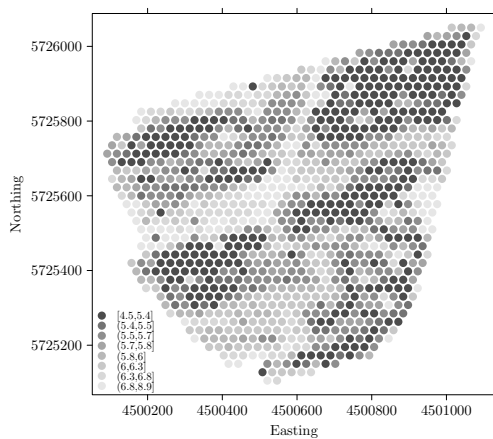


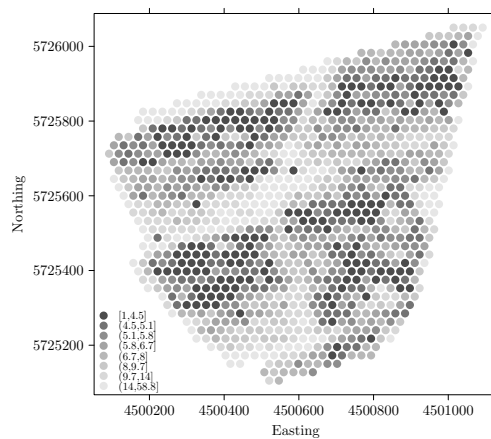
Figure 4.34: F611, HACC-SPATIAL on REIP32 and REIP49 value, high spatial contiguity setting, starting with initial 200 clusters, 7 clusters shown. Apart from the largest zone in the center of the field, which contains mostly medium REIP values, the darker areas in the southwestern and northeastern part of the field have low REIP values. The same holds for the larger zone in the northwest of the field and the zone on the southeastern border. HACC-SPATIAL tolerates irregularities on the field, such as the “hole” east of the vertical center strip and the non-straight borders, especially in the southwestern part.

In the F550 data set, soil sampling variables are available. The variables PH, P, K, MG can be expected to be highly correlated (positively or negatively, cp. Figure 4.35). A visual inspection shows that a spatial structure is emergent, with four to six visible areas, separated by another almost cross-shaped area in the center. This structure shall now be discovered by HACC-SPATIAL. The effect of the spatial contiguity parameter has been demonstrated in the previous section by setting it to a high and a low spatial contiguity value. Values in between those borders are possible and shift the point at which the switch from hard to soft spatial contiguity constraint occurs. In the following, the spatial contiguity parameter is not of primary concern, therefore different settings are not compared further.

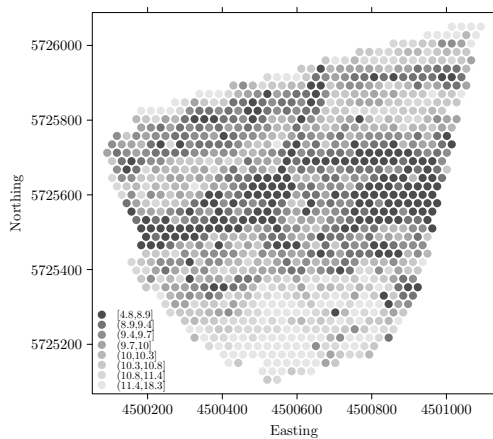
Figure 4.36 (in combination with Fig. 4.35) illustrates the course of HACC-SPATIAL when applied on the four soil sampling variables. Figures 4.36c and 4.36d clearly show one large zone spanning most of the field. In addition, the southwestern part of the field must be distinctly different from this zone. Especially in Figure 4.36c, a further zone (consisting



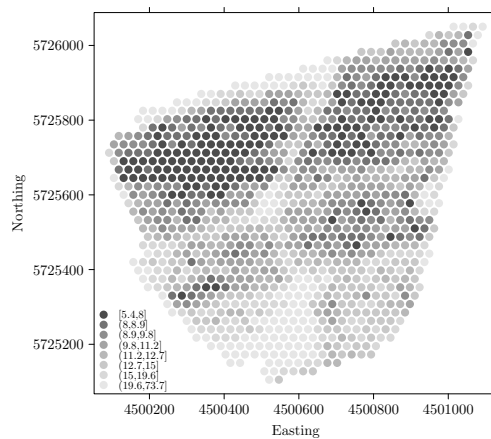
(a) F550, pH value



(b) F550, P concentration



(c) F550, Mg concentration



(d) F550, K concentration

Figure 4.35: F550, four soil sampling variables, used as input for HACC-SPATIAL: pH value, P, Mg, K concentration (left to right, top to bottom). There are a few zones which are clearly visually distinguishable, those are briefly described on the opposing page in Figure 4.36.

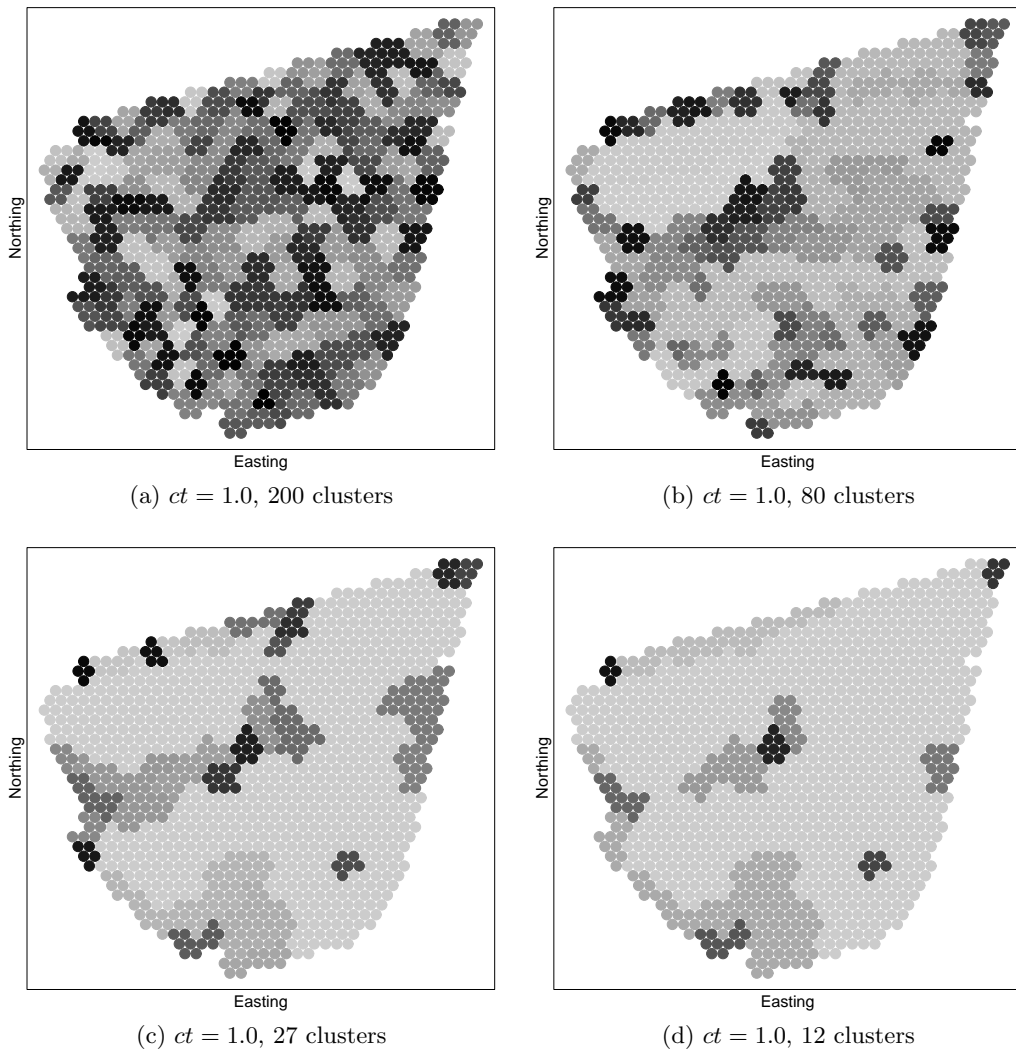


Figure 4.36: HACC-SPATIAL on the soil sampling variables of F550 (cp. Figure 4.35 on opposing side), high contiguity setting, starting at 200 clusters (a), with different numbers of clusters yielding insight into the field's zones: 80 (b), 27 (c), 12 (d). The largest zone keeps emerging quite early, while even towards the end of the algorithm the field borders and the southwestern part of the field are rather stable and therefore differ considerably from the rest of the field. The center of the field can also be distinguished quite early, but is also merged rather early with the largest zone. By examining the cluster hierarchy, this can be discovered.

of numerous clusters) protrudes from the western side of the field towards the center. There are additional smaller parts on the borders of the field which are distinctly different. In a rather coarse categorization, these zones can be characterized as follows:

- 1: largest zone:** low PH, low P, low MG, low K
- 2: zone protruding from the west:** high PH, high P, low MG, high K
- 3: southwestern zone:** high PH, high P, high MG, high K
- 4: northern field border:** high PH, high P, high MG, high K
- 5: eastern field border:** high PH, high P, low MG, high K

Zones 3 and 4 are rather similar, as well as zones 2 and 5, and may be merged manually for practical purposes. This is similar to what has been done in [Murray and Shyy, 2000], where a classification of property crime rates is followed by a grouping of these into six classes. However, in [Murray and Shyy, 2000] the data exhibit different spatial density and only one variable is considered, while in the above setup four variables are used.

Furthermore, the algorithm identifies a few rather small clusters even towards the end of the clustering. These can therefore be expected to differ considerably from the adjacent clusters or the surrounding zone. The small cluster in the southeastern part of the field (containing six data vectors) may either be an artifact of the tessellation phase of the algorithm or may have distinct properties: the cluster groups values with a high MG and high K concentration, but a variety of PH and P values. The northeastern and northwestern field corners are also identified as distinct zones and can be treated accordingly.

A comparison to the result obtained without the initial tessellation (cp. Figure 4.27) shows that both trials result in similar zones, depending on which step of the clustering is used for examining the zones. In the previous run, three zones were generated directly, while manual labour would be expected in the current version. For an exploratory algorithm that allows for an overview of the data, this result is quite encouraging.

4.7.7 Zones with Initial Tessellation, Incomplete Data

The F610 data set provides a test case for HACC-SPATIAL regarding its capabilities and its tolerance in dealing with incomplete data sets. In the F610 data set, strips of data are missing, which may occur in practice. Although these missing data strips can in principle be predicted using geostatistical methods like kriging, HACC-SPATIAL may be applied directly. The apparent electrical conductivity variable (cp. Fig. A.7b) shows a spatial structure that should be discovered by HACC-SPATIAL. Figure 4.37a shows that the initial k -Means clustering ignores the strips where the data are missing. It still creates initial areas which are similar in size. However, these areas now differ in the numbers of data records they contain. Nevertheless, the algorithm succeeds in delineating appropriate zones which reflect the heterogeneity on the field. Those are shown in Figure 4.37b.

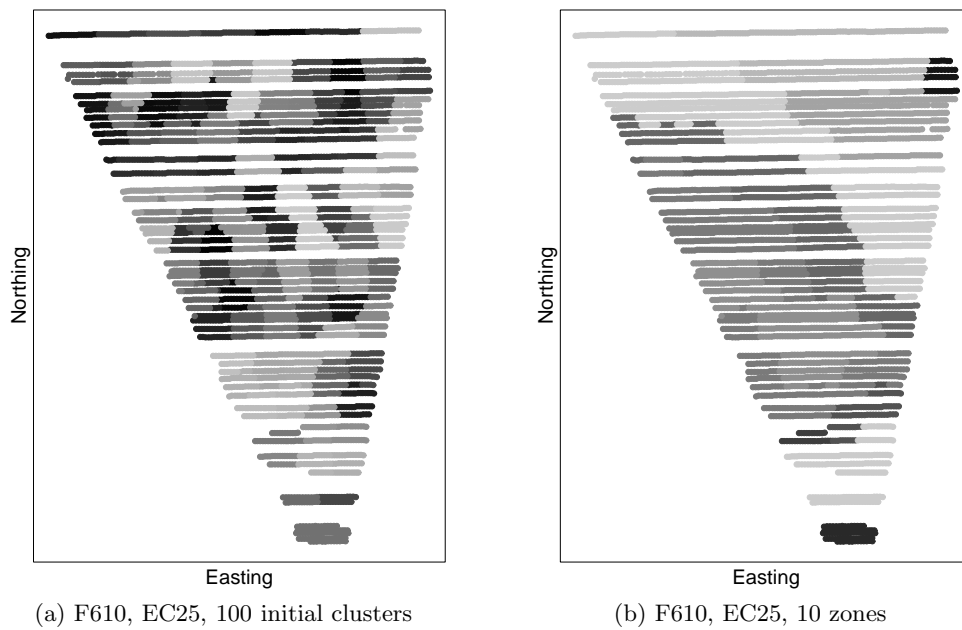


Figure 4.37: HACCP-SPATIAL on F610, high spatial contiguity, using the variable EC25 (cp. Figure A.7b on Page 182), starting with 100 clusters (a), and showing 10 zones (b). This demonstrates the ability of HACCP-SPATIAL to work on incomplete data where whole strips of data are missing. The final zones reflect the heterogeneity of the field and are similar to a choropleth map where the borders of the zones correspond to contour lines.

4.8 Heuristic Parameter Guidelines for HACC-spatial

HACC-SPATIAL features two parameters that influence the course and the result of the algorithm: k for the initial tessellation and the spatial contiguity constraint threshold. Recommendations for these parameters are provided below; both depend on the data set characteristics and must be set manually, which is not an issue in the exploratory task described here. The recommendations are based on the experimental evaluation in the previous section and on the experience both with HACC-SPATIAL and the PA data sets.

initial tessellation k The setting of k depends on the field variables' heterogeneity. For spatially rather homogeneously distributed field variables, this can be set to a value up to $\frac{N}{100}$, where N is the number of available data records. For rather heterogeneous data sets such as the ones encountered here, it may be set to as high as $\frac{N}{10}$. For the influencable variables, such as N fertilizer, this initial tessellation should be omitted.

Finally, k determines the granularity of the zones, i.e. how well those align with field variables' characteristics. A higher k allows for finer resolution. This resolution may also ultimately be limited by the best spatial resolution that the available farming equipment can achieve.

contiguity ratio threshold This parameter determines heuristically when to switch off the spatial contiguity constraint during the course of the algorithm. According to Equation 4.1 (Page 111), the mean distances of adjacent and non-adjacent clusters are computed in each step, while the ratio of these values defines the spatial contiguity. Higher spatial contiguity can be achieved by setting the threshold to a higher value, while setting it to a lower value enables an earlier switch from the constrained to the unconstrained algorithm.

In the results presented in the previous section, the high spatial contiguity is achieved by setting the threshold to ≥ 1.0 , while low spatial contiguity is achieved by setting the threshold to ≤ 0.3 . Values in between these bounds shift the algorithm step at which the constraint is switched off (cp. Figure 4.32). As pointed out earlier, this particular step for switching can also be determined using a fixed setting or a different heuristic approach. Both full and no spatial contiguity can be achieved easily.

4.9 Summary

This chapter presented a hierarchical agglomerative clustering approach with a spatial constraint for the task of management zone delineation in precision agriculture. Based on the literature review from major precision agriculture publications and the approaches towards management zone delineation, the key shortcomings of existing approaches were identified in Section 4.3. Most prominently, the spatial information in the data sets was neglected. Hence, a few requirements for a novel zone delineation approach based on spatial clustering were outlined in Section 4.4. From the clustering point of view, existing algorithms were examined regarding their capabilities in dealing with the specific kind of data encountered here in Section 4.5. None of those algorithms was found to be suitable to the task of management zone delineation.

Therefore, HACC-SPATIAL, a novel algorithm based on hierarchical agglomerative constrained clustering was devised and implemented in Section 4.6. Results on the data sets available in this thesis were presented in Section 4.7 and the effects of the two main parameters of the algorithm were investigated. HACC-SPATIAL incorporates a spatial contiguity constraint into hierarchical agglomerative clustering and is therefore well-suited to be used in an exploratory clustering approach such as management zone delineation.

HACC-SPATIAL allows for setting a geospatial precedence for data records by appropriately reducing the search space using a spatial constraint. In its first phase, the hierarchical agglomerative algorithm only considers geospatially adjacent data records or clusters for merging. In its second phase, the search space is broadened to also include geospatially non-adjacent data records or clusters. The switch from the first phase to the second phase occurs when a user-influencable *contiguity threshold* is reached. This is computed as the ratio between the mean distances of adjacent and non-adjacent clusters. By setting this threshold accordingly, the user can decide gradually how important spatial adjacency is to his particular clustering task. On the one hand, HACC-SPATIAL can emulate an unconstrained traditional hierarchical agglomerative algorithm. On the other hand, it can enforce the spatial contiguity constraint.

Furthermore, HACC-SPATIAL has an additional feature which exploits spatial autocorrelation. It allows for an initial tessellation of the site into small groups of data records, without actually considering their variables' values, but rather by clustering them according to their coordinates. The underlying assumption is that adjacent records are rather similar due to spatial autocorrelation. This additional tessellation leads to a better tolerance of HACC-SPATIAL against erroneous data sets and a computational speedup. This tessellation can also be seen as a smoothing step which lowers the effect a single data record may have on the clustering. However, it has also been shown that, if the underlying assumption of spatial autocorrelation fails, the initial tessellation gives worse results than the version of HACC-SPATIAL starting without the tessellation.

4.9.1 Requirements Revisited

In the following, HACC-SPATIAL is examined with regard to the clustering and zone delineation requirements recorded in Sections 4.4.1 and 4.4.2.

H1: multi-level By definition of hierarchical agglomerative clustering, it creates a multi-level hierarchy (dendrogram) of clusters which may be examined to reveal clusters and their subclusters.

H2: exploratory nature By using the generated hierarchy, a user can easily explore the outcome of the clustering. With or without the initial tessellation, a “quick first view” of the data can be generated.

H3: incorporate spatial proximity HACC-SPATIAL by design incorporates spatial proximity through a spatial constraint imposed on the cluster merging candidates in its first phase. Only those clusters which are spatially adjacent may then be merged. However, this constraint may be switched off to allow for broadening the search space.

- H4: efficiency and effectiveness** Since HACC-SPATIAL with an average linkage criterion involves the same routines and substeps as a traditional hierarchical agglomerative clustering, the same computational bounds hold. For typical implementations, the runtime is in $O(N^3)$, while the best current implementation achieves $O(N^2)$ ⁴, where N is the number of data records.
- H5: no density differences in data** HACC-SPATIAL does not try to exploit density differences and is therefore suited to the PA data sets in this thesis. It rather exploits the spatially dense data records' distribution. With the initial tessellation, it is also able to accommodate for "regular irregularities" such as missing data strips (e.g. F610 data set).
- H6: spatial contiguity of clusters** This requirement has been accommodated for by the first phase of HACC-SPATIAL, which only allows spatially adjacent clusters to be merged.
- S1: only few parameters / understandability** The two parameters *contiguity threshold* and optionally the initial tessellation have an intuitive meaning and can easily be explained.
- S2: flexible number of management zones** As with **H1**, **H2** and **H4**, HACC-SPATIAL generates a dendrogram of clusters, which can easily be manually examined to yield different numbers of management zones.

4.9.2 Future Work

Once HACC-SPATIAL finishes, the resulting dendrogram should be examined further. A chosen clustering may easily be examined using frequent itemset mining. Numerical variables can be converted to a three- or five-value categorical scale and the resulting frequent sets could be generated in an approach similar to the result in Figure 4.36, where this was done manually. This would lead to a coarse "first-look" exploration which may be refined further during subsequent steps. This is closely related to [Shekhar et al., 2001], where association rules are turned into spatial co-location rules. In a different approach which was not applicable here, spatial association rules were directly generated in [Koperski and Han, 1995]. For clusterings based on multiple variables, HACC-SPATIAL can be easily customized to cater for different importances of single variables in the clustering. For example, the EC25 variable may be the main factor for the zones, while an additional sensor variable S1 or a certain correction by using past YIELD variables may also be known to influence the zones in a practical setup. Technically, this only requires setting appropriate weights in the internal distance calculations of HACC-SPATIAL. An important practical issue is the availability of soil sampling data. Acquiring these data is invasive and usually prohibitively expensive to be done each season. It should therefore be investigated whether the mineral content can be inferred from other variables which can be acquired more cheaply, such as remote sensing data. The above management zone delineation approach may then be used to investigate zones on these data sets.

⁴fastcluster: <http://math.stanford.edu/~muellner/fastcluster.html> (last visited 2011-10-18)

Chapter 5

Conclusions

5.1 Thesis Summary

This study revolved around a few highly related topics. *Precision agriculture* (PA) itself is interdisciplinary and involves knowledge in agriculture, computer science and geostatistics. PA is clearly about to become a data-driven approach to agriculture. This requires appropriate algorithms and technologies to deal with PA tasks in a consistent and efficient way. In its first chapter, this thesis introduced precision agriculture and laid the groundwork for understanding the remainder of the thesis. Since PA data sets are at the very core of this work, the sets and the variables therein were explained. In addition to the actual PA data, additional digital elevation model (DEM) data were acquired and terrain attributes were derived from those DEM data. They were integrated into the spatial PA data sets accordingly to facilitate analyses. Related geostatistical topics such as spatial autocorrelation and data preprocessing were also detailed in Chapter 2. The two main topics of this thesis, making up Chapter 3 and 4, were briefly laid out in Chapter 1.

The first question covered by this thesis was concerned with a traditional task in agriculture: yield prediction. With nowadays' technology and ever larger data collections in PA, the question of data reduction will become ever more important. In a yield prediction setup, which of the variables are actually the most important or interesting ones for this task? And, as a subquestion, when it comes to regression modeling for yield prediction, which regression model(s) should be chosen?

While high-resolution spatial data sets keep growing, the overwhelming amount of small-scale information may lead to issues with discovering the important relationships between data variables on an agricultural site. This may be either seen as a chance or a curse – with *data mining* being concerned with the “chance” part of this dichotomy. While traditionally a farmer knew his sites rather well from experience and could thus point out specific areas that needed different treatments than others, the information available in PA requires appropriate techniques to deal with the data. One of the tasks in (precision) agriculture is “management zone delineation”, where the problem is to create homogeneous zones on a site which are then treated differently from each other, for different purposes.

5.2 Results of the Spatial Variable Importance Approach

Chapter 3 dealt with the particular task of yield prediction (YP), based on PA data sets. From a data mining point of view, YP is a multivariate regression task, where yield is the response variable and numerous predictors exist. Hence, a few of the most frequently used regression models were introduced. However, in the context of spatial data sets, the typically used cross-validation approach exhibits a few pitfalls, most notably caused by the involved random sampling to create the training and test sets. Due to spatial autocorrelation, geospatially adjacent data records are likely to be rather similar to each other and therefore the random sampling had to be adapted to avoid underestimating the predictive error of any regression model. This was achieved using a spatial clustering approach, which involved a simple, yet effective run of k -Means on the data sets' coordinates. This spatial tessellation then allowed for reusing the cross-validation in a spatial setting. The regression models could be run using a few predictor variables to predict yield, answering the first question which of the models is actually the most appropriate for YP in PA. Without further tweaking and parameter space searches, the linear model, as well as support vector regression and bagging turned out to be consistently of high performance, resulting in a low predictive error. The neural network used in the dissertation which partly led to this thesis typically performed less well.

The second basic idea of this chapter toward assessing a single variable's importance in a yield prediction setup is as follows: based on a trained regression model in a cross-validation setup, measure the increase in prediction error on the test set after one variable in the test set has been permuted. Important variables are expected to return a rather high error increase, while less important variables are expected to have no influence on the model. It is assumed that a variable which starkly influences the model's prediction is also important for influencing the actual yield in practice. This spatial variable importance (SVI) approach could clearly be validated using a few subsets of the PA data sets where specific fertilization strategies were carried out. Further results on the importance of single variables were provided in Section 3.8, with either anticipated or potentially novel and useful results, especially regarding particular terrain attributes. This approach can thus readily be used to assess further variables' importance (such as novel sensor data) or even to reverse-engineer proprietary fertilization strategies.

5.3 Results of the Management Zone Delineation Approach

Chapter 4 revolved around management zone delineation. Based on a detailed literature review of existing management zone delineation (MZD) approaches, mainly using the major precision agriculture journals, there are a few shortcomings that existing approaches have. Nevertheless, the requirements are typically stated quite clearly. Those requirements were collected and grouped accordingly. Based on these requirements and the assumption that MZD is essentially a clustering problem on spatial data sets, the existing algorithms in the area of data mining were reviewed and shortly evaluated as to their usefulness in the context of MZD. It was found that none of the existing algorithms fulfill the special requirements.

MZD, as stated above, is viewed as a spatial clustering problem: given a data set of georeferenced data vectors, find zones (clusters) which are (mostly) contiguous with the property that the similarity of data vectors within a zone is high, while the similarity between zones is low. To solve this problem, a novel algorithm HACC-SPATIAL was devised, which is based on hierarchical agglomerative clustering and incorporates a spatial constraint. The underlying idea is rather simple: to find the aforementioned zones, the algorithm starts initially with each data vector as a single cluster. In each subsequent step, the two clusters which are most similar are merged. However, to achieve spatial contiguity, only those clusters may be merged which are (in addition to being similar in feature space) geospatially adjacent. Since the requirement is for the zones to be mostly contiguous rather than strictly contiguous, a user-definable threshold has been introduced. It simply determines when HACC-SPATIAL switches off the spatial contiguity constraint.

On the one hand, HACC-SPATIAL can behave like a standard hierarchical agglomerative clustering algorithm, by switching off the constraint. On the other hand, it can strictly enforce spatial contiguity, which may, however, be too strict and lead to undesirable results. By setting the contiguity threshold accordingly, the user can determine the behaviour of HACC-SPATIAL and compare different results.

Nevertheless, the pure version of HACC-SPATIAL may lead to non-smooth maps, which do not directly convey novel information to an agriculture expert. Furthermore, the spatial data sets typically exhibit spatial autocorrelation. Therefore, the idea was to exploit this spatial autocorrelation by including an initial clustering phase before the actual HACC-SPATIAL algorithm, which creates contiguous initial zones. This was a heuristic assumption, but typically lead to much smoother maps which directly allowed for discovering novel information and knowledge about the underlying processes on the site.

Numerous examples and trial runs of HACC-SPATIAL demonstrated its ability to find contiguous zones. The effects of the spatial contiguity threshold were presented and the outcomes of using an initial clustering were laid out.

List of Algorithms

1	<i>k</i> -means, adapted from [MacQueen, 1967].	36
2	<i>k</i> -Nearest-Neighbor Regression, adapted from [Mitchell, 1997]	40
3	Random forest algorithm	45
4	Spatial Variable Importance	49
5	HACC-SPATIAL	114

List of Tables

3.1	Results for cross-validation on F440/F611	46
3.2	Regression formulae for the PA data sets	50
3.3	F440, SVI results, starplots	53
3.4	F440sorte1, SVI results, starplots	54
3.5	F440sorte2, SVI results, starplots	55
3.6	F550 w/o YIELD2003, SVI results, starplots	57
3.7	F550 w/ YIELD2003, SVI results, starplots	58
3.8	F610, SVI results, starplots	60
3.9	F611, SVI results, starplots	62
3.10	F631, SVI results, starplots	64
3.11	Model rankings for the spatial cross-validation approach	66
4.1	Overview of HACC-SPATIAL experiments	115
A.1	F440, descriptive statistics, correlation coefficients with YIELD	176
A.2	F550, descriptive statistics, correlation coefficients with YIELD	181
A.3	F610, descriptive statistics, correlation coefficients with YIELD	184
A.4	F611, descriptive statistics, correlation coefficients with YIELD	188
A.5	F631, descriptive statistics, correlation coefficients with YIELD	192
B.1	Overview of RMSE/SVI figures per data set	193
C.1	Regression models, details for the R packages	241

List of Figures

1.1	Data Mining cycle in PA	3
2.1	Growing stages of winter wheat	9
2.2	REIP shifting	12
2.3	Overview on the fields of study	17
2.4	Temporal relationships in data sets	22
2.5	Moran scatter plots	25
2.6	Variograms for two predictors on two fields	27
2.7	Contour maps for F440	28
3.1	Chapter outline	31
3.2	Cross-validation and regression setup	34
3.3	Comparison of grid vs. voronoi approach	36
3.4	k -means clustering on F440	37
3.5	Schematic of a feedforward neural network	42
3.6	Schematic of a single perceptron unit	43
3.7	Spatial variable importance approach for regression	47
3.8	Effect of SORTe variable in F440	51
4.1	Exploratory spatial clustering for MZD, example	72
4.2	Spatially coherent regions from [Lark, 1998]	75
4.3	Management Zones from [Franzen and Nanna, 2006]	76
4.4	Management Zones from [Derby et al., 2007]	78
4.5	Productivity zones from [Kitchen et al., 2005]	81
4.6	Management zones from [Xin-Zhong et al., 2009]	87
4.7	Management zone approaches surveyed	90
4.8	Management zones compared with yield levels	90
4.9	Spatial Objects for Clustering with DBSCAN	94
4.10	STING algorithm result for 2D spatial data	96
4.11	Spatial Objects for Clustering with CLARANS	97
4.12	Spatial Proximity in AMOEBA	98
4.13	Spatial Objects for Clustering with AMOEBA	99
4.14	Region Growing on PA data	100
4.15	Clustering with ICEAGE	101
4.16	Regionalization with SKATER	102

4.17	Regionalization with REDCAP	103
4.18	F550, tessellation based on geographic location	109
4.19	Single linkage on spatial data	110
4.20	Mean distances during clustering, contiguity threshold 0.5	112
4.21	Mean distances during clustering, contiguity threshold 0.8	113
4.22	HACC-SPATIAL on F611 using EC25	116
4.23	HACC-SPATIAL on F440, REIP49, different thresholds	117
4.24	HACC-SPATIAL on F631 using EC25	118
4.25	HACC-SPATIAL on F440, YIELD07, $ct \in \{0.5, 1.0\}$	119
4.26	HACC-SPATIAL on F611, using REIP32/REIP49	121
4.27	F550, effect of spatial contiguity constraint	122
4.28	HACC-SPATIAL on F610 using EC25	123
4.29	HACC-SPATIAL on F611 using EC25	125
4.30	HACC-SPATIAL on F440 using REIP49	126
4.31	HACC-SPATIAL on F631 using EC25	127
4.32	HACC-SPATIAL on F440 using REIP32	128
4.33	HACC-SPATIAL on F440 using N1	130
4.34	HACC-SPATIAL on F611, using REIP32/REIP49	131
4.35	F550, soil sampling variables	132
4.36	HACC-SPATIAL on F550 using soil sampling variables	133
4.37	HACC-SPATIAL on F610 using EC25	135
A.1	F440: N1,N2,N3, EC25	174
A.2	F440: REIP32, REIP49, YIELD07	175
A.3	F550: YIELD03, YIELD04, YIELD07	177
A.4	F550: REIP32, REIP49, EC25	178
A.5	F550: N1,N2,N3	179
A.6	F550: pH, P, Mg, K	180
A.7	F610: YIELD07, EC25	182
A.8	F610: N1,N2,N3	183
A.9	F611: REIP32, REIP49	185
A.10	F611: EC25, YIELD07	186
A.11	F611: N1,N2,N3	187
A.12	F631: YIELD07, EC25	189
A.13	F631: N1,N2	190
A.14	F631: N3	191
B.1	RMSE for F440 and its subsets (by strategy)	194
B.2	F440, all strategies, models and SVI	195
B.3	F440, strategy “low, constant”, models and SVI	196
B.4	F440, strategy “constant”, models and SVI	197
B.5	F440, strategy “neural network”, models and SVI	198
B.6	F440, strategy “sensor”, models and SVI	199
B.7	RMSE for F440sorte1 and its subsets (by strategy)	200
B.8	F440sorte1, all strategies, models and SVI	201

B.9	F440sorte1, strategy “low, constant”, models and SVI	202
B.10	F440sorte1, strategy “constant”, models and SVI	203
B.11	F440sorte1, strategy “neural network”, models and SVI	204
B.12	F440sorte1, strategy “sensor”, models and SVI	205
B.13	RMSE for F440sorte2 and its subsets (by strategy)	206
B.14	F440sorte2, all strategies, models and SVI	207
B.15	F440sorte2, strategy “low, constant”, models and SVI	208
B.16	F440sorte2, strategy “constant”, models and SVI	209
B.17	F440sorte2, strategy “neural network”, models and SVI	210
B.18	F440sorte2, strategy “sensor”, models and SVI	211
B.19	RMSE for F550 and its subsets (by strategy, without YIELD2003)	212
B.20	F550, all strategies, without YIELD2003, models and SVI	213
B.21	F550, strategy “company”, without YIELD2003, models and SVI	214
B.22	F550, strategy “mapping”, without YIELD2003, models and SVI	215
B.23	F550, strategy “N-trial”, without YIELD2003, models and SVI	216
B.24	F550, strategy “sensor”, without YIELD2003, models and SVI	217
B.25	RMSE for F550 and its subsets (by strategy, with YIELD2003)	218
B.26	F550, all strategies, with YIELD2003, models and SVI	219
B.27	F550, strategy “company”, with YIELD2003, models and SVI	220
B.28	F550, strategy “mapping”, with YIELD2003, models and SVI	221
B.29	F550, strategy “N-trial”, with YIELD2003, models and SVI	222
B.30	F550, strategy “sensor”, with YIELD2003, models and SVI	223
B.31	RMSE for F610 and its subsets (by strategy)	224
B.32	F610, all strategies, models and SVI	225
B.33	F610, strategy “constant”, models and SVI	226
B.34	F610, strategy “neural network”, models and SVI	227
B.35	F610, strategy “N-trial”, models and SVI	228
B.36	F610, strategy “sensor 1”, models and SVI	229
B.37	F610, strategy “sensor 2”, models and SVI	230
B.38	RMSE for F611 and its subsets (by strategy)	231
B.39	F611, all strategies, models and SVI	232
B.40	F611, strategy “constant”, models and SVI	233
B.41	F611, strategy “neural network”, models and SVI	234
B.42	F611, strategy “sensor”, models and SVI	235
B.43	RMSE for F631 and its subsets (by strategy)	236
B.44	F631, all strategies, models and SVI	237
B.45	F631, strategy “constant”, models and SVI	238
B.46	F631, strategy “neural network”, models and SVI	239
B.47	F631, strategy “N-trial”, models and SVI	240

List of Abbreviations

AK	Available Potassium
AN	Available Nitrogen
ANN	Artificial Neural Network
AP	Available Phosphorus
CEC	Cation Exchange Capacity
DEM	Digital Elevation Model
DM	Data Mining
EC_a	(Apparent) Electrical Conductivity
GAM	Generalized Additive Model
GIS	Geographical Information System
LM	Linear Model
ML	Machine Learning
MZD	Management Zone Delineation
NDVI	Normalized Difference Vegetation Index
NN	Neural Network
OM	Organic Matter
PA	Precision Agriculture
pH	pH Value
SSM	Site-Specific (Crop) Management
SVM	Support Vector Machine
SVR	Support Vector Regression

Bibliography

- V. Adamchuk, J. Hummel, M. Morgan, and S. Upadhyaya. On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, 44(1):71 – 91, 2004. 13
- V. I. Adamchuk and J. Mulliken. Ec05-705 precision agriculture: Site-specific of soil ph. Technical report, University of Nebraska, Lincoln, Nebraska, USA, 2005. 13
- W. Adler, A. Brenning, S. Potapov, M. Schmid, and B. Lausen. Ensemble classification of paired data. *Computational Statistics & Data Analysis*, 55(5):1933 – 1941, 2011. ISSN 0167-9473. 65
- C. Aggarwal and P. Yu. Data mining techniques for associations, clustering and classification. In *PAKDD '99: Proceedings of the Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining*, pages 13–23. Springer-Verlag, 1999. ISBN 3-540-65866-1. 5
- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD '98: Proc. of the 1998 ACM SIGMOD Int. Conf. on Management of Data*, pages 94–105, New York, NY, USA, 1998. ACM. ISBN 0-89791-995-5. 95
- I. Alcantarilla, N. Zarraoa, and J. Caro. On EGNOS and WAAS Performance. In *Proceedings of the 61st Annual Meeting of The Institute of Navigation*, pages 774–782, Cambridge, MA, 2005. The Institute of Navigation. 10
- M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In A. Delis, C. Faloutsos, and S. Ghandeharizadeh, editors, *SIGMOD Conference*, pages 49–60. ACM Press, 1999. ISBN 1-58113-084-8. 94
- L. Anselin. *Spatial Econometrics*, pages 310–330. Basil Blackwell, Oxford, 2001. 23
- L. Anselin, R. Bongiovanni, and J. Lowenberg-DeBoer. A spatial econometric approach to the economics of site-specific nitrogen management in corn production. *American Journal of Agricultural Economics*, 86:675–687, 2004. 18
- K. J. Archer and R. V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249 – 2260, 2008. ISSN 0167-9473. 65

- R. M. Assuncao, M. C. Neves, G. Camara, and C. D. C. Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006. 101, 102
- L. Auret and C. Aldrich. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105(2):157 – 170, 2011. ISSN 0169-7439. 65
- M. Baatz and A. Schäpe. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. In *Angewandte Geographische Informationsverarbeitung, AGIT-Symposium Salzburg*, volume 200, pages 12–23, Karlsruhe, 2000. Herbert Wichmann Verlag. 85
- M. Bachmaier and M. Gandorfer. A conceptual framework for judging the precision agriculture hypothesis with regard to site-specific nitrogen application. *Precision Agriculture*, 10(2):95–110, April 2009. 32
- S. Begiebing, M. Schneider, H. Bach, and P. Wagner. Assessment of in-field heterogeneity for determination of the economic potential of precision farming. In J. Stafford, editor, *Proceedings of the 7th European Conference on Precision Agriculture*, pages 811–818, Netherlands, 2007. Wageningen Academic Publishers. 6
- L. Bel, D. Allard, J. Laurent, R. Cheddadi, and A. Bar-Hen. Cart algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics & Data Analysis*, 53(8):3082 – 3093, 2009. ISSN 0167-9473. doi: 10.1016/j.csda.2008.09.012. 65
- K. J. Beven and M. J. Kirkby. A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Journal*, 24(1):43–69, 1979. 21
- J. Biermacher, F. Epplin, B. Brorsen, J. Solie, and W. Raun. Economic feasibility of site-specific optical sensing for managing nitrogen fertilizer for growing wheat. *Precision Agriculture*, 10(3):213–230, June 2009. 15, 32
- R. S. Bivand, E. J. Pebesma, and V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. Use R. Springer, New York, 2008. ISBN 978-0-387-78170-9. 28
- J. Böhner, R. Köthe, O. Conrad, J. Gross, A. Ringeler, and T. Selige. *Soil regionalisation by means of terrain analysis and process parameterisation*, pages 213–222. European Union, EUR 20398, 2002. 21
- R. G. Bongiovanni, C. W. Robledo, and D. M. Lambert. Economics of site-specific nitrogen management for protein content in wheat. *Comput. Electron. Agric.*, 58(1):13–24, August 2007. ISSN 0168-1699. 15
- B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992. 43
- L. Breiman. Bagging predictors. Technical report, Department of Statistics, Univ. of California, Berkeley, 1994. 44

- L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 44
- L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 10 2001. 44, 45
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. 5, 41
- A. Brenning. Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Science*, 5(6):853–862, 2005. 43
- A. Brenning. Benchmarking classifiers to optimally integrate terrain analysis and multispectral remote sensing in automatic rock glacier detection. *Remote Sensing of Environment*, 113(1):239–247, January 2009. ISSN 00344257. 65, 66
- A. Brenning and B. Lausen. Estimating error rates in the classification of paired organs. *Statistics in Medicine*, 27(22):4515–4531, 2008. 65
- A. Brenning, K. Kaden, and S. Itzerott. Comparing classifiers for crop identification based on multitemporal Landsat TM/ETM data. In *Proceedings of the 2nd workshop of the EARSeL Special Interest Group Remote Sensing of Land Use and Land Cover*, pages 64–71, September 2006. 65
- A. Brenning, H. Piotraschke, and P. Leithold. Geostatistical analysis of on-farm trials in precision agriculture. In J. M. Ortiz and X. Emery, editors, *GEOSTATS 2008, Proceedings of the Eighth International Geostatistics Congress*, volume 2, pages 1131–1136, 12 2008. 18
- A. Brock, S. M. Brouder, G. Blumhoff, and B. S. Hofmann. Defining yield-based management zones for corn-soybean rotations. *Agronomy Journal*, 97(4):1115–1128, 2005. 67, 82
- P. Bühlmann. Bootstraps for time series. Technical report, ETH Zürich, 2001. 65
- J. Byfuglien and A. Nordgård. Region-building – a comparison of methods. *Norsk Geografisk Tidsskrift*, 27(2):127–151, 1973. 105
- H. Cetin, J. Pafford, and T. Mueller. Precision agriculture using hyperspectral remote sensing and gis. In *Proceedings of 2nd International Conference on Recent Advances in Space Technologies, 2005 (RAST 2005)*, pages 70 – 77, june 2005. 11
- M. W. Chang, B. J. Chen, and C. J. Lin. Eunate network competition: Electricity load forecasting. Technical report, National Taiwan University, 2001. 43
- J. Choo, R. Jiamthaphaksin, C.-s. Chen, O. Celepcikay, C. Giusti, and C. Eick. Mosaic: A proximity graph approach for agglomerative clustering. In *Data Warehousing and Knowledge Discovery*, pages 231–240, 2007. 99
- I. Clark. *Practical Geostatistics*. Elsevier Applied Science, London, 1979. 26
- S. Clemencon, M. Depecker, and N. Vayatis. Adaptive partitioning schemes for bipartite ranking. *Machine Learning*, 83:31–69, 2011. ISSN 0885-6125. 65

- W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, sep 1988. 40
- R. N. Colwell. *History and Place of Photographic Interpretation*, pages 3–47. American Society for Photogrammetry and Remote Sensing (ASPRS), 2nd edition, 1997. 10
- D. L. Corwin and S. M. Lesch. Application of soil electrical conductivity to precision agriculture: Theory, principles, and guidelines. *Agronomy Journal*, 95(3):455–471, May 2003. 13
- D. L. Corwin and R. E. Plant. Applications of apparent soil electrical conductivity in precision agriculture. *Computers and Electronics in Agriculture*, 46(1-3):1 – 10, 2005. ISSN 0168-1699. Applications of Apparent Soil Electrical Conductivity in Precision Agriculture. 13
- D. L. Corwin, S. M. Lesch, P. J. Shousea, R. Soppeb, and J. E. Ayars. Identifying soil properties that influence cotton yield using soil sampling directed by apparent soil electrical conductivity. *Agronomy Journal*, 95:352–364, 2003. 13
- N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993. 23, 26
- J. J. Daniels. Ground penetrating radar fundamentals. Technical report, The Ohio State University, 2000. 13
- M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1: 131–156, 1997. 47
- I. Davidson and S. Ravi. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In A. Jorge, editor, *Proceedings of PKDD 2005*, volume 3721 of *Lecture Notes in Artificial Intelligence*, pages 59–70, Berlin, Heidelberg, 2005. Springer. 106
- M. J. de Smith, M. F. Goodchild, and P. A. Longley. *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*. Troubador Publishing Ltd, The Winchelsea Press, Leicester, UK, 3rd edition, 2009. 6, 20, 24
- N. E. Derby, F. X. M. Casey, and D. W. Franzen. Comparison of nitrogen management zone delineation methods for corn grain yield. *Agronomy Journal*, 99:405–414, 2007. 76, 78, 147
- A. Dobermann, J. L. Ping, V. I. Adamchuk, G. C. Simbahan, and R. B. Ferguson. Classification of Crop Yield Variability in Irrigated Production Fields. *Agronomy Journal*, 95 (5):1105–1120, 2003. 84
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, and U. M. Fayyad, editors, *2nd Int. Conf. on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996. 94

- M. Ester, H.-P. Kriegel, and J. Sander. Knowledge discovery in spatial databases. In *23rd German Conference on AI (KI99)*, volume 1701 of *LNCS*, pages 61–74. Springer, 1999. 6
- V. Estivill-Castro and I. Lee. AMOEBA: Hierarchical clustering based on spatial proximity using delaunay diagram. In *Proc. of the Ninth International Symposium on Spatial Data Handling*, pages 7a.26–7a.41, Beijing, China, 2000. 96
- V. Estivill-Castro and I. Lee. Multi-level clustering and its visualization for exploratory spatial analysis. *GeoInformatica*, 6(2):123–152, June 2002. 91, 92, 98, 99
- I. S. Evans, T. Hengl, and P. Gorsevski. Applications in geomorphology. In Hengl and Reuter [2009], pages 497–525. 14
- M. H. Ezrin, M. S. M. Amin, A. R. Anuar, and W. Aimrun. Rice yield prediction using apparent electrical conductivity of paddy soils. *European Journal of Scientific Research*, 37(4):575–590, 2009. 13
- FAO Trade and Market Division. Food outlook, global market analysis. Technical report, Food and Agriculture Organization of the United Nations (FAO), 2010. 9
- A. Farag and R. M. Mohamed. Regression using support vector machines. Technical report, University of Louisville, 2004. 44
- U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996a. 2, 5, 6
- U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39:27–34, November 1996b. ISSN 0001-0782. 5, 6
- M. M. Fischer. Regional taxonomy : A comparison of some hierarchic and non-hierarchic strategies. *Regional Science and Urban Economics*, 10(4):503 – 537, 1980. 105
- D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, 1987. 107
- A. S. Fotheringham, M. Charlton, and C. Brunson. *Geographically Weighted Regression: the analysis of spatially varying relationships*. Wiley, New York, 2002. 40, 65
- D. W. Franzen and T. Nanna. Comparison of nitrogen management zone delineation methods. In *Proceedings of the North Central Extension-Industry Soil Fertility Conference*, volume 19, Des Moines, IA, 2003. 76
- D. W. Franzen and T. Nanna. Use of data layering to address changes in nitrogen management zone delineation. In *USDA Forest Service Proceedings*, 2006. 76, 77, 147
- J. J. Fridgen, N. R. Kitchen, K. A. Sudduth, S. T. Drummond, W. J. Wiebold, and C. W. Fraisse. Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation. *Agronomy Journal*, 96(1):100–108, 2004. 80

- H. Frohlich and A. Zell. Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. In *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, volume 3, pages 1431–1436, jul 2005. doi: 10.1109/IJCNN.2005.1556085. 66
- K. R. Gabriel and R. R. Sokal. A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18(3):259–278, 1969. 99
- C. M. Gold and P. R. Remmele. Voronoi methods in GIS. In *Algorithmic Foundations of Geographic Information Systems*, pages 21–35, London, UK, 1997. Springer. 108
- M. Goodchild, L. Anselin, R. Appelbaum, and B. Harthorn. Toward spatially integrated social science. *International Regional Science Review*, 23:139–159, 2000. 23
- A. D. Gordon. A survey of constrained classification. *Computational Statistics & Data Analysis*, 21:17–29, 1996. 105
- GRASS Development Team. *Geographic Resources Analysis Support System (GRASS GIS) Software*. Open Source Geospatial Foundation, USA, 2010. URL <http://grass.osgeo.org>. 11
- M. S. Grewal, L. R. Weill, and A. P. Andrews. *Global Positioning Systems, Inertial Navigation and Integration*. Wiley & Sons, New York, USA, 2001. 10
- D. A. Griffith. *Spatial Autocorrelation and Spatial Filtering*. Advances in Spatial Science. Springer, New York, 2003. ISBN 978-3-540-00932-0. 23
- S. Gruber and S. Peckham. Land-surface parameters and objects in hydrology. In Hengl and Reuter [2009], pages 171–194. 14
- F. Guastafarro, A. Castrignano, D. D. Benedetto, D. Sollitto, A. Troccoli, and B. Cafarelli. A comparison of different algorithms for the delineation of management zones. *Precision Agriculture*, 11(6):600–620, 2010. 74
- S. Gunn. Support vector machines for classification and regression. Technical Report, School of Electronics and Computer Science, University of Southampton, Southampton, U.K., 1998. 43, 44
- D. Guo. Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). *International Journal of Geographical Information Science*, 22(7): 801–823, 2008. 102, 103
- D. Guo and J. Mennis. Spatial data mining and geographic knowledge discovery – an introduction. *Computers, Environment and Urban Systems*, 33(6):403–408, 2009. 6
- D. Guo, D. Peuquet, and M. Gahegan. Opening the black box: interactive hierarchical clustering for multivariate spatial patterns. In *GIS '02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pages 131–136, New York, NY, USA, 2002. ACM. 100

- D. Guo, D. J. Peuquet, and M. Gahegan. ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. *Geoinformatica*, 7(3):229–253, 2003. 100, 101
- G. Guyot, F. Baret, and D. J. Major. High spectral resolution: Determination of spectral shifts between the red and the near infrared. *International Archives of Photogrammetry and Remote Sensing*, 11:750–760, 1988. 11
- M. T. Hagan. *Neural Network Design (Electrical Engineering)*. Thomson Learning, December 1995. ISBN 0534943322. 42
- J. Han, M. Kamber, and A. K. H. Tung. Spatial clustering methods in data mining: A survey. In H. Miller and J. Han, editors, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001. 105
- D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining (Adaptive Computation and Machine Learning)*. The MIT Press, August 2001. ISBN 026208290X. 32
- J. W. Hansen. Realizing the potential benefits of climate prediction to agriculture: issues, approaches, challenges. *Agricultural Systems*, 74:309–330, 2002. 16
- T. J. Hastie. *Generalized additive models*, chapter Generalized Additive Models. Wadsworth & Brooks/Cole, 1991. 39
- S. Haykin. *Neural Networks: A Comprehensive Foundation (2nd Edition)*. Prentice Hall, July 1998. ISBN 0132733501. 42
- H. Heege, S. Reusch, and E. Thiessen. Prospects and results for optical systems for site-specific on-the-go control of nitrogen-top-dressing in germany. *Precision Agriculture*, 9(3):115–131, June 2008. 11, 12
- T. Hengl and H. I. Reuter, editors. *Geomorphometry*, volume 33 of *Developments in Soil Science*. Elsevier, Amsterdam, 2009. 15, 157, 158, 162, 163, 164, 165
- A. Hornung, R. Khosla, R. Reich, D. Inman, and D. G. Westfall. Comparison of site-specific management zones: Soil-color-based and yield-based. *Agronomy Journal*, 98(2):407–415, 2006. 86
- C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification. Technical report, National Taiwan University, Taipeh 106, Taiwan, 2008. 44
- J. Hüter, F. Kloepfer, and U. Klöble. *Elektronik, Satelliten und Co. Precision Farming in der Praxis*. KTBL, KTBL e.V. Darmstadt, Germany, 2005. 8
- J. Iqbal, J. J. Read, A. J. Thomasson, and J. N. Jenkins. Relationships between soil-landscape and dryland cotton lint yield. *Soil Science Society of America*, 69(3): 872–882, 2005. 68
- O. Ivanciuc. *Applications of Support Vector Machines in Chemistry*, volume 23, pages 291–400. Wiley-VCH, Weinheim, 2007. 43

- G. M. Jacquez. Spatial cluster analysis. In S. Fotheringham and J. Wilson, editors, *The Handbook of Geographic Information Science*, pages 395–416. Blackwell Publishing, 2008. 73
- A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Survey*, 31(3):264–323, 1999. ISSN 0360-0300. 109
- D. Jaynes, T. Kasper, T. Colvin, and D. James. Cluster analysis of spatiotemporal corn yield patterns in an iowa field. *Agronomy Journal*, 95(3):574–586, 2003. 77
- D. B. Jaynes, T. S. Colvin, and T. C. Kaspar. Identifying potential soybean management zones from multi-year yield data. *Computers and Electronics in Agriculture*, 46(1–3):309–327, 2005. ISSN 0168-1699. Applications of Apparent Soil Electrical Conductivity in Precision Agriculture. 67, 77
- J. R. Jensen. *Remote Sensing of the Environment: An Earth Resource Perspective*. Prentice Hall, 2nd edition, 2006. 10
- Á. Jiménez, J. Lázaro, and J. Dorronsoro. Finding optimal model parameters by discrete grid search. In E. Corchado, J. Corchado, and A. Abraham, editors, *Innovations in Hybrid Intelligent Systems*, volume 44 of *Advances in Soft Computing*, pages 120–127. Springer Berlin / Heidelberg, 2007. 66
- M. Kanevski, V. Timonin, and A. Pozdnukhov. *Machine Learning for Spatial Environmental Data*. Psychology Press, Taylor and Francis Group, London, UK, 2009. 65
- T. C. Kaspar, T. S. Colvin, D. B. Jaynes, D. L. Karlen, D. E. James, and D. W. Meek. Relationship between six years of corn yields and terrain attributes. *Precision Agriculture*, 4:87–101, 2003. 68, 77
- L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Interscience, New York, 1990. 73
- M. Khanna, O. F. Epouhe, and R. Hornbaker. Site-specific crop management: Adoption patterns and incentives. *Review of Agricultural Economics*, 21(2):455–472, 1999. 8
- R. Khosla, D. Inman, D. G. Westfall, R. M. Reich, M. Frasier, M. Mzuku, B. Koch, and A. Hornung. A synthesis of multi-disciplinary research in precision agriculture: site-specific management zones in the semi-arid western Great Plains of the USA. *Precision Agriculture*, 9:85–100, 2008. 73, 86
- R. Khosla, D. G. Westfall, R. M. Reich, J. S. Mahal, and W. J. Gangloff. Spatial variation and site-specific management zones. In Oliver [2010b], pages 195–219. 89, 90
- J. A. King, P. M. R. Dampney, R. M. Lark, H. C. Wheeler, R. I. Bradley, and T. R. Mayr. Mapping potential crop management zones within fields: Use of yield-map series and patterns of soil physical properties identified by electromagnetic induction sensing. *Precision Agriculture*, 6:167–181, 2005. 67, 81, 82

- N. Kitchen, K. Sudduth, D. Myers, S. Drummond, and S. Hong. Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. *Computers and Electronics in Agriculture*, 46(1-3):285 – 308, 2005. ISSN 0168-1699. Applications of Apparent Soil Electrical Conductivity in Precision Agriculture. 13, 80, 81, 147
- D. Klein, S. D. Kamvar, and C. D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 307–314, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. 107
- A. Knudby, A. Brenning, and E. LeDrew. New approaches to modelling fish-habitat relationships. *Ecological Modelling*, 221:503–511, 2010a. 47, 66
- A. Knudby, E. LeDrew, and A. Brenning. Predictive mapping of reef fish species richness, diversity and biomass in zanzibar using IKONOS imagery and machine-learning techniques. *Remote Sensing of Environment*, 114:1230–1241, June 2010b. 43
- K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *SSD '95: Proceedings of the 4th International Symposium on Advances in Spatial Databases*, pages 47–66, London, UK, 1995. Springer-Verlag. 138
- S. B. Kotsiantis and P. E. Pintelas. Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications*, 1:73–81, 2004. 105
- A. N. Kravchenko and D. G. Bullock. Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal*, 92:75–83, 2000. 68
- J. Kumhálová, F. Kumhála, M. Kroulík, and Štěpánka Matějková. The impact of topography on soil properties and yield and the effects of weather conditions. *Precision Agriculture*, 12:813–830, 2011. 68
- D. M. Lambert, J. Lowenberg-DeBoer, and R. Bongiovanni. Spatial regression models for yield monitor data: a case study from argentina. In *Proceedings of the Agricultural Economics Association Annual Meeting*. Agricultural Economics Association, Montreal, Canada, 2003. 18
- J. L. Lambert. Editorial. *Computers and Electronics in Agriculture*, 1:1–4, 1985. 7
- P. Langley. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, pages 140–144. AAAI Press, 1994. 47
- R. M. Lark. Forming spatially coherent regions by classification of multi-variate data: an example from the analysis of maps of crop yield. *International Journal of Geographical Information Science*, 12(1):83–98, 1998. 74, 75, 147
- R. M. Lark, J. V. Stafford, and H. C. Bolam. Limitations on the spatial resolution of yield mapping for combinable crops. *Journal of Agricultural Engineering Research*, 66(3):183 – 193, 1997. ISSN 0021-8634. 14

- S. F. Ledgard and K. E. Giller. *Atmospheric N₂ Fixation as an Alternative N Source*, chapter 12, pages 443–486. Nitrogen Fertilization in the Environment. Marcel Dekker, New York, 1995. 14
- V. Lemaire and F. Clairot. An input variable importance definition based on empirical data probability distribution. In I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh, editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 509–516. Springer Berlin / Heidelberg, 2006. ISBN 978-3-540-35487-1. 65
- X. Li and C. Claramunt. A spatial entropy-based decision tree for classification of geographical information. *Transactions in GIS*, 10(3):451–467, 2006. 110
- Y. Li, Z. Shi, F. Li, and H.-Y. Li. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. *Comput. Electron. Agric.*, 56(2):174–186, 2007. ISSN 0168-1699. 83
- J. Liu, J. R. Miller, D. Haboudane, and E. Pattey. Exploring the relationship between red edge parameters and crop variables for precision agriculture. In *2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 2, pages 1276–1279, 2004. 11
- L. Ma, L. R. Ahuja, and T. W. Bruulsema. *Quantifying and understanding plant nitrogen uptake for systems modeling*. CRC Press, 2009. 14
- R. A. MacMillan and P. A. Shary. Landforms and landform elements in geomorphometry. In Hengl and Reuter [2009], pages 227–254. 14
- J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, 1967. University of California Press. 36, 143
- C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. 110
- C. R. Margules, D. P. Faith, D. Belbin, and L. Belbin. An adjacency constraint in agglomerative hierarchical classifications of geographic data. *Environment & Planning A*, 17: 397–412, 1985. 106, 108
- G. Matheron. Principles of geostatistics. *Economic Geology*, 58(8):1246–1266, 1963. 25
- U. Meier. *Entwicklungsstadien mono- und dikotyler Pflanzen*. Biologische Bundesanstalt für Land- und Forstwirtschaft, Braunschweig, Germany, 2001. 9
- I. Mejía-Guevara and A. Kuri-Morales. Evolutionary feature and parameter selection in support vector regression. In *Lecture Notes in Computer Science*, volume 4827, pages 399–408. Springer, Berlin, Heidelberg, 2007. 43, 44
- Y. Miao, D. J. Mulla, and P. C. Robert. Identifying important factors influencing corn yield and grain quality variability using artificial neural networks. *Precision Agriculture*, 7:117–135, 2006. 50, 68

- E. M. Middleton, P. K. E. Campbell, J. E. McMurtrey, L. A. Corp, L. M. Butcher, and E. W. Chappelle. “Red edge” optical properties of corn leaves from different nitrogen regimes. In *2002 IEEE International Geoscience and Remote Sensing Symposium*, volume 4, pages 2208–2210, 2002. 11
- H. J. Miller and J. Han, editors. *Geographic Data Mining and Knowledge Discovery*. CRC / Taylor & Francis, 2009. 6
- T. M. Mitchell. *Machine Learning*. McGraw-Hill Science/Engineering/Math, March 1997. ISBN 0070428077. 32, 39, 40, 41, 143
- E. R. C. Morales and Y. Y. Mendizabal. A new contiguity-constrained agglomerative hierarchical clustering algorithm for image segmentation. In P. Meseguer, L. Mandow, and R. M. Gasca, editors, *Proc. of CAEPIA 2009*, volume 5988 of *LNAI*, pages 261–270, Berlin, Heidelberg, 2010. Springer. 106
- P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37:17–33, 1950. 24
- G. Mountrakis, J. Im, and C. Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66:247–259, 2011. 65
- A. T. Murray and T.-K. Shyy. Integrating attribute and space characteristics in choropleth display and spatial data mining. *International Journal of Geographical Information Science*, 14(7):649–667, 2000. 134
- D. Nauck, C. Borgelt, F. Klawonn, and R. Kruse. *Neuro-Fuzzy-Systeme — Von den Grundlagen Neuronaler Netze zu modernen Fuzzy-Systemen*. Vieweg, Wiesbaden, Germany, 2003. 42, 47
- J. J. Neeteson. *Nitrogen Management for Intensively Grown Arable Crops and Field Vegetables*, chapter 7, pages 295–326. CRC Press, Haren, The Netherlands, 1995. ISBN 978-0-824789947. 19
- A. Nelson, H. I. Reuter, and P. Gessler. DEM production methods and sources. In Hengl and Reuter [2009], pages 65–85. 10
- R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *VLDB’94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 144–155. Morgan Kaufmann Publishers Inc., 1994. 96, 97
- R. T. Ng and J. Han. Clarans: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5):1003–1016, 2002. 95
- K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(110), 2010. 65
- V. Olaya. Basic land-surface parameters. In Hengl and Reuter [2009], pages 141–169. 14

- V. Olaya and O. Conrad. Geomorphometry in SAGA. In Hengl and Reuter [2009], pages 293–308. 20
- J. D. Olden, M. K. Joy, and R. G. Death. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178(3-4):389 – 397, 2004. ISSN 0304-3800. 65
- M. A. Oliver. An overview of geostatistics and precision agriculture. In Oliver [2010b], pages 1–34. 25, 26
- M. A. Oliver, editor. *Geostatistical Applications for Precision Agriculture*. Springer, 1st edition, 2010b. 160, 164
- S. Oppel and F. Huettmann. Using a random forest model and public data to predict the distribution of prey for marine wildlife management. In *Spatial Complexity, Informatics, and Wildlife Conservation*, pages 151–163. Springer, 2009. 65
- R. A. Ortega and O. A. Santibáñez. Determination of management zones in corn (*zea mays* l.) based on soil fertility. *Computers and Electronics in Agriculture*, 58(1):49 – 59, 2007. ISSN 0168-1699. Precision Agriculture in Latin America. 83, 84
- A. Papritz and A. Stein. Spatial prediction by linear kriging. In A. Stein, F. van der Meer, and B. Gorte, editors, *Spatial Statistics for Remote Sensing*, chapter 6, pages 83–113. Kluwer Academic Publishers, Dordrecht, Boston, London, 1999. 16
- J. M. Pena-Barragan, F. Lopez-Granados, M. Jurado-Exposito, and L. Garcia-Torres. Sunflower yield related to multi-temporal aerial photography, land elevation and weed infestation. *Precision Agriculture*, 11(5):568–585, 2010. 68
- C. Perruchet. Constrained agglomerative hierarchical classification. *Pattern Recognition*, 16:213–217, 1983. 106
- A. Persson, P. Pilesjö, and L. Eklundh. Spatial influence of topographical factors on yield of potato (*solanum tuberosum* l.) in Central Sweden. *Precision Agriculture*, 6(4):341–357, 2005. 68
- C.-G. Pettersson, M. Söderström, and H. Eckersten. Canopy reflectance, thermal stress, and apparent soil electrical conductivity as predictors of within-field variability in grain yield and grain protein of malting barley. *Precision Agriculture*, 7:343–359, 2006. 67, 68
- R. E. Plant. Site-specific management: the application of information technology to crop production. *Computers and Electronics in Agriculture*, 30(1-3):9 – 29, 2001. 6
- C. Potvin. Anova – experimental layout and analysis. In Scheiner and Gurevitch [2001], pages 63–76. 18
- J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, March 1986. 41
- R. J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., January 1993. 41

- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL <http://www.R-project.org>. ISBN 3-900051-07-0. 111
- W. R. Raun, J. B. Solie, G. V. Johnson, M. L. Stone, R. W. Mullen, K. W. Freeman, W. E. Thomason, and E. V. Lukina. Improving Nitrogen Use Efficiency in Cereal Grain Production with Optical Sensing and Variable Rate Application. *Agronomy Journal*, 94(4):815–820, 2002. 80
- M. Reichardt and C. Jürgens. Adoption and future perspective of precision farming in germany: results of several surveys among different agricultural target groups. *Precision Agriculture*, 10:73–94, 2009. 8
- H. I. Reuter and K.-C. Kersebaum. Applications in precision agriculture. In Hengl and Reuter [2009], pages 623–636. 14
- H. I. Reuter, A. Giebel, and O. Wendroth. Can landform stratification improve our understanding of crop yield variability? *Precision Agriculture*, 6(6):521–537, 2005. 68
- P. Roudier, B. Tisseyre, H. Poilvé, and J.-M. Roger. Management zone delineation using a modified watershed algorithm. *Precision Agriculture*, 9(5):233–250, 2008. 85
- G. Ruß. Data mining of agricultural yield data: A comparison of regression models. In P. Perner, editor, *Advances in Data Mining – Applications and Theoretical Aspects*, volume 5633 of *LNAI*, pages 24–37. Springer, July 2009. 32
- G. Ruß and A. Brenning. Spatial variable importance assessment for yield prediction in precision agriculture. In P. R. Cohen, N. M. Adams, and M. R. Berthold, editors, *Proceedings of IDA2010*, volume 6065 of *LNCS*, pages 184–195, Heidelberg, 2010a. Springer. 46, 69
- G. Ruß and A. Brenning. Data mining in precision agriculture: Management of spatial information. In E. Hüllermeier, R. Kruse, and F. Hoffmann, editors, *Computational Intelligence for Knowledge-Based Systems Design*, volume 6178 of *LNAI*, pages 350–359, Berlin, Heidelberg, 2010b. Springer. 46, 48, 69
- G. Ruß and R. Kruse. Regression models for spatial data: An example from precision agriculture. In P. Perner, editor, *Advances in Data Mining. Applications and Theoretical Aspects*, volume 6171 of *LNAI*, pages 450–463, Berlin, Heidelberg, July 2010. Springer. 46
- G. Ruß and R. Kruse. Exploratory hierarchical clustering for management zone delineation in precision agriculture. In P. Perner, editor, *Proceedings of ICDM 2011*, volume 6870 of *LNAI*, pages 161–173, Berlin, Heidelberg, Aug. 2011a. Springer. 71
- G. Ruß and R. Kruse. Machine learning methods for spatial clustering on precision agriculture data. In A. Kofod-Petersen, F. Heintz, and H. Langseth, editors, *Eleventh Scandinavian Conference on Artificial Intelligence*, *Frontiers in Artificial Intelligence and Applications*, pages 40–49, Amsterdam, Netherlands, May 2011b. IOS Press. 71

- G. Ruß and M. Schneider. Hierarchical spatial clustering for management zone delineation in precision agriculture. In I. Bichindaritz, P. Perner, and G. Ruß, editors, *Advances in Data Mining*, pages 95–104, Leipzig, July 2010. IBAI Publishing. 71
- G. Ruß, R. Kruse, M. Schneider, and P. Wagner. Estimation of neural network parameters for wheat yield prediction. In M. Bramer, editor, *Artificial Intelligence in Theory and Practice II*, volume 276 of *IFIP International Federation for Information Processing*, pages 109–118. Springer Boston, July 2008a. 32, 42, 46
- G. Ruß, R. Kruse, M. Schneider, and P. Wagner. Optimizing wheat yield prediction using different topologies of neural networks. In J. L. Verdegay, M. Ojeda-Aciego, and L. Magdalena, editors, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-08)*, pages 576–582. University of Málaga, June 2008b. 32
- G. Ruß, R. Kruse, P. Wagner, and M. Schneider. Data mining with neural networks for wheat yield prediction. In P. Perner, editor, *Advances in Data Mining – Medical Applications, E-Commerce, Marketing, and Theoretical Aspects*, volume 5077 of *LNAI*, pages 47–56, Berlin, Heidelberg, July 2008c. Springer Verlag. 32, 42
- G. Ruß, R. Kruse, and M. Schneider. A clustering approach for management zone delineation in precision agriculture. In R. Khosla, editor, *Proceedings of ICPA 2010*, Denver, July 2010a. International Society of Precision Agriculture. 71
- G. Ruß, R. Kruse, M. Schneider, and P. Wagner. Using advanced regression models for determining optimal soil heterogeneity indicators. In H. Locarek-Junge and C. Weihs, editors, *Classification as a Tool for Research, Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 463–471, Berlin, Heidelberg, New York, June 2010b. Springer. 32
- M. R. Sabatini and G. Palmerini. RTO-AG-160-V21 – Differential Global Positioning System (DGPS) for Flight Testing. *RTO-AG-160 AC/323(SCI-135)*, 21, 2008. 10
- M. Sandri and P. Zuccolotto. Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms. *Statistics and Computing*, 20:393–407, 2010. 65
- S. M. Scheiner and J. Gurevitch, editors. *Design and Analysis of Ecological Experiments*. Oxford University Press, New York, 2nd edition, 2001. 18, 164
- A. R. Schepers, J. F. Shanahan, M. A. Liebig, J. S. Schepers, S. H. Johnson, and A. Luchiari. Appropriateness of Management Zones for Characterizing Spatial Variability of Soil Properties and Irrigated Corn Yields across Years. *Agronomy Journal*, 96(1):195–203, 2004. 79, 80
- U. Schmidhalter, T. Selige, and J. Bobert. Geophysikalische und fernerkundliche Ermittlung teilflächenspezifischer Ertragspotenziale auf der Grundlage des Wasserhaushalts. In

- Managementsystem für den ortsspezifischen Pflanzenbau. Verbundprojekt pre agro, Abschlußbericht*, pages 239–291. KTBL, 2004. 11
- U. Schmidhalter, F.-X. Maidl, H. Heuwinkel, M. Demmel, H. Auernhammer, P. Noack, and M. Rothmund. *Perspectives for Agroecosystem Management*, chapter 2.3 – Precision Farming - Adaptation of land use management to small scale heterogeneity, pages 121–199. Elsevier, 2008. 6
- M. Schneider, G. Weigert, and P. Wagner. Teilflächenspezifische N-Düngung nach datenbankgestützten Entscheidungsregeln. In *Land- und Ernährungswirtschaft im Wandel: Aufgaben und Herausforderungen für die Agrar- und Umweltinformatik*. Gesellschaft für Informatik in der Land-, Forst- und Ernährungswirtschaft, 2006. 32
- S. Schroedl, K. L. Wagstaff, S. Rogers, P. Langley, and C. Wilson. Mining GPS traces for map refinement. *Data Mining and Knowledge Discovery*, 9:59–87, 2004. 107
- B. G. Sears, B. Mijatovic, T. G. Mueller, and R. I. Barnhisel. Interpreting yield variability with electrical conductivity and terrain attributes across a central kentucky landscape. *Crop Management*, 2005. 68, 69
- S. Shekhar, Y. Huang, W. Wu, and C. Lu. *Data Mining for Scientific and Engineering Applications*, chapter 1 – What’s Spatial about Spatial Data Mining: Three Case Studies. Kluwer Academic Publishers, 2001. 138
- G. Sinai, D. Zaslavsky, and P. Golany. The effect of soil surface curvature on moisture and yield - beer sheba observations. *Soil Science*, 132:367–375, 1981. 14
- A. J. Smola and B. Schölkopf. A tutorial on support vector regression. Technical report, Statistics and Computing, 1998. 43
- A. Soliman, R. Brown, and R. Heck. Separating near surface thermal inertia signals from a thermal time series by standardized principal component analysis. *International Journal of Applied Earth Observation and Geoinformation*, 13(4):607 – 615, 2011. 11
- M. Sommer and M. Wehrhan. Methods for an integrative, non-invasive site analysis to characterise site properties relevant for crop production in spatially heterogeneous agricultural areas. In *Forschungsverbundprojekt preagro II – Informationsgeleitete Pflanzenproduktion mit Precision Farming als zentrale inhaltliche und technische Voraussetzung für eine nachhaltige Entwicklung der landwirtschaftlichen Landnutzung. Zwischenbericht.*, pages 151–163, 2005. 10
- X. Song, J. Wang, W. Huang, L. Liu, G. Yan, and R. Pu. The delineation of agricultural management zones with high resolution remotely sensed data. *Precision Agriculture*, 10: 471–487, 2009. 88, 89
- R. Sørensen, U. Zinko, and J. Seibert. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. *Hydrology and Earth System Sciences Discussions*, pages 1807–1834, 2005. 21

- N. A. Spence. A multifactor uniform regionalization of British counties on the basis of employment data for 1961. *Regional Studies*, 2(1):87–104, 1968. 105
- M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer, June 1999. 16
- C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1): 25, 2007. ISSN 1471-2105. 47, 67
- C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008. 47, 51
- C. Strobl, J. Malley, and G. Tutz. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4):323 – 348, 2009. ISSN 1082-989X. 65
- P. R. Tozer and B. J. Isbister. Is it economically feasible to harvest by management zone? *Precision Agriculture*, 8:151–159, 2007. 74
- N. Tremblay, M. Y. Bouroubi, B. Panneton, S. Guillaume, P. Vigneault, and C. Belec. Development and validation of fuzzy logic inference to determine optimum rates of N for corn on the basis of field and crop features. *Precision Agriculture*, 11(6):621–635, 2010. 68
- M. Velandia, R. Rejesus, and E. Segarra. Economics of management zone delineation in cotton precision agriculture. In *IAAE conference proceedings*, Gold Coast, Australia, 2006. 74
- P. Wagner and M. Schneider. Economic benefits of neural network-generated site-specific decision rules for nitrogen fertilization. In J. V. Stafford, editor, *Proceedings of the 6th European Conference on Precision Agriculture*, pages 775–782, 2007. 15
- P. Wagner and G. Weigert. Development and evaluation of decision rules for site-specific nitrogen fertilization of wheat. In *Programme book of the joint conference of ECPA-ECPLF*, page 666, 2003. 32
- K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584, San Francisco, 2001. Morgan Kaufmann Publishers Inc. 105
- K. L. Wagstaff. *Intelligent Clustering with Instance-Level Constraints*. PhD thesis, Cornell University, 2002. 107
- K. L. Wagstaff, D. Mazzone, and S. Sain. HARVIST: A system for agricultural and weather studies using advanced statistical models. In *Proceedings of the Earth-Sun Systems Technology Conference*, 2005. 107

- W. Wang, J. Yang, and R. Muntz. Sting: A statistical information grid approach to spatial data mining. In *Proceedings of the 23rd VLBD conference, Athens, Greece, 1997*, pages 186–195. Morgan Kaufmann Publishers Inc., 1997. 95, 96
- R. Webster and P. A. Burrough. Computer-based soil mapping of small areas from sample data, II. Classification smoothing. *Soil Science*, 23(2):222–234, 1972a. 105, 106
- R. Webster and P. A. Burrough. Computer-based soil mapping of small areas from sample data, I. Multivariate classification and ordination. *Soil Science*, 23(2):210–221, 1972b. 105
- R. Webster and M. A. Oliver. *Geostatistics for Enviromental Scientists*. Wiley, UK, 2nd edition, 2007. 25
- G. Weigert. *Data Mining und Wissensentdeckung im Precision Farming - Entwicklung von ökonomisch optimierten Entscheidungsregeln zur kleinräumigen Stickstoff-Ausbringung*. PhD thesis, TU München, 2006. 11, 18, 31, 32, 65, 69
- G. Weigert and P. Wagner. Optimierung standortangepasster N-Düngung über teilflächen-spezifische Ertragsprognosen. In H.-J. Budde, R. A. E. Müller, and U. Birkner, editors, *Referate der 24. GIL-Jahrestagung*, pages 161–164, Göttingen, 2003. Gesellschaft für Informatik in der Land-, Forst- und Ernährungswirtschaft. ISBN 932987-05-5. 31
- G. Weigert, P. Wagner, and H. Linseisen. Development of decision rules for site-specific N-fertilization by the application of data mining techniques. In *Programme book of the joint conference of ECPA-ECPLF*, page 327, 2003. 32
- C. Weihs and G. Szepannek. Distances in classification. In *Proceedings of ICDM'2009*, volume 5633 of *Lecture Notes in Artificial Intelligence*, pages 1–12, Berlin, Heidelberg, 2009. Springer. 111
- C. Wood, G. Mullins, and B. Hajek. Phosphorous in agriculture. Technical report, Soil Quality Institute, USDA, Auburn, AL, USA, 1994. 13
- W. Xin-Zhong, L. Guo-Shun, H. Hong-Chao, W. Zhen-Hai, L. Qing-Hua, L. Xu-Feng, H. Wei-Hong, and L. Yan-Tao. Determination of management zones for a tobacco field based on soil fertility. *Computers and Electronics in Agriculture*, 65(2):168 – 175, 2009. ISSN 0168-1699. 87, 88, 147
- X. Xu, M. Ester, H.-P. Kriegel, and J. Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *Proc. of 14th Int. Conference on Data Engineering (ICDE'98)*, 1998. 94
- K.-S. Yang, R. Yang, and M. Kafatos. A feasible method to find areas with constraints using hierarchical depth-first clustering. In *Proceedings of the 13th International Conference on Scientific and Statistical Database Management*, pages 257–262, Los Alamitos, CA, USA, 2001. IEEE Computer Society. 105
- J. C. Zadoks, T. T. Chang, and C. F. Konzak. A decimal code for the growth stages of cereals. *Weed Research*, 14(6):415–421, 1974. 9

- K. R. Zalik and B. Zalik. A sweep-line algorithm for spatial clustering. *Advances in Engineering Software*, 40(6):445 – 451, 2009. 98
- K. Zeitouni. A survey of spatial data mining methods databases and statistics point of views. *Data warehousing and web engineering*, pages 229–242, 2002. 105
- L. W. Zevenbergen and C. R. Thorne. Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12:47–56, 1987. 20
- J. Zhu and G. D. Morgan. A nonparametric procedure for analyzing repeated measures of spatially correlated data. *Environmental and Ecological Statistics*, 11:431–443, 2004. ISSN 1352-8505. 65

Appendices

Appendix A

Data Set Details

This chapter serves as a reference for the data sets used in this thesis. The five data sets consist of a number of variables, whose descriptive statistics and graphical plots are presented here. The figures are typically referenced in the main part of the thesis. The variables resulting from the digital elevation models are not shown. In addition to the presented figures, a statistical summary of the respective data set's variables is given, also providing each variable's correlation with the respective yield (response) variable.

A.1 F440, Variable Plots and Descriptive Statistics

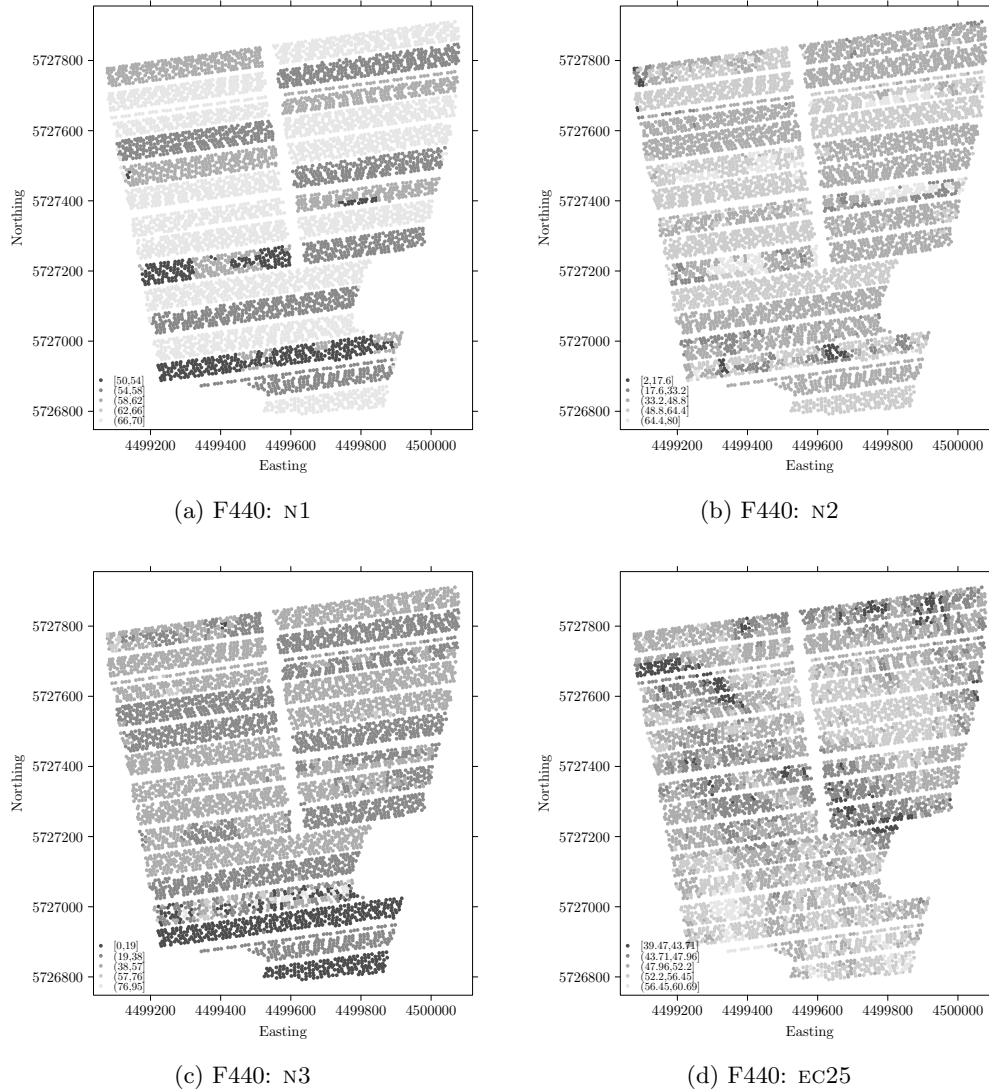
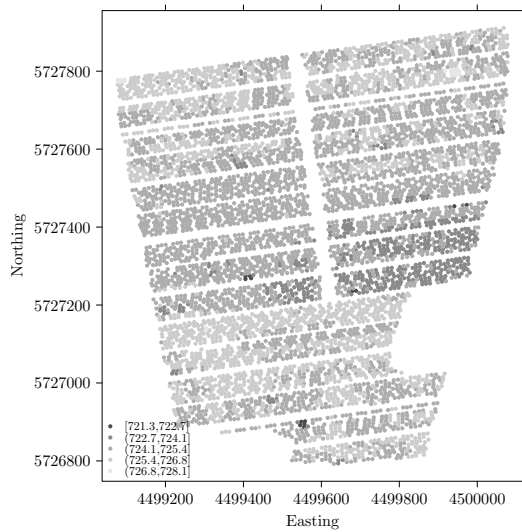
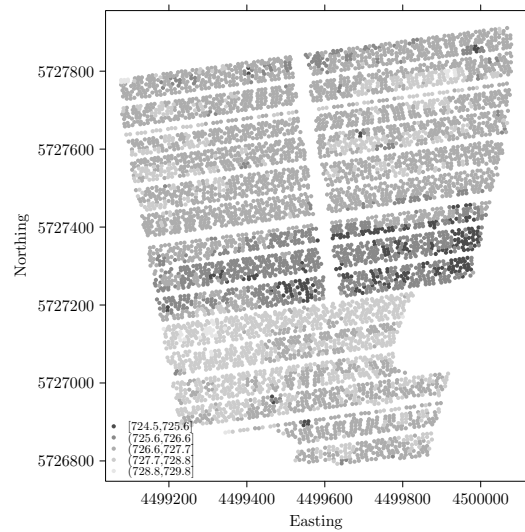


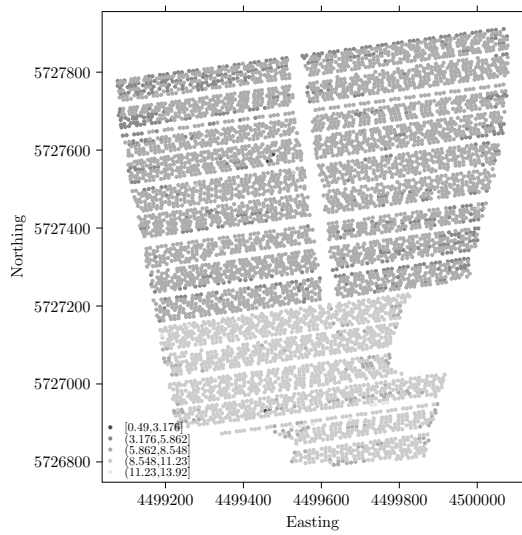
Figure A.1: F440: N1,N2,N3, EC25



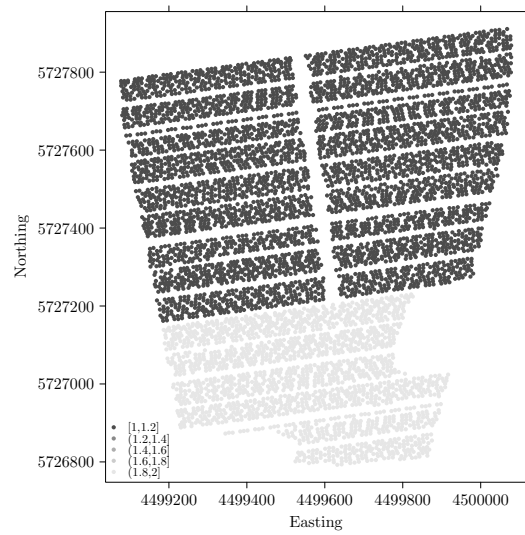
(a) F440: REIP32



(b) F440: REIP49



(c) F440: YIELD07



(d) F440: SORTE

Figure A.2: F440: REIP32, REIP49, YIELD07

	minimum	median	average	maximum	<i>COR</i> <i>Pearson</i>	<i>COR</i> <i>Spearman</i>	Levels
YIELD07	0.490000	6.890000	7.3706	13.9200	1.00	1.00	612
EC25	39.470000	50.220000	50.1273	60.6900	0.46	0.43	851
N3	0.000000	40.000000	37.9775	95.0000	-0.31	-0.20	50
N2	2.000000	48.000000	47.6015	80.0000	-0.07	0.02	45
N1	50.000000	70.000000	63.5681	70.0000	-0.08	-0.07	4
REIP32	721.330000	725.190000	725.1090	728.1400	0.40	0.38	367
REIP49	724.500000	727.340000	727.2038	729.8200	0.53	0.55	397
CATCHMENT.AREA	2.011141	77.572838	308.3326	50525.5983	0.13	0.14	6446
CATCHMENT.SLOPE	0.000232	0.005142	0.0051	0.0129	0.30	0.23	6446
CATCHMENT.AREA.MOD	42.316549	2876.880915	7351.6910	194625.2808	0.05	0.05	6446
WETNESS.INDEX	7.359758	11.747549	11.6374	16.0460	0.02	0.04	6446
SLOPE	0.002429	0.327796	0.3329	0.8316	0.21	0.18	6446
CURVATURE	-0.000538	0.000006	0.0000	0.0006	-0.06	-0.07	6446
CURVATURE.PLAN	-0.000383	-0.000002	-0.0000	0.0004	-0.06	-0.06	6446
CURVATURE.PROFILE	-0.000305	0.000009	0.0000	0.0004	-0.03	-0.04	6446

Table A.1: F440, descriptive statistics, correlation coefficients with YIELD

A.2 F550, Variable Plots and Descriptive Statistics

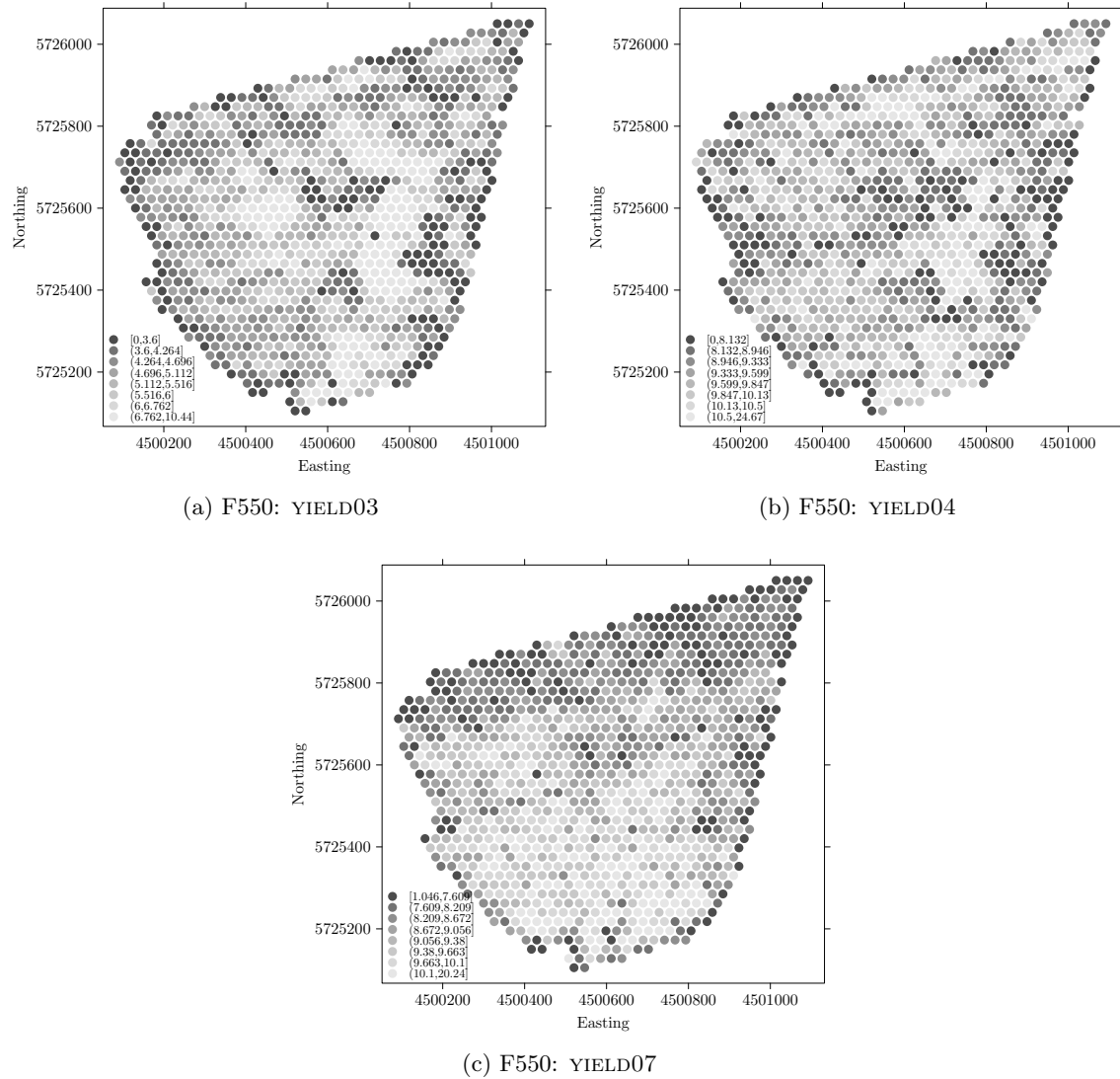
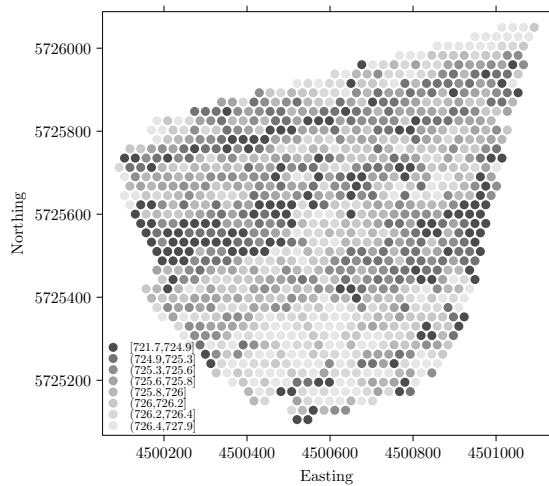
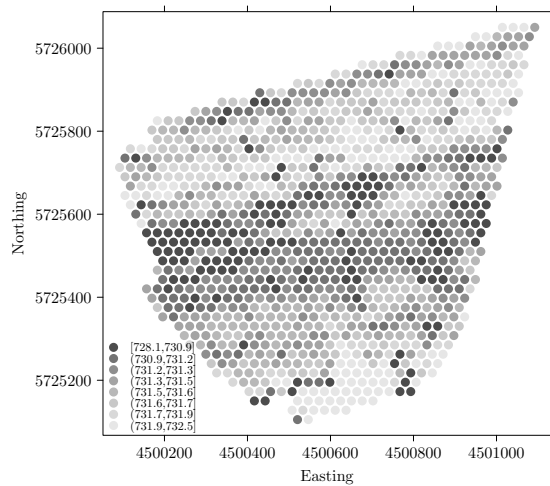


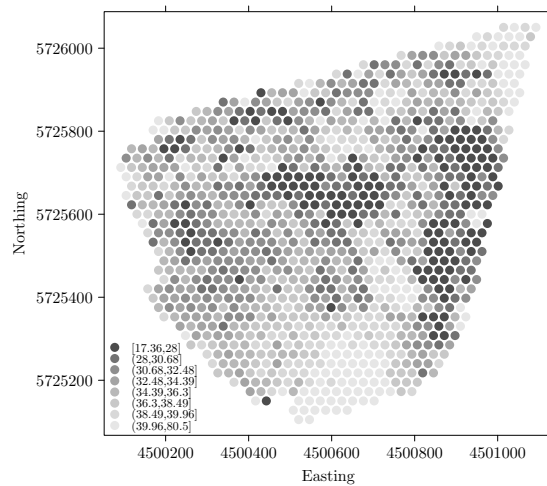
Figure A.3: F550: YIELD03, YIELD04, YIELD07



(a) F550: REIP32

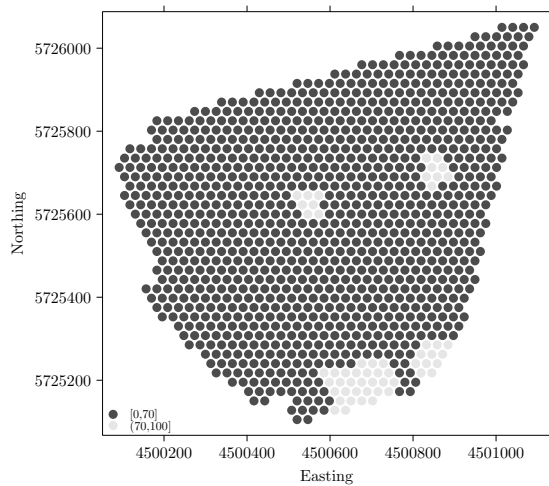


(b) F550: REIP49

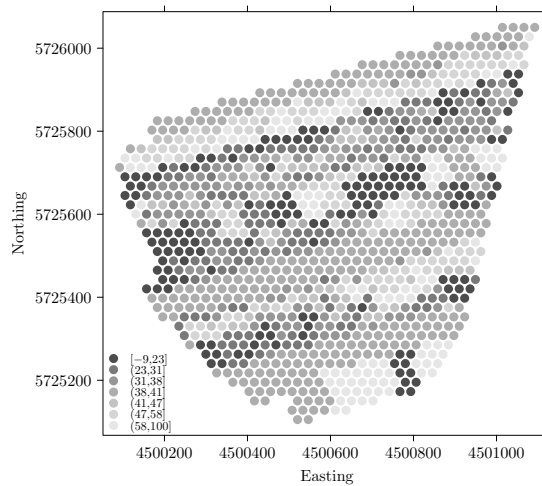


(c) F550: EC25

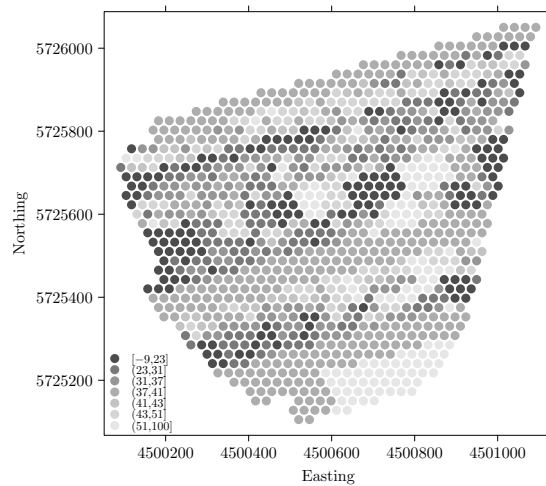
Figure A.4: F550: REIP32, REIP49, EC25



(a) F550: N1

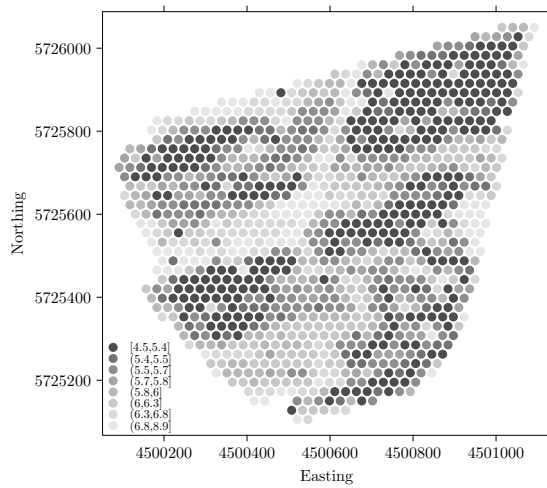


(b) F550: N2

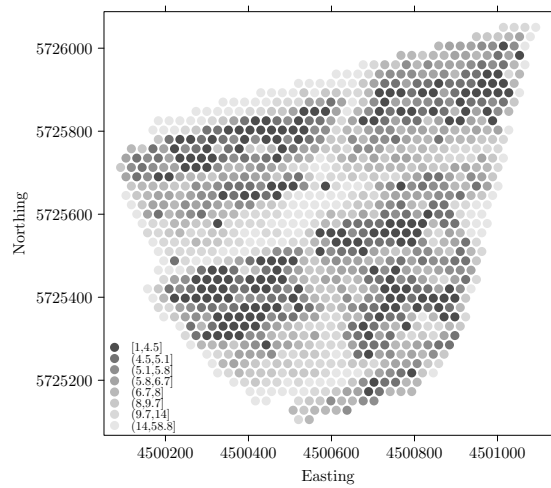


(c) F550: N3

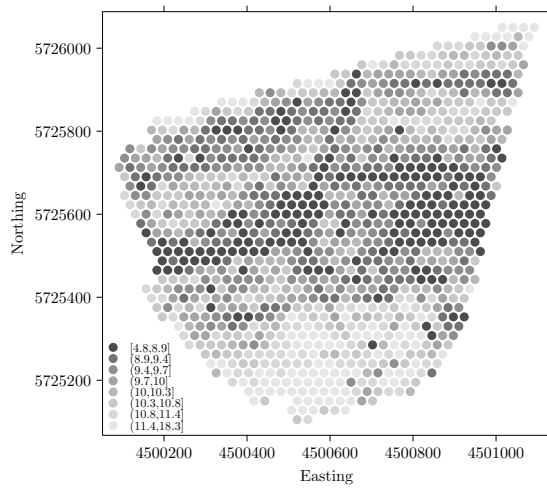
Figure A.5: F550: N1,N2,N3



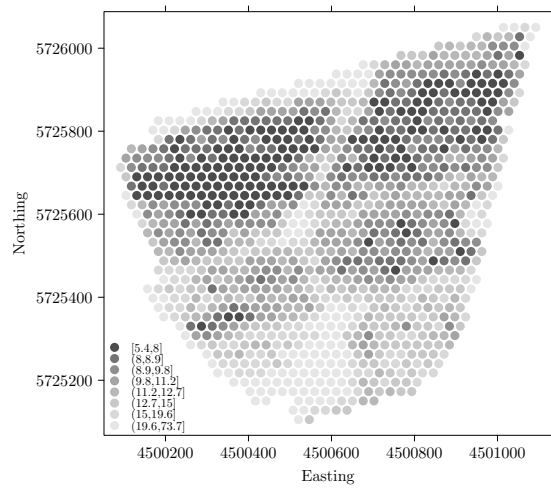
(a) F550: pH value



(b) F550: Phosphorus



(c) F550: Magnesium



(d) F550: Potassium

Figure A.6: F550: pH, P, Mg, K

	minimum	median	average	maximum	$COR_{Pearson}$	$COR_{Spearman}$	Levels
YIELD04	0.000000	9.639162	9.2909	24.6713	1.00	1.00	978
YIELD03	0.000000	5.111930	5.0985	10.4423	0.34	0.44	971
EC25	19.290000	34.530000	34.7323	80.5000	0.07	0.20	605
N3	2.000000	40.000000	39.9513	100.0000	0.12	0.15	66
N2	2.000000	40.000000	41.7266	100.0000	0.13	0.18	72
N1	30.000000	60.000000	60.3877	100.0000	0.07	0.06	7
REIP32	722.644135	725.833010	725.7551	727.9293	0.12	0.05	1000
REIP49	728.588624	731.490040	731.4579	732.5247	0.16	0.22	1001
CATCHMENT.AREA	1.535021	50.407056	133.8512	11068.3987	0.06	0.30	1006
CATCHMENT.SLOPE	0.000172	0.009093	0.0100	0.0325	-0.01	0.01	1006
CATCHMENT.AREA.MOD	6.483051	1255.078227	4553.7905	152021.8243	0.13	0.31	1006
WETNESS.INDEX	5.230815	10.760694	10.5721	15.9159	0.20	0.31	1006
SLOPE	0.011845	0.519174	0.5991	2.3376	-0.08	-0.12	1006
CURVATURE	-0.001694	-0.000011	-0.0000	0.0018	-0.19	-0.32	1006
CURVATURE.PLAN	-0.001035	-0.000009	-0.0000	0.0011	-0.18	-0.25	1006
CURVATURE.PROFILE	-0.001141	-0.000010	-0.0000	0.0012	-0.13	-0.27	1006

Table A.2: F550, descriptive statistics, correlation coefficients with YIELD

A.3 F610, Variable Plots and Descriptive Statistics

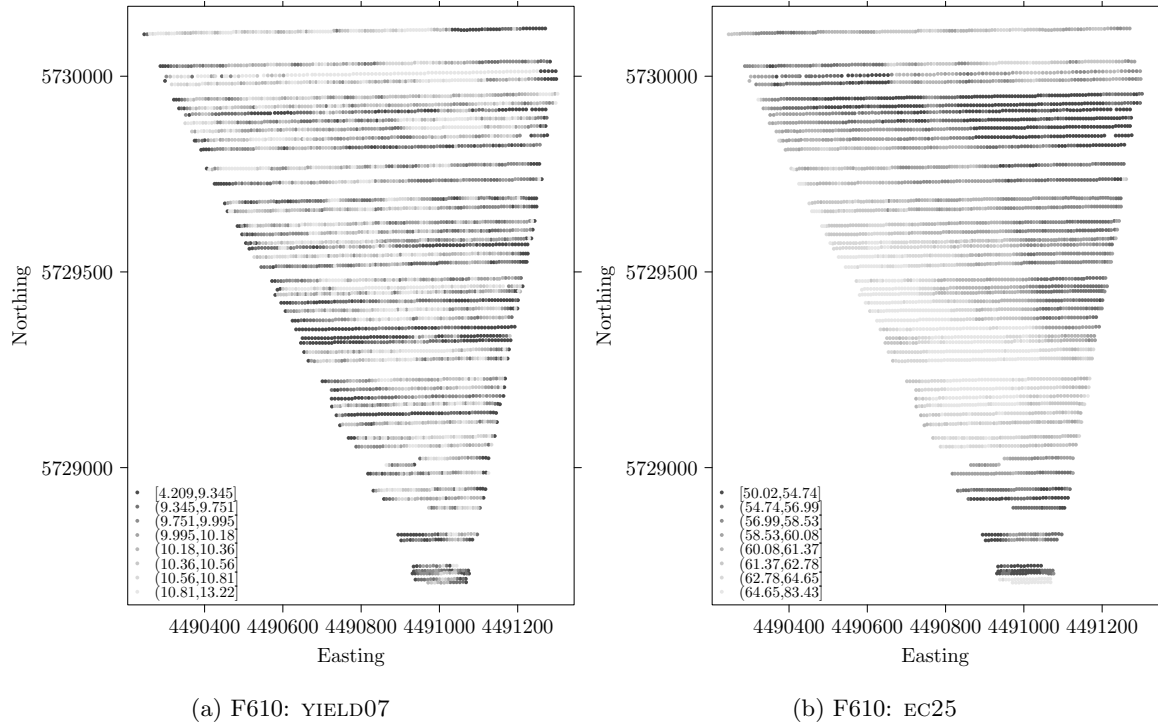
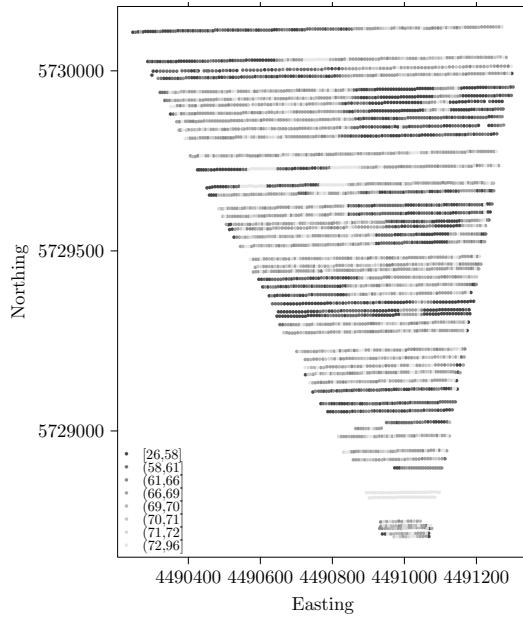
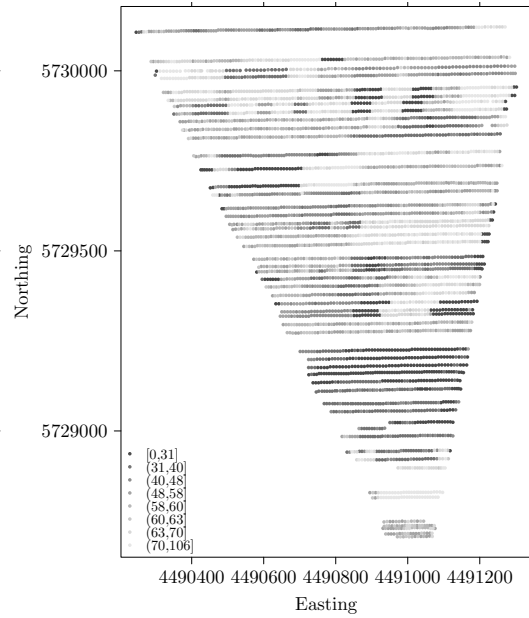


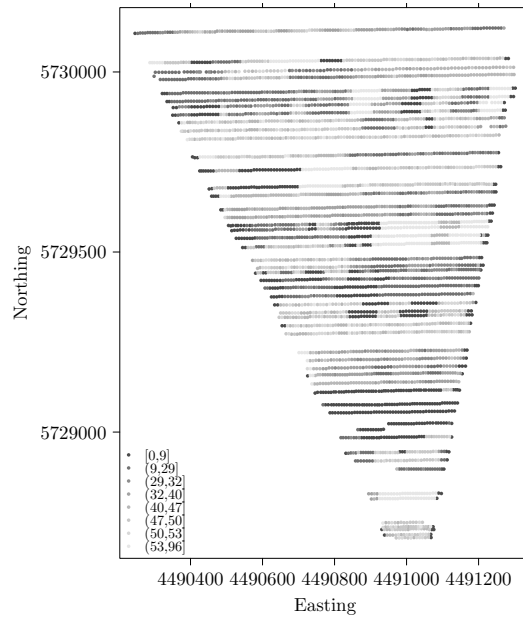
Figure A.7: F610: YIELD07, EC25



(a) F610: N1



(b) F610: N2



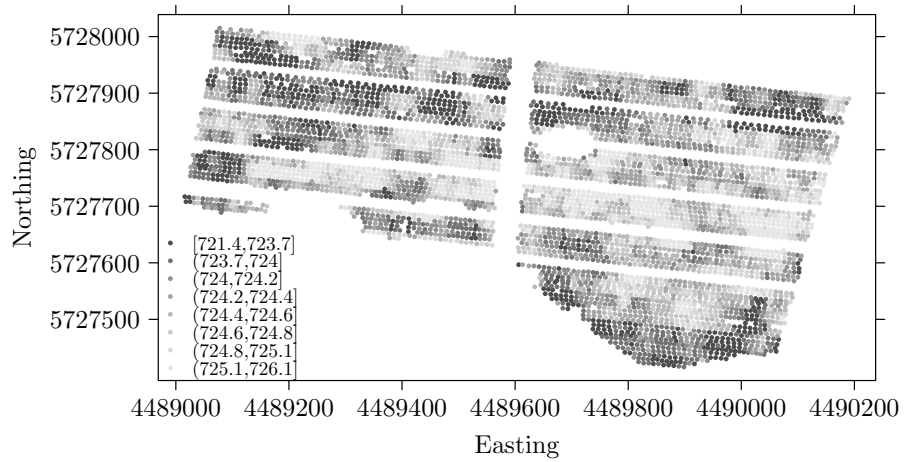
(c) F610: N3

Figure A.8: F610: N1,N2,N3

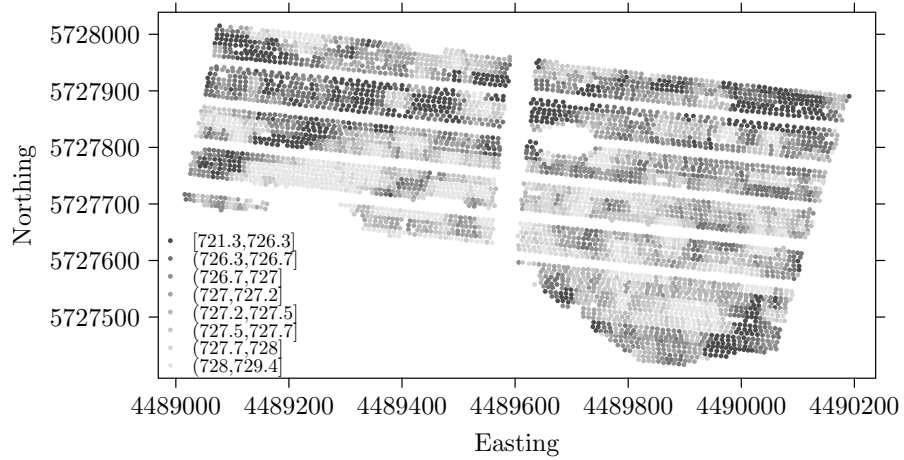
	minimum	median	average	maximum	$COR_{Pearson}$	$COR_{Spearman}$	Levels
YIELD07	4.208895	10.184023	10.0816	13.2213	1.00	1.00	2951
EC25	50.020000	60.060000	59.8532	83.4300	-0.01	-0.00	846
N3	0.000000	39.000000	35.1944	96.0000	0.04	-0.02	84
N2	0.000000	57.000000	50.7260	106.0000	0.09	0.00	95
N1	26.000000	68.000000	64.6757	96.0000	0.10	0.08	65
CATCHMENT.AREA	1.127921	66.335663	224.3115	10942.1789	-0.01	0.01	3996
CATCHMENT.SLOPE	0.000174	0.004349	0.0046	0.0115	0.06	0.04	3996
CATCHMENT.AREA.MOD	21.494001	6553.091303	10420.9736	76218.6085	-0.05	-0.06	3996
WETNESS.INDEX	6.784818	12.608011	12.3342	15.1807	-0.05	-0.06	3996
SLOPE	0.008623	0.250229	0.2679	0.7975	0.09	0.05	3996
CURVATURE	-0.000616	-0.000005	-0.0000	0.0006	-0.02	-0.01	3996
CURVATURE.PLAN	-0.000379	0.000002	-0.0000	0.0003	-0.01	-0.02	3996
CURVATURE.PROFILE	-0.000337	-0.000008	-0.0000	0.0004	-0.03	-0.01	3996

Table A.3: F610, descriptive statistics, correlation coefficients with YIELD

A.4 F611, Variable Plots and Descriptive Statistics

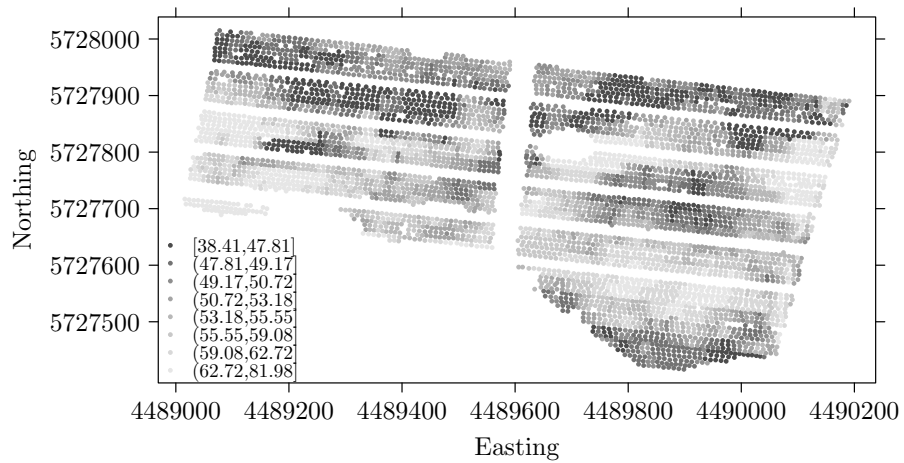


(a) F611: REIP32

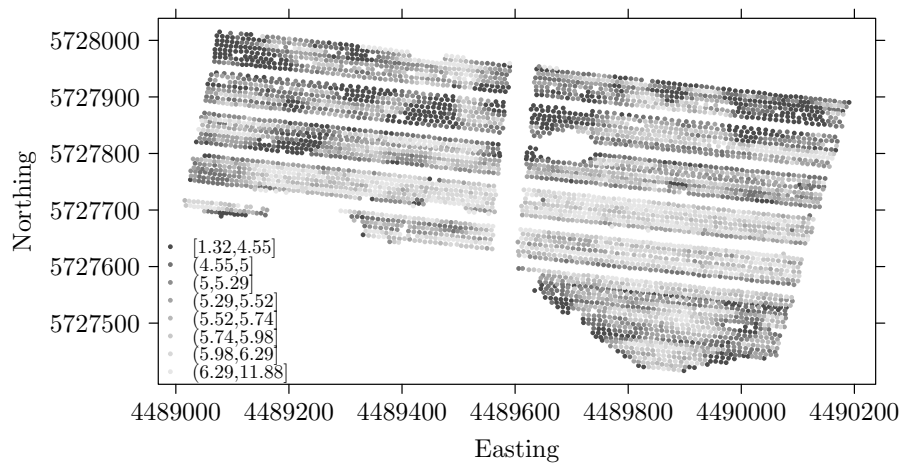


(b) F611: REIP49

Figure A.9: F611: REIP32, REIP49

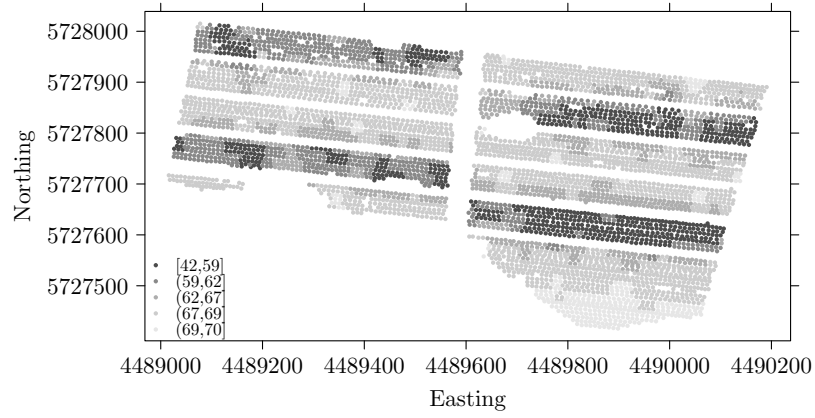


(a) F611: EC25

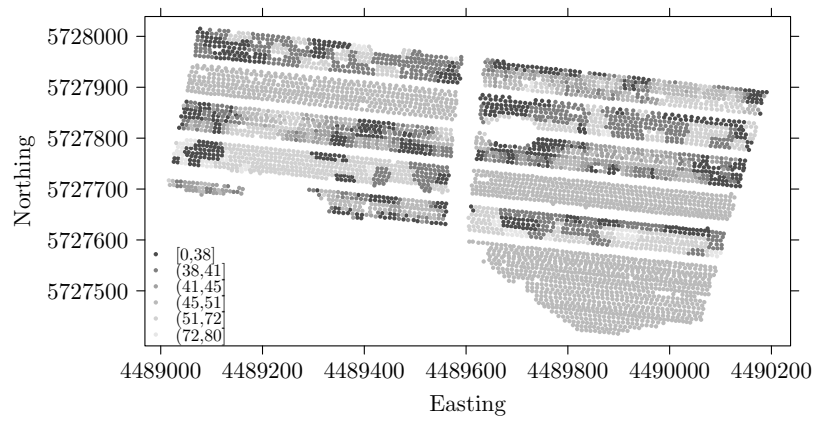


(b) F611: YIELD07

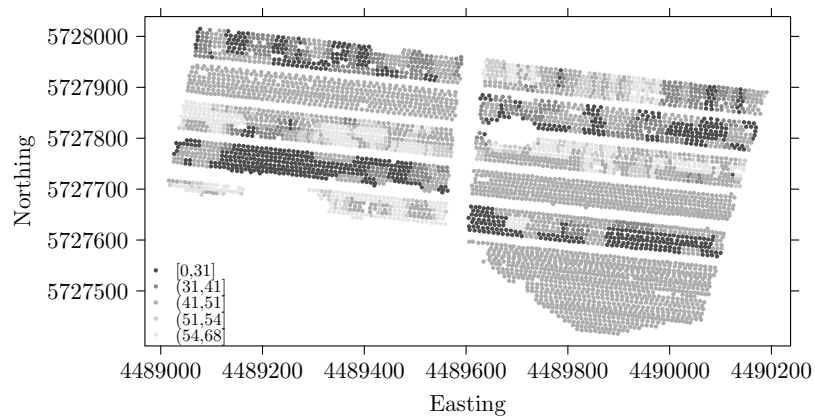
Figure A.10: F611: EC25, YIELD07



(a) F611: N1



(b) F611: N2



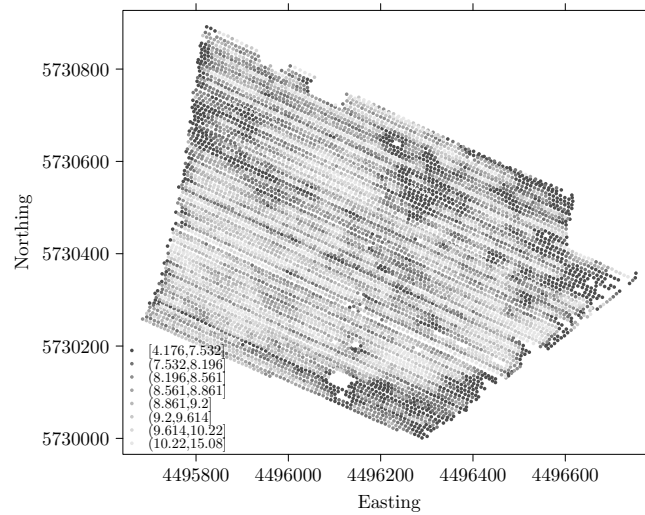
(c) F611: N3

Figure A.11: F611: N1,N2,N3

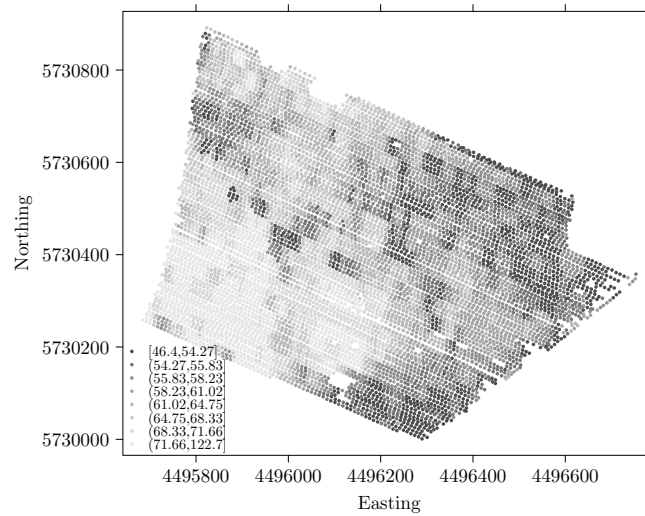
	minimum	median	average	maximum	<i>COR</i> _{Pearson}	<i>COR</i> _{Spearman}	Levels
YIELD07	1.320000	5.510000	5.4224	11.8800	1.00	1.00	465
EC25	38.410000	53.170000	54.4351	81.9800	0.31	0.29	1199
N3	0.000000	50.000000	45.6097	68.0000	0.03	0.05	49
N2	0.000000	50.000000	47.8903	80.0000	0.15	0.06	44
N1	42.000000	68.000000	65.0891	70.0000	-0.12	-0.17	26
REIP32	721.410000	724.415000	724.3705	726.0900	0.52	0.48	308
REIP49	721.300000	727.230000	727.1204	729.4100	0.61	0.52	437
CATCHMENT.AREA	2.567747	143.405428	210.3313	42253.1319	0.08	0.38	4970
CATCHMENT.SLOPE	0.000608	0.039677	0.0385	0.0813	-0.14	-0.16	4970
CATCHMENT.AREA.MOD	6.729962	187.089904	4316.9496	118921.3251	0.03	0.30	4970
WETNESS.INDEX	4.801615	8.243747	8.8857	15.7145	0.19	0.32	4970
SLOPE	0.020075	1.728570	1.7593	5.4382	-0.38	-0.35	4970
CURVATURE	-0.001343	-0.000113	-0.0000	0.0037	-0.30	-0.22	4970
CURVATURE.PLAN	-0.001246	0.000022	0.0000	0.0015	-0.30	-0.23	4970
CURVATURE.PROFILE	-0.001226	-0.000106	-0.0001	0.0029	-0.21	-0.16	4970

Table A.4: F611, descriptive statistics, correlation coefficients with YIELD

A.5 F631, Variable Plots and Descriptive Statistics

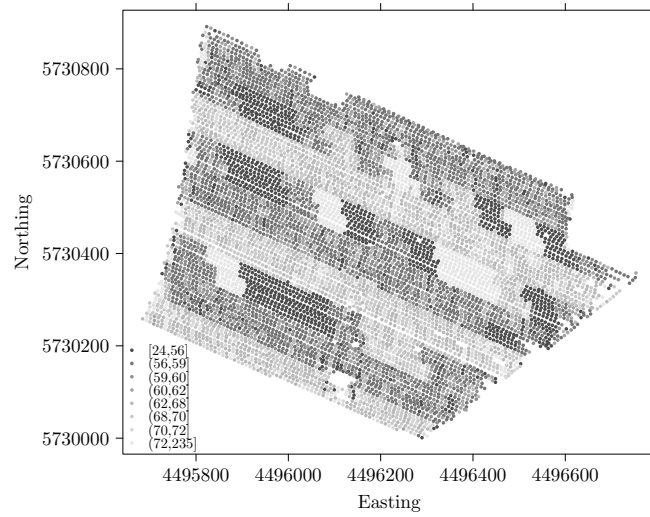


(a) F631: YIELD07

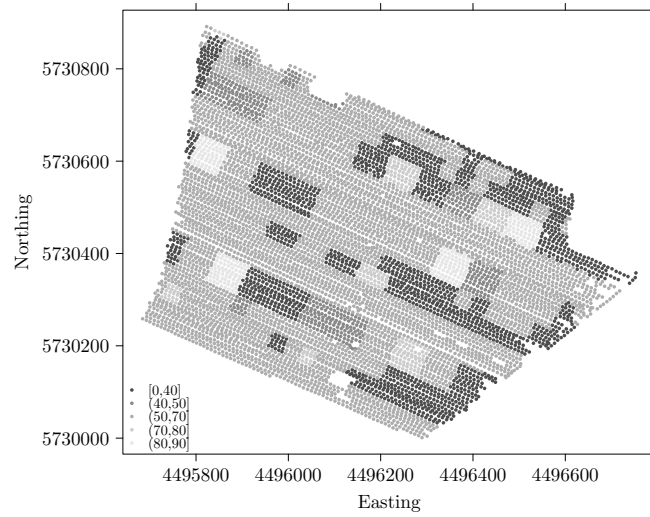


(b) F631: EC25

Figure A.12: F631: YIELD07, EC25



(a) F631: N1



(b) F631: N2

Figure A.13: F631: N1,N2

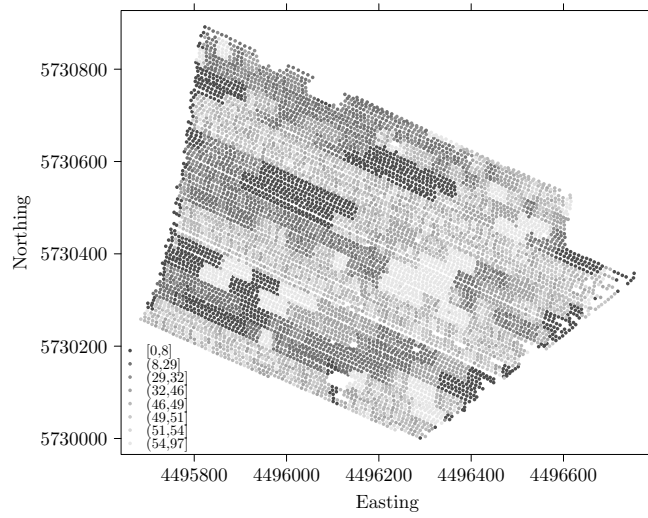


Figure A.14: F631: N3

	minimum	median	average	maximum	<i>COR</i> _{Pearson}	<i>COR</i> _{Spearman}	Levels
YIELD07	4.176233	8.859837	8.8639	15.0772	1.00	1.00	5199
EC25	46.400000	61.000000	63.0716	122.7300	0.29	0.36	1290
N3	0.000000	45.000000	37.1992	97.0000	0.00	-0.04	83
N2	0.000000	60.000000	54.7916	90.0000	0.21	0.22	9
N1	24.000000	61.000000	62.1341	235.0000	0.01	0.03	74
CATCHMENT.AREA	2.161526	162.181608	392.5565	5974.9762	0.17	0.32	7847
CATCHMENT.SLOPE	0.000993	0.015964	0.0157	0.0253	0.07	0.07	7847
CATCHMENT.AREA.MOD	8.578191	272.148114	2344.6815	300603.1376	0.02	0.25	7847
WETNESS.INDEX	5.883951	9.078069	9.3097	16.3687	0.21	0.25	7847
SLOPE	0.064432	0.794242	0.8288	1.9167	-0.16	-0.15	7847
CURVATURE	-0.000817	-0.000027	0.0000	0.0011	-0.35	-0.31	7847
CURVATURE.PLAN	-0.000722	-0.000009	0.0000	0.0007	-0.30	-0.28	7847
CURVATURE.PROFILE	-0.000471	-0.000021	-0.0000	0.0006	-0.28	-0.27	7847

Table A.5: F631, descriptive statistics, correlation coefficients with YIELD

Appendix B

Spatial Variable Importance Plots

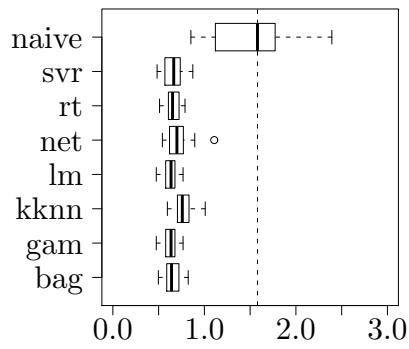
This appendix provides the detailed results of the spatial variable importance (SVI) approach summarized in Chapter 3. Furthermore, the regression models can be compared according to their predictive performance. The models' RMSE is measured in t ha^{-1} . The SVI is also based on the yield prediction errors and has the same unit. The units are not mentioned further.

For the RMSE comparisons for each site and sub-site, the *naive* model serves as a reference model providing a baseline which the regression models compete against. Each RMSE comparison figure (such as B.1) first gives the results of using the complete data set (Figure B.1a) before showing the results of using subsets of the data set according to specific fertilization strategies (e.g. Figures B.1b to B.1e).

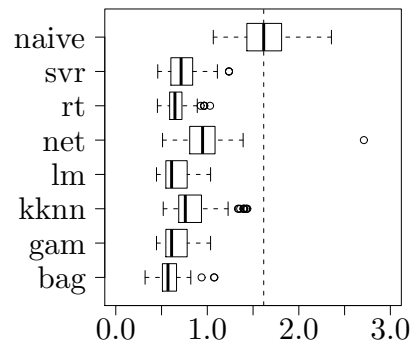
Each of the RMSE figures (e.g. B.1a to B.1e) is associated with a collection of SVI plots (e.g. B.2 to B.6) where the SVI is displayed according to one of the respective regression models. Those figures depict the mean RMSE increase of each model per variable.

data set	RMSE figures	SVI figures
F440	B.1	B.2 – B.6
F440sorte1	B.7	B.8 – B.12
F440sorte2	B.13	B.14 – B.18
F550 w/o yield	B.19	B.20 – B.24
F550 w yield	B.25	B.26 – B.30
F610	B.31	B.32 – B.37
F611	B.38	B.39 – B.42
F631	B.43	B.44 – B.47

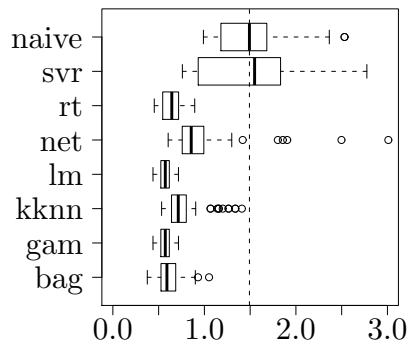
Table B.1: Overview of RMSE/SVI figures per data set



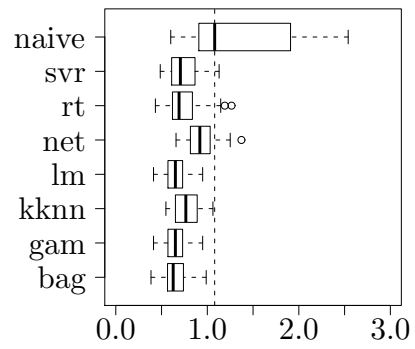
(a) f440, all strategies, rmse of models



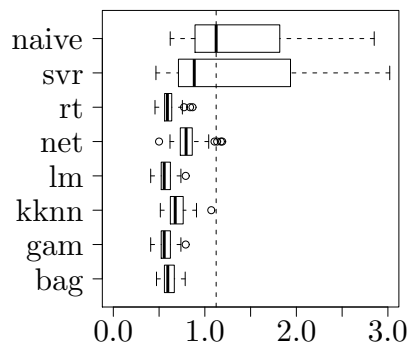
(b) f440, "low constant fertilization"



(c) f440, strategy "constant fertilization"



(d) f440, strategy "neural network"



(e) f440, strategy "sensor"

Figure B.1: RMSE for F440 and its subsets (by strategy)

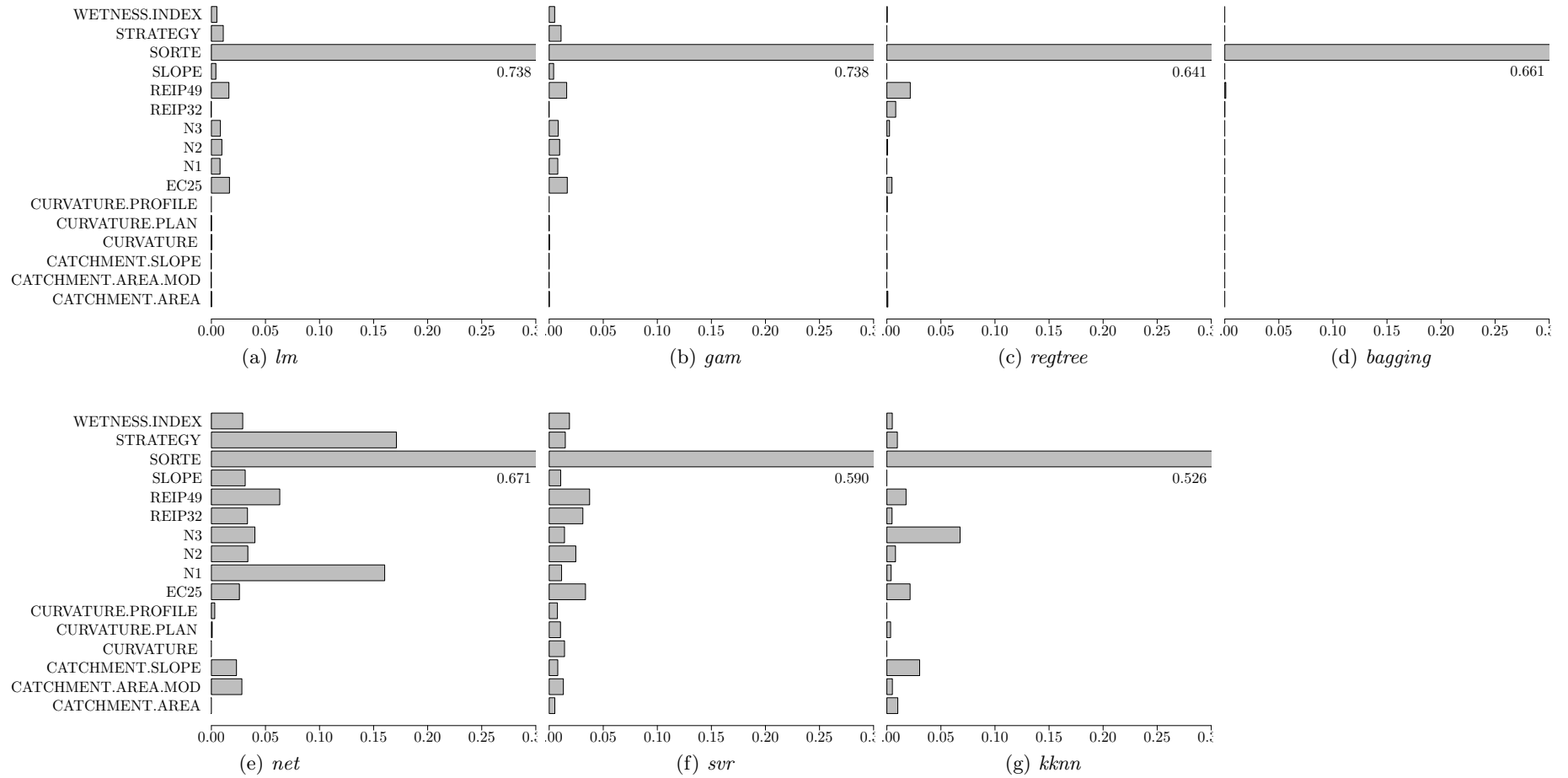


Figure B.2: F440, all strategies, regression models and spatial variable importance

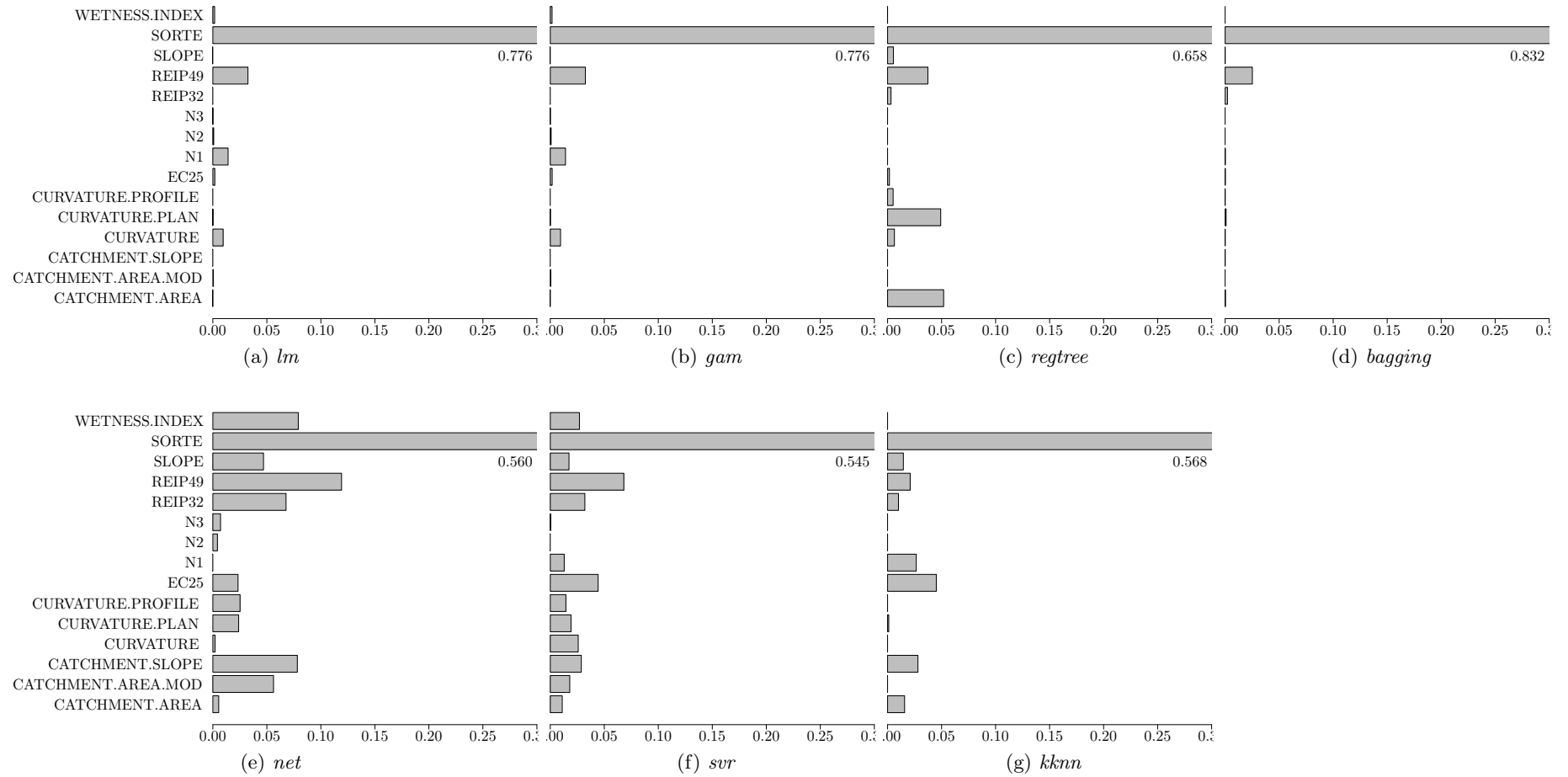


Figure B.3: F440, strategy “low, constant”, regression models and spatial variable importance

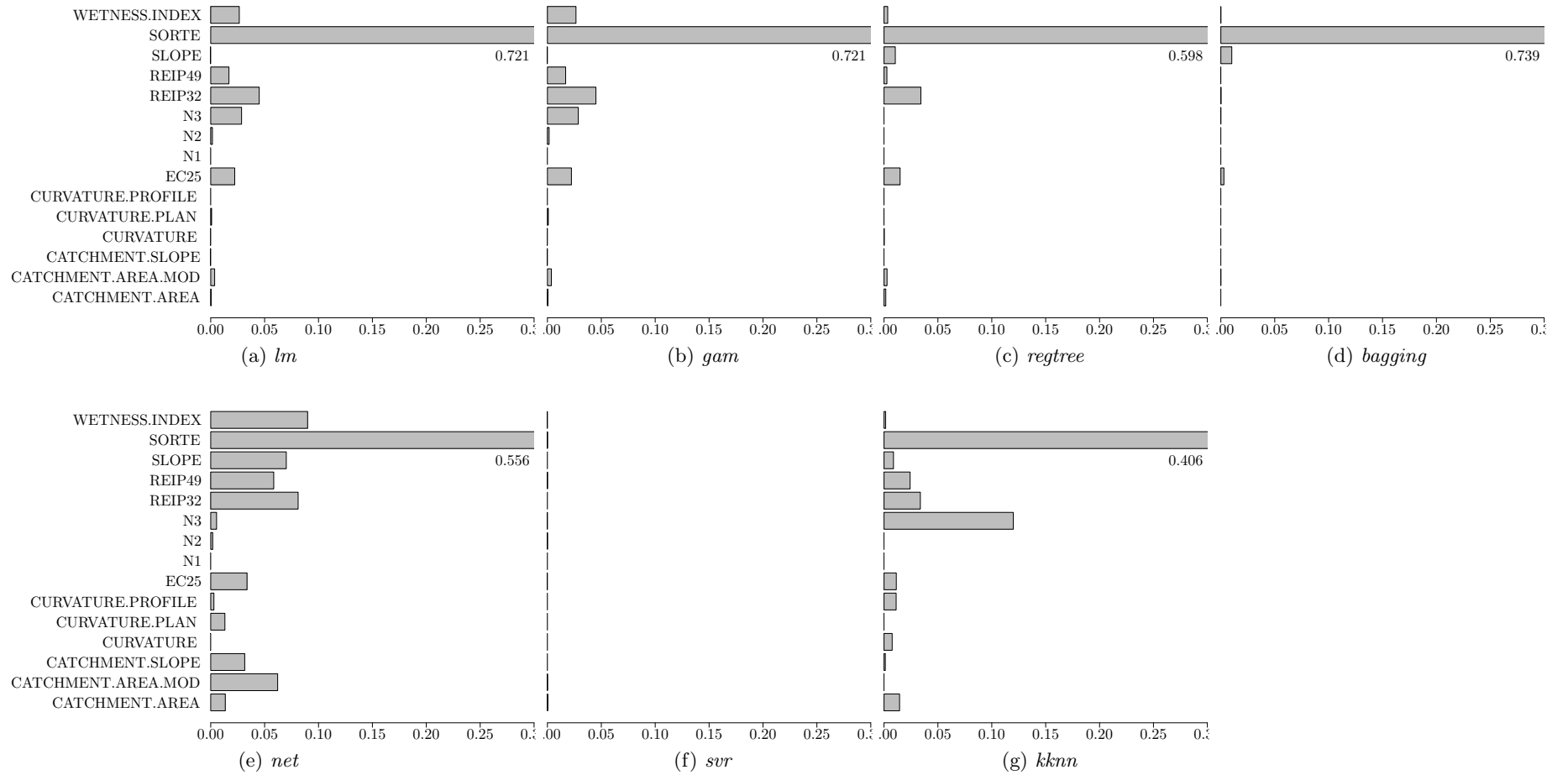


Figure B.4: F440, strategy “constant”, regression models and spatial variable importance

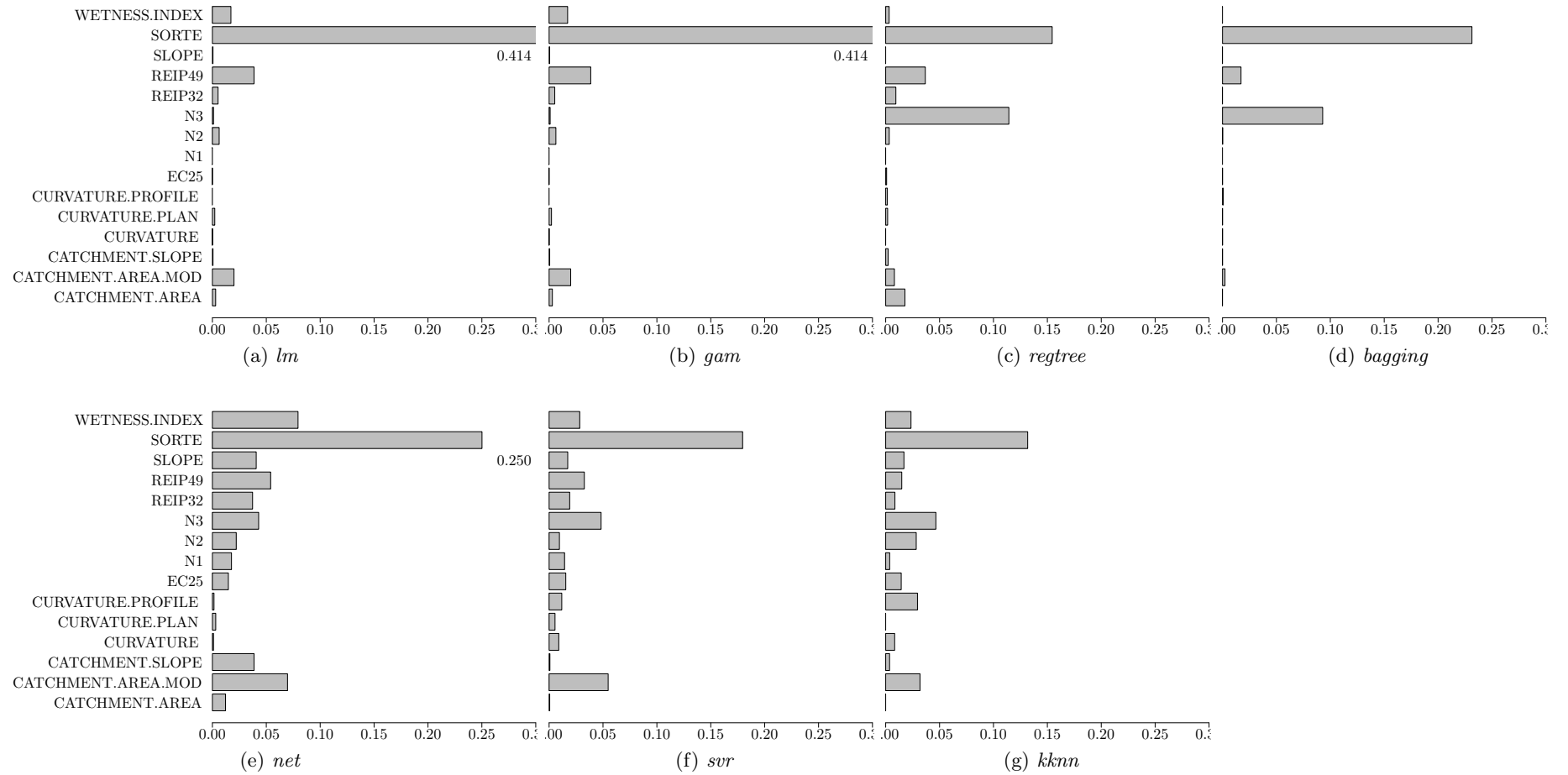


Figure B.5: F440, strategy “neural network”, regression models and spatial variable importance

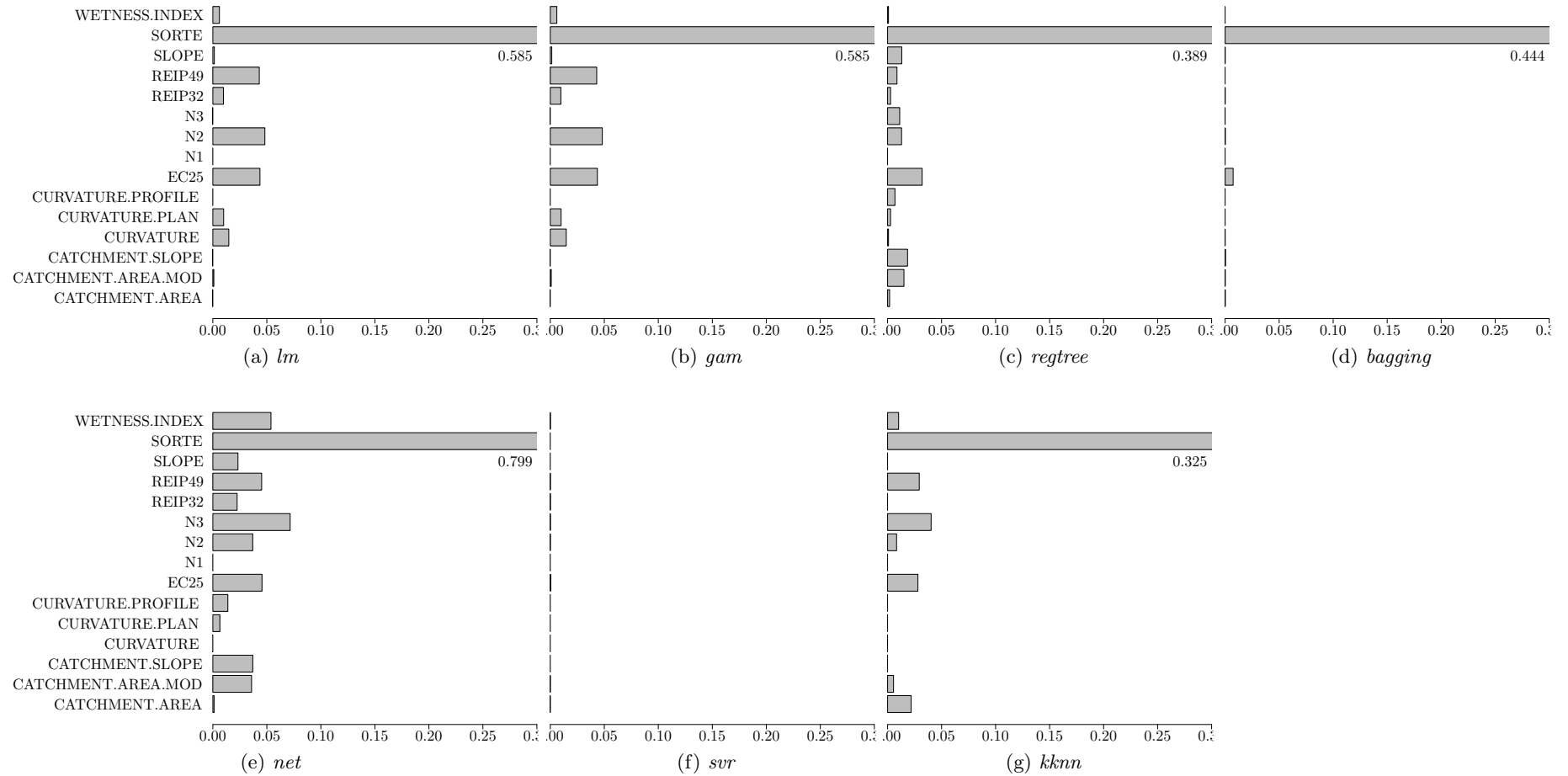
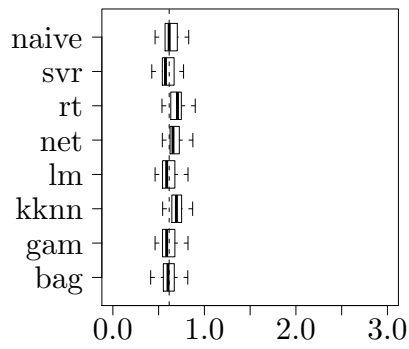
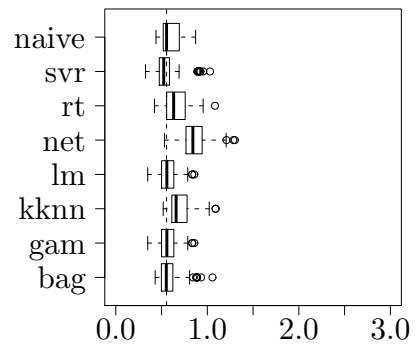


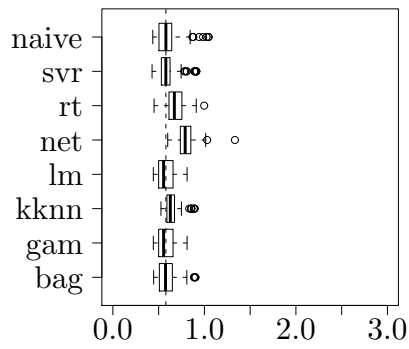
Figure B.6: F440, strategy “sensor”, regression models and spatial variable importance



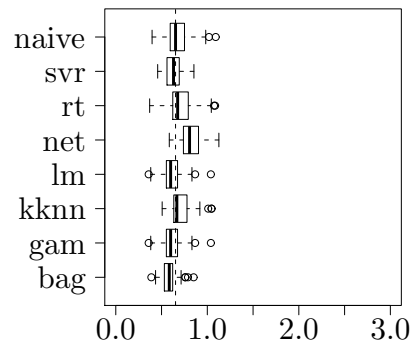
(a) f440sorte1, all strategies, rmse of models



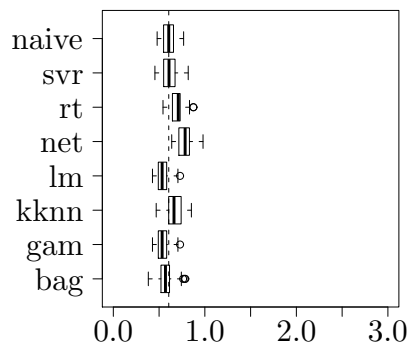
(b) f440sorte1, "low constant fertilization"



(c) f440sorte1, strategy "constant fertilization"



(d) f440sorte1, strategy "neural network"



(e) f440sorte1, strategy "sensor"

Figure B.7: RMSE for F440sorte1 and its subsets (by strategy)

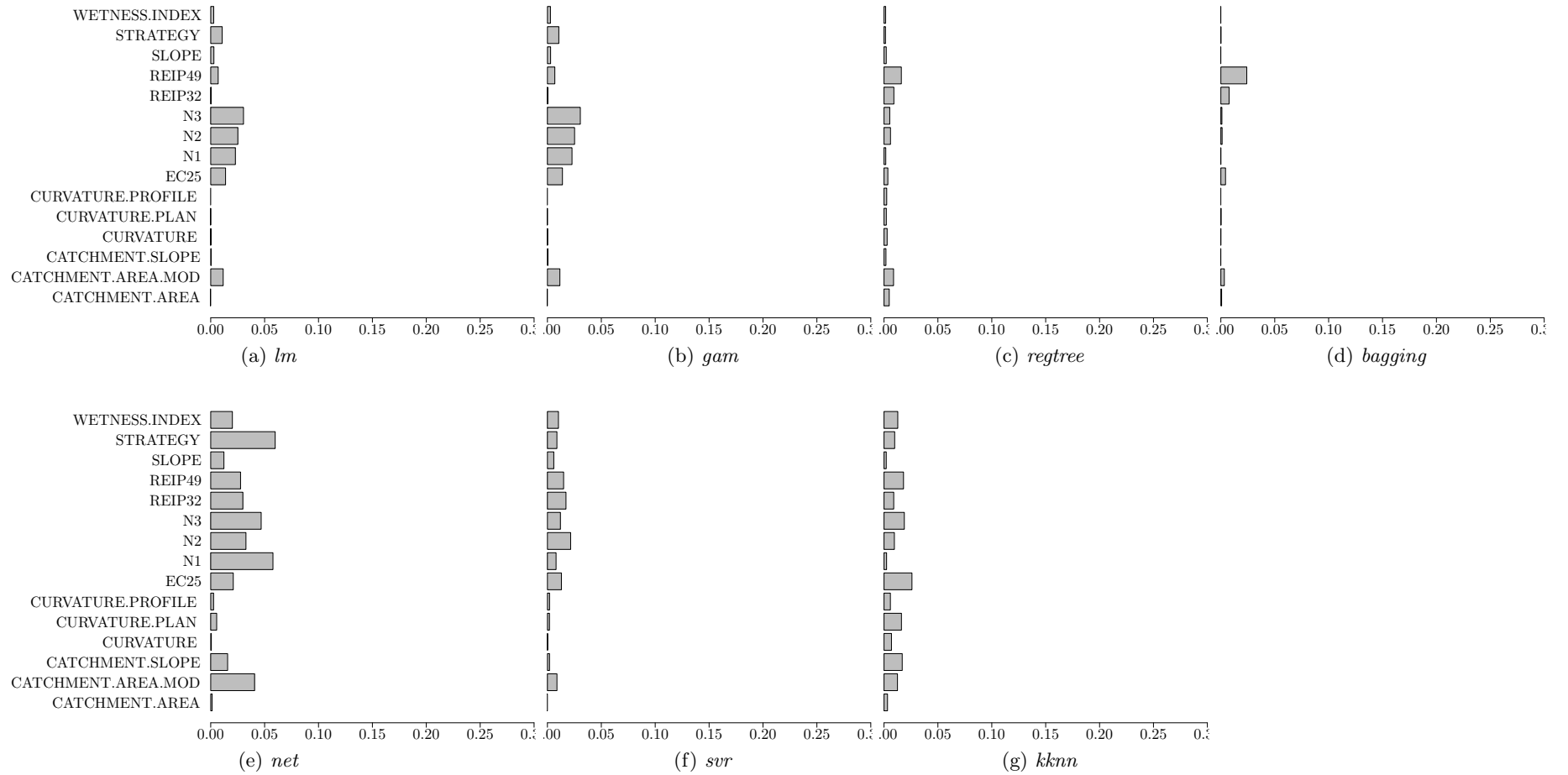


Figure B.8: F440sorte1, all strategies, regression models and spatial variable importance

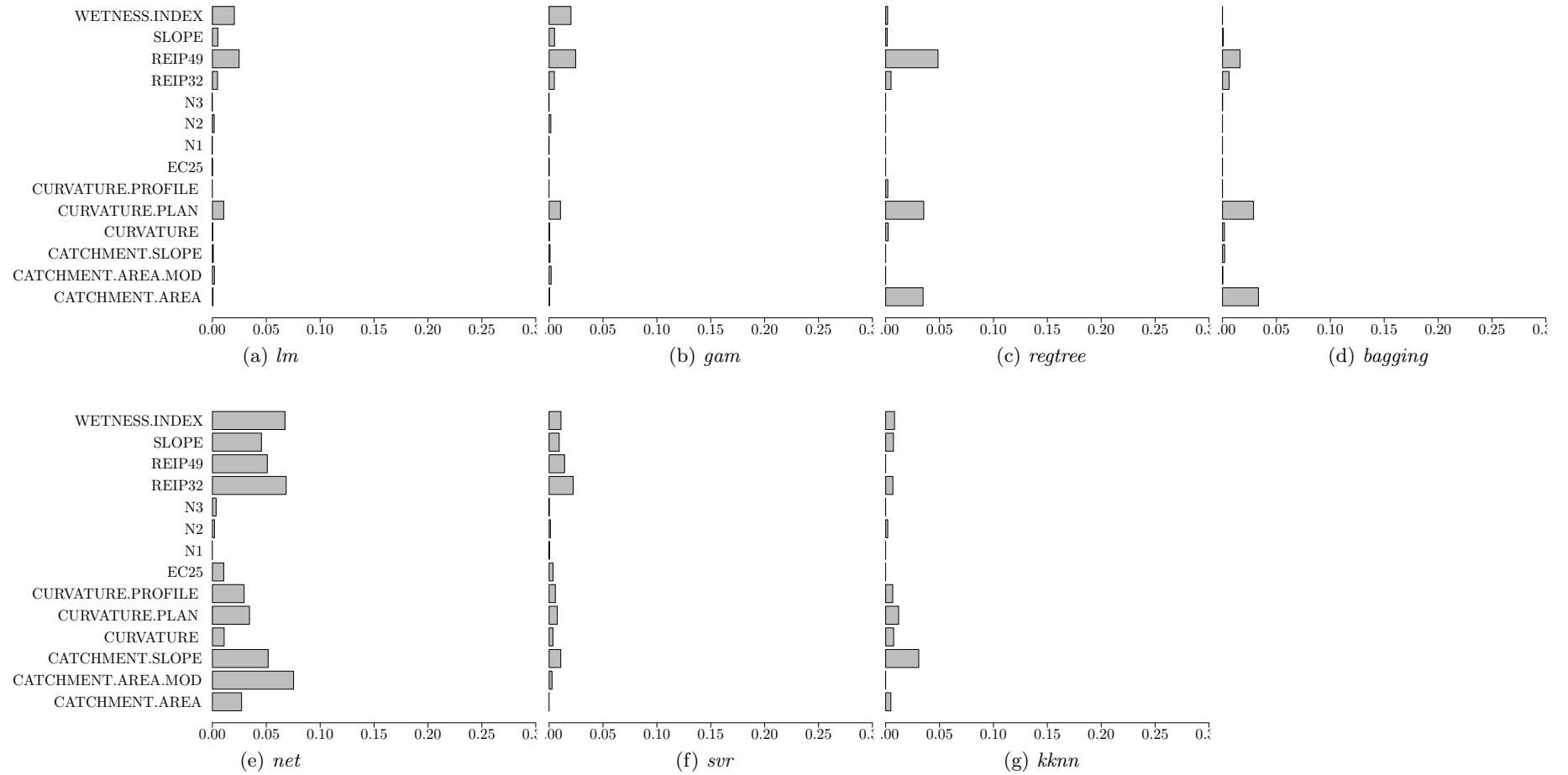


Figure B.9: F440sorte1, strategy “low, constant”, regression models and spatial variable importance

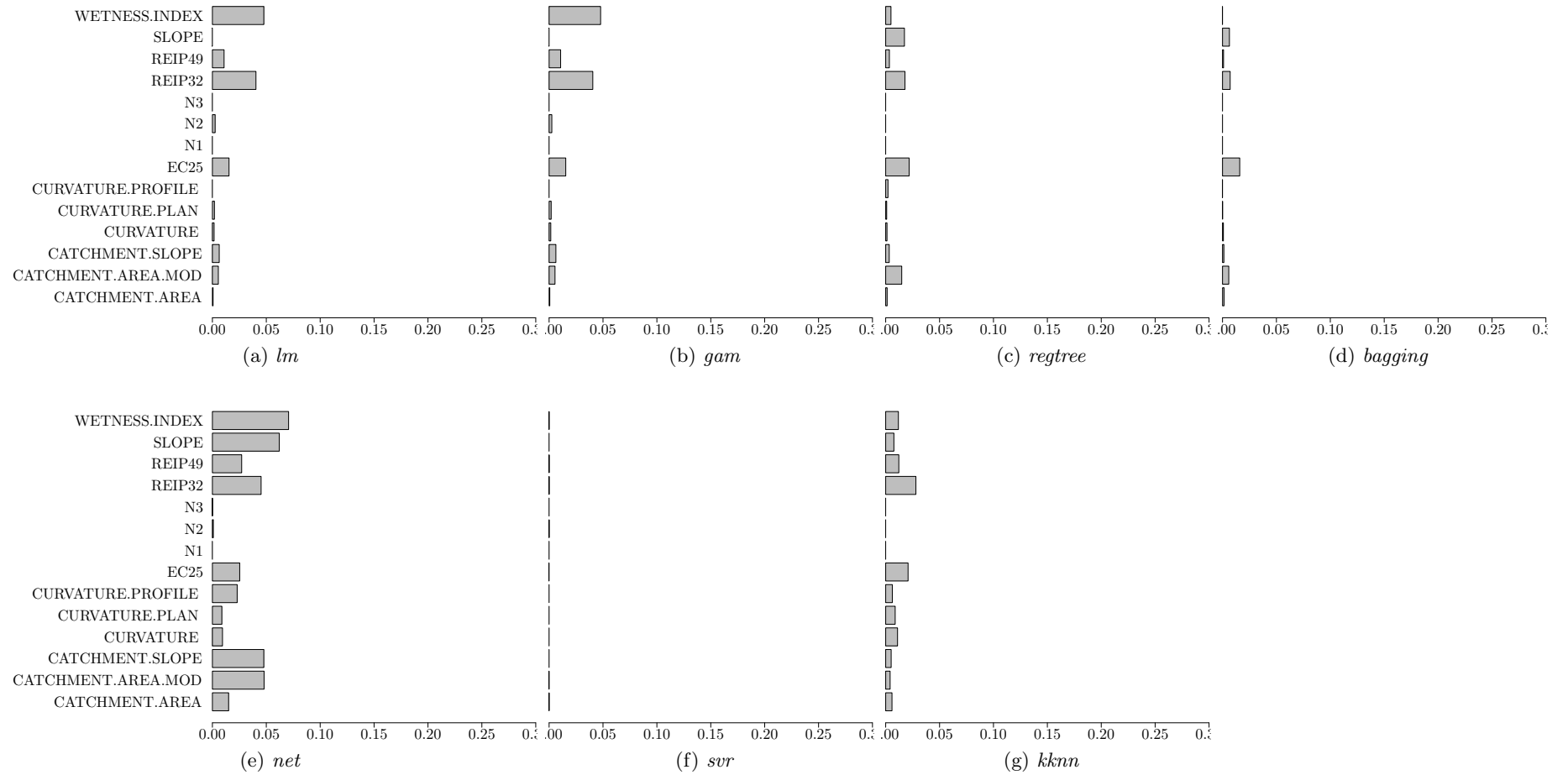


Figure B.10: F440sorte1, strategy “constant”, regression models and spatial variable importance

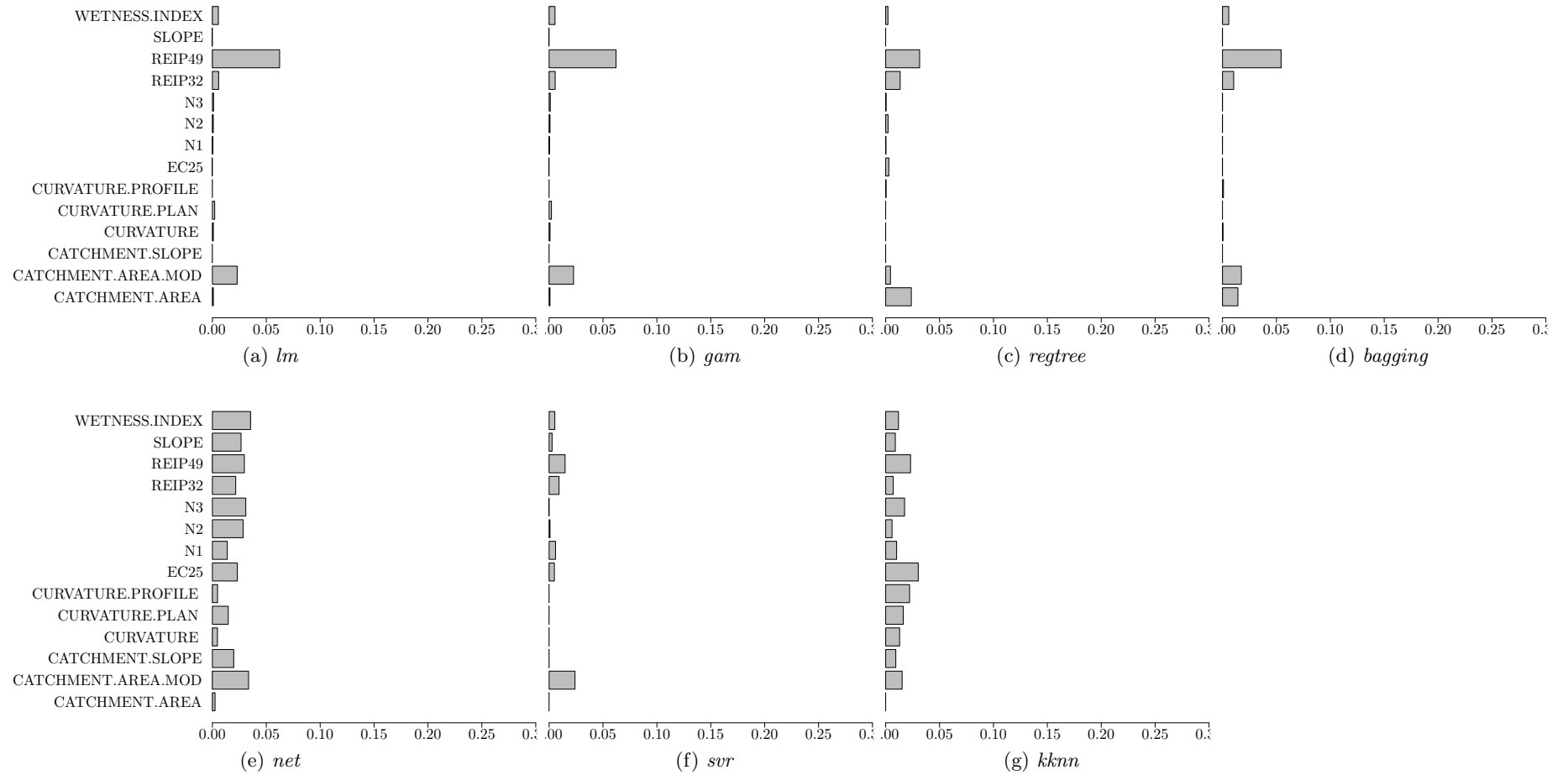


Figure B.11: F440sorte1, strategy “neural network”, regression models and spatial variable importance

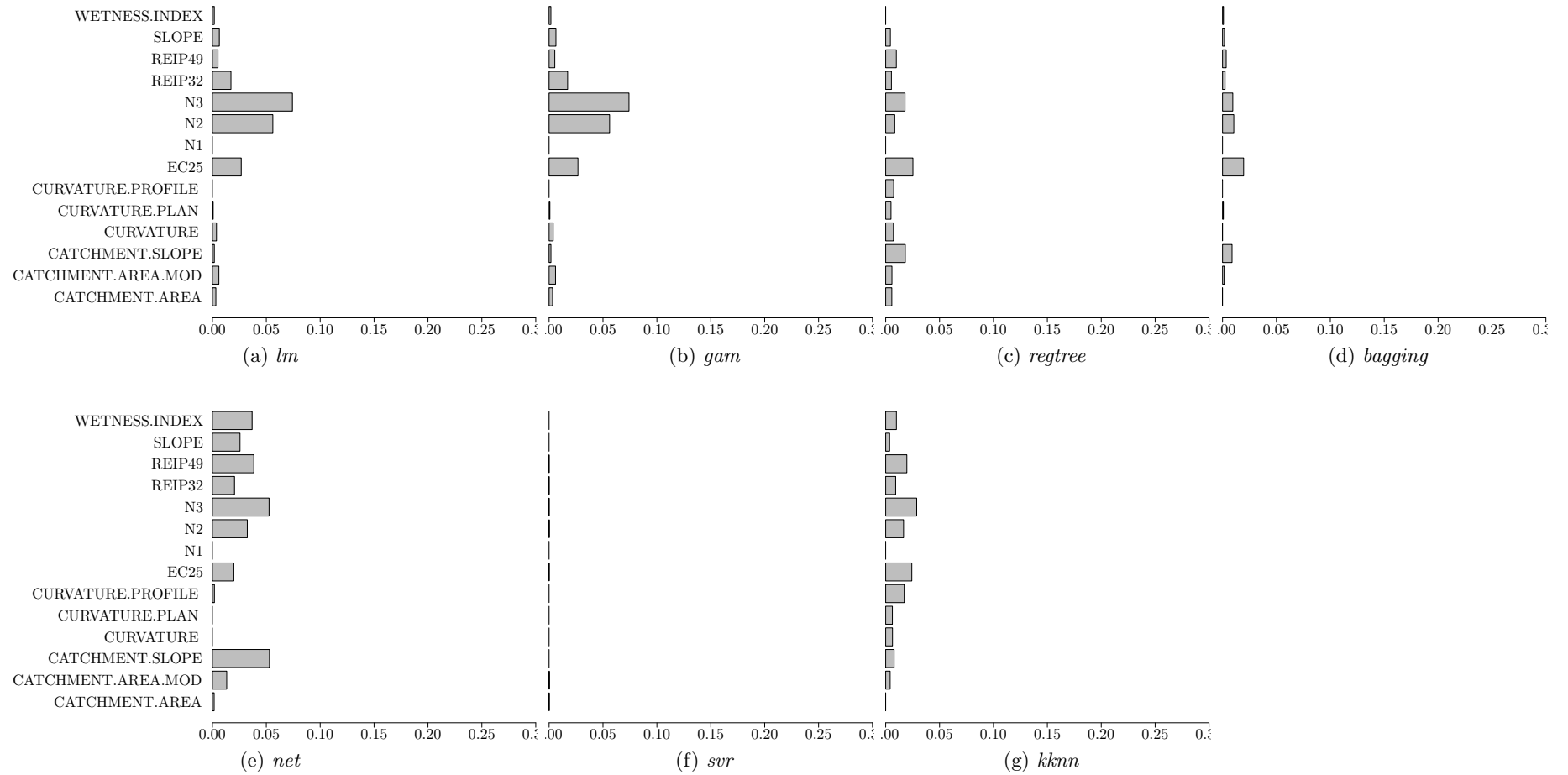
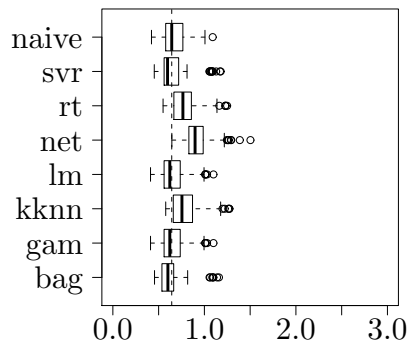
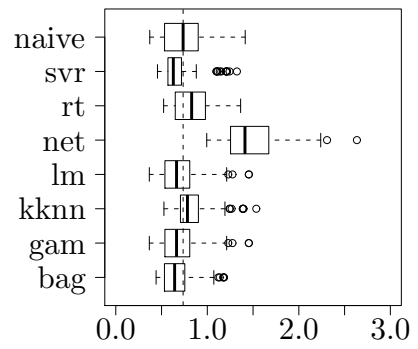


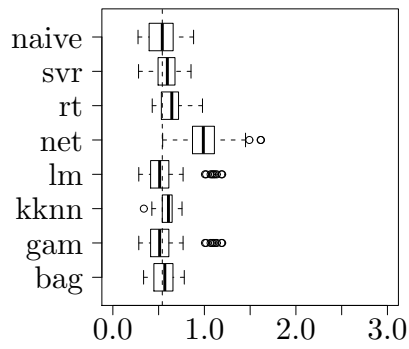
Figure B.12: F440sorte1, strategy "sensor", regression models and spatial variable importance



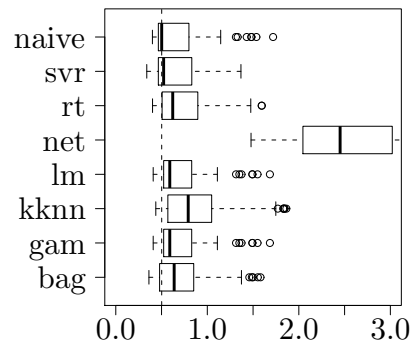
(a) f440sorte2, all strategies, rmse of models



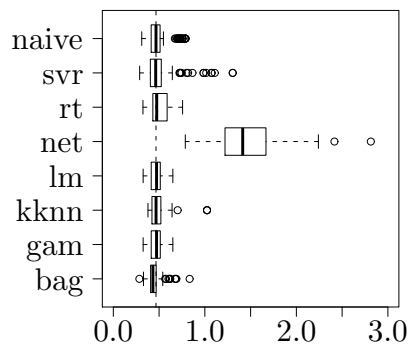
(b) f440sorte2, "low constant fertilization"



(c) f440sorte2, strategy "constant fertilization"



(d) f440sorte2, strategy "neural network"



(e) f440sorte2, strategy "sensor"

Figure B.13: RMSE for F440sorte2 and its subsets (by strategy)

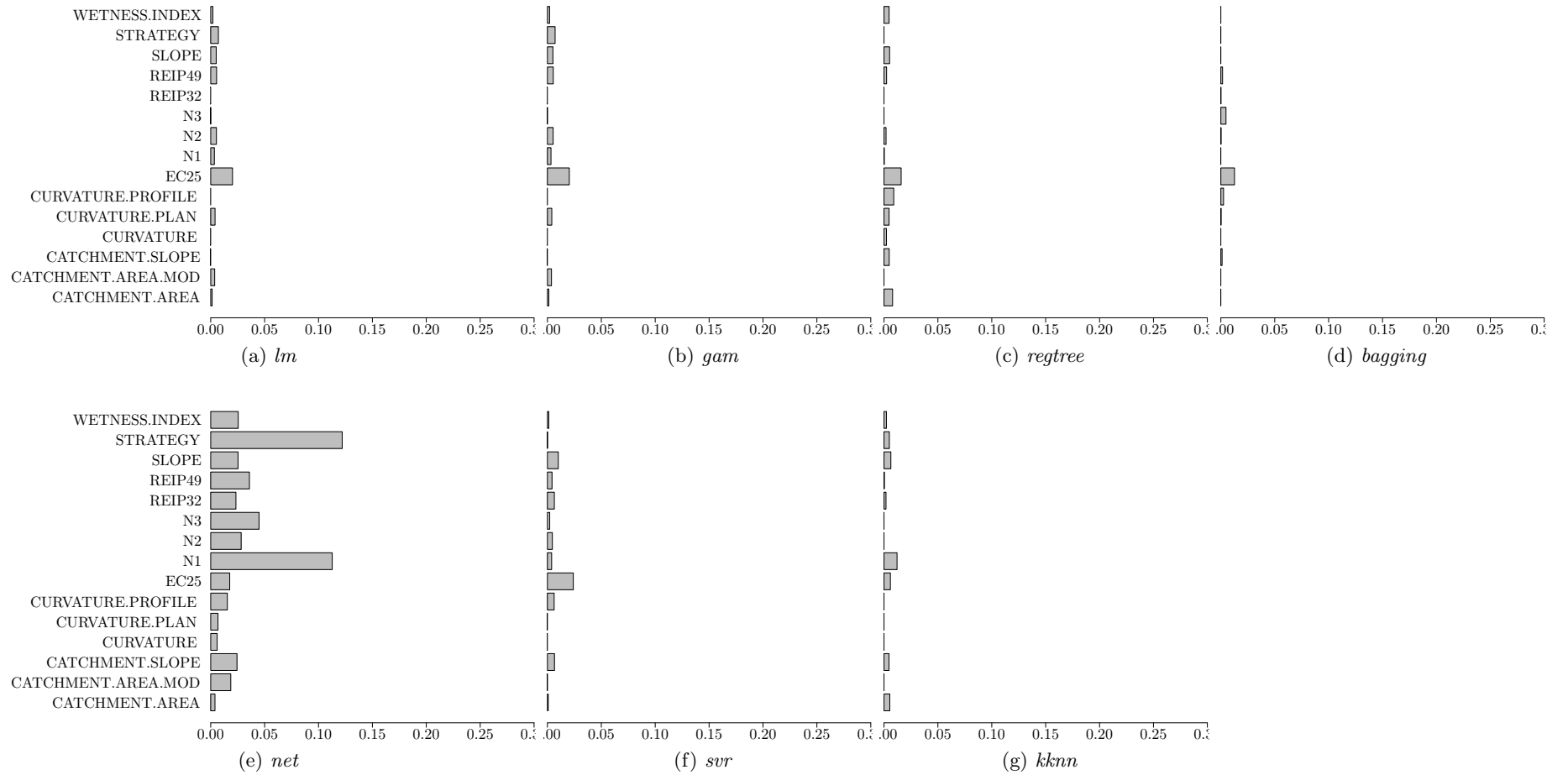


Figure B.14: F440sorte2, all strategies, regression models and spatial variable importance

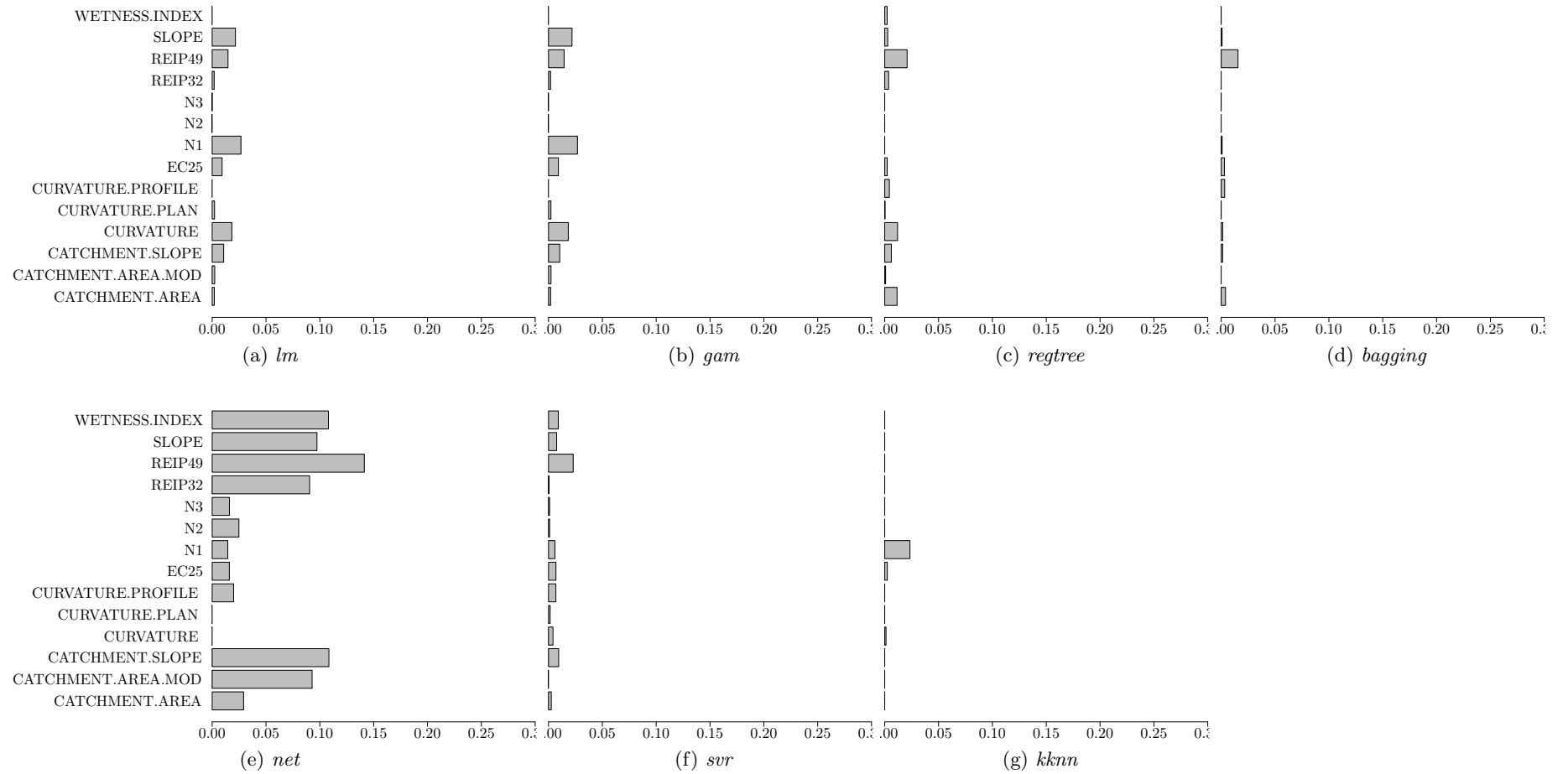


Figure B.15: F440sorte2, strategy “low, constant”, regression models and spatial variable importance

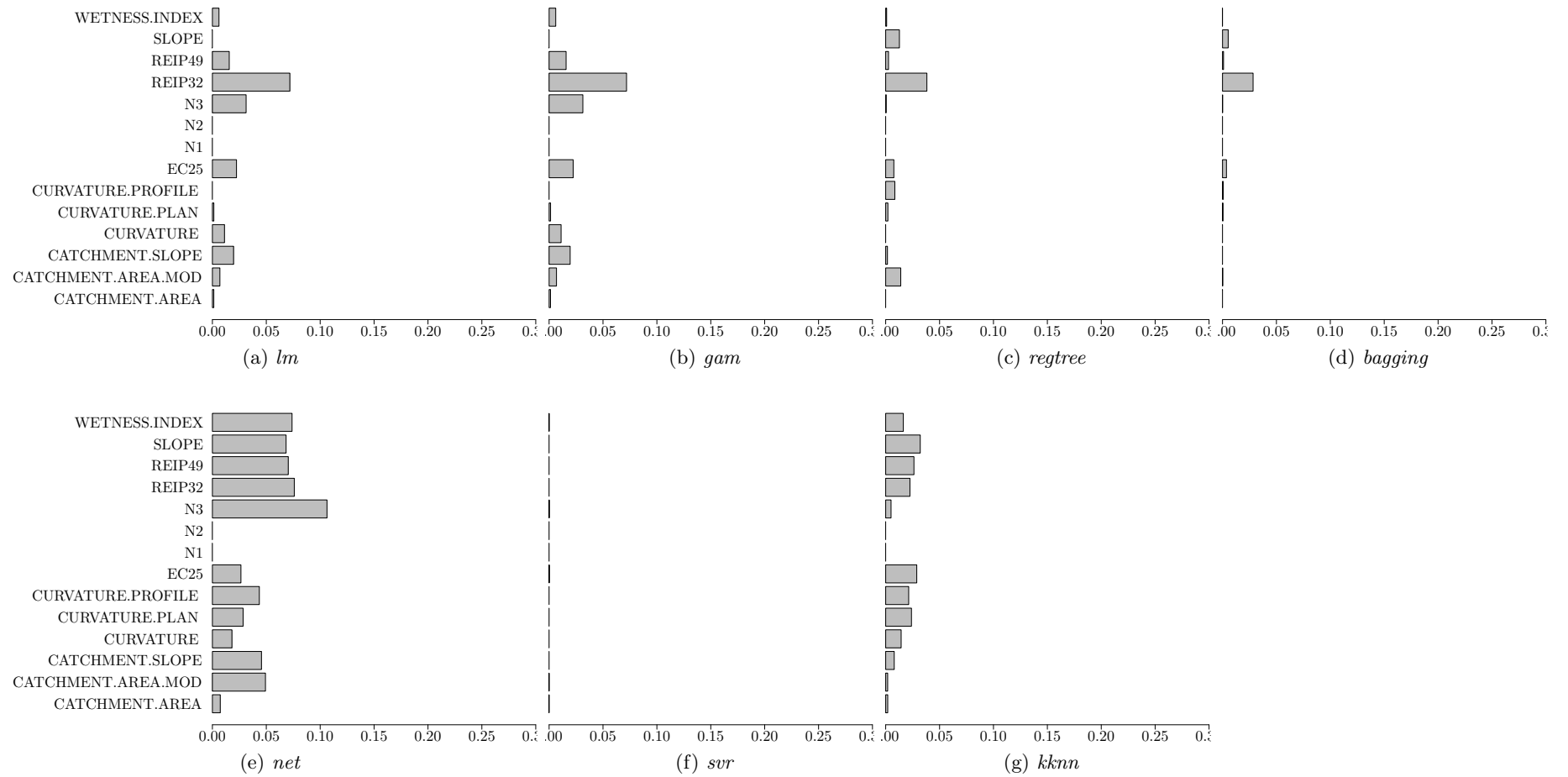


Figure B.16: F440sorte2, strategy “constant”, regression models and spatial variable importance

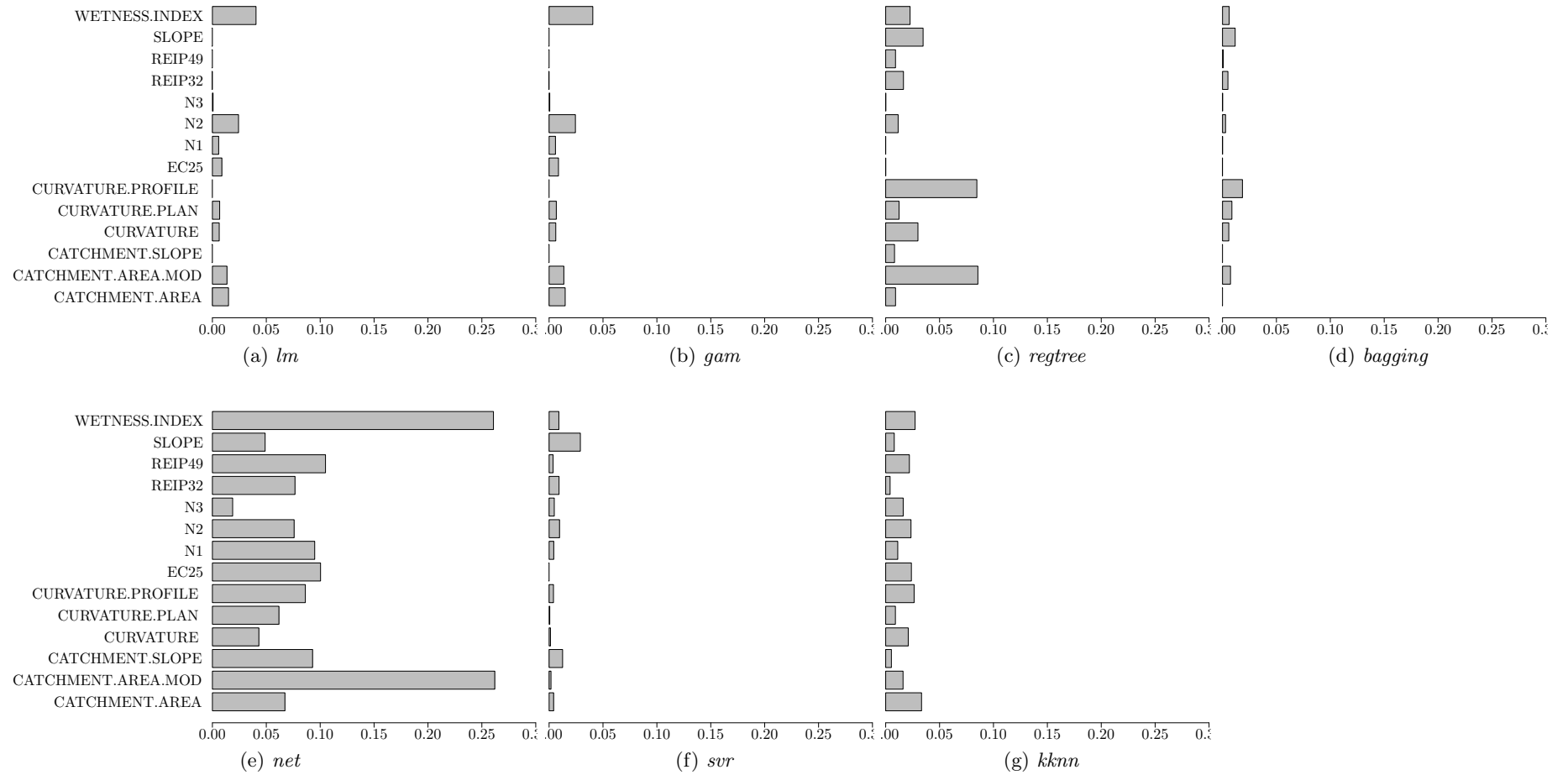


Figure B.17: F440sorte2, strategy “neural network”, regression models and spatial variable importance

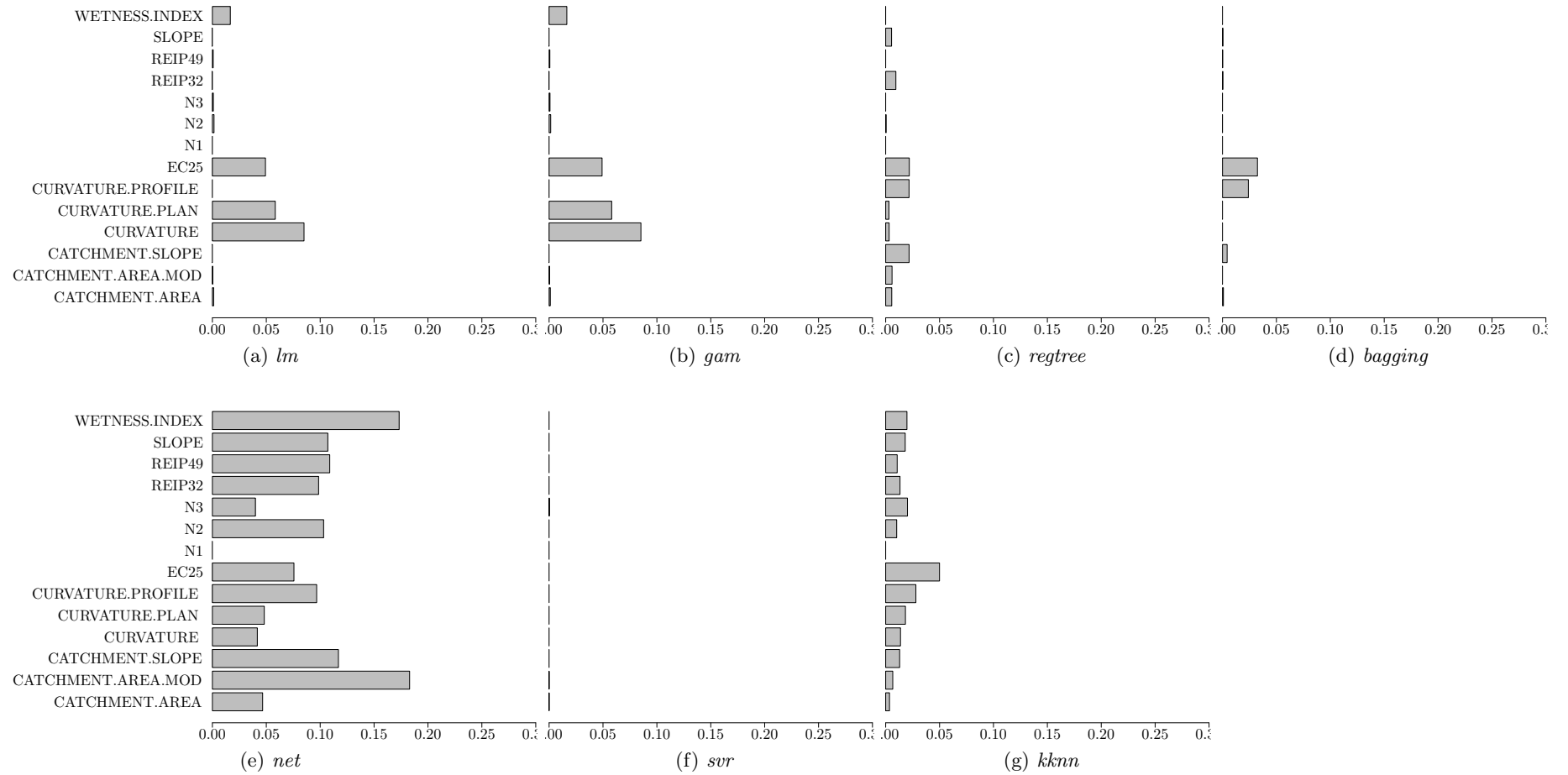
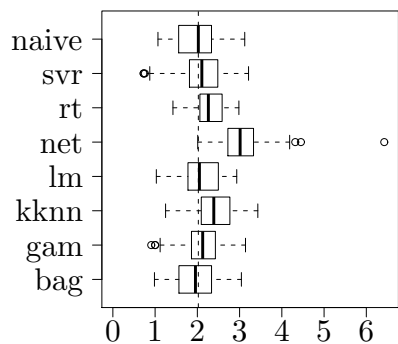
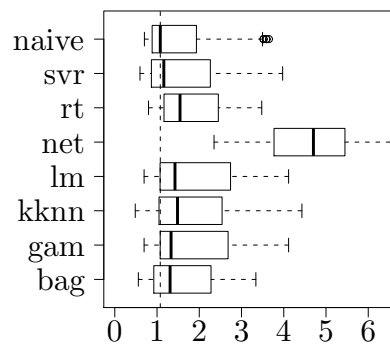


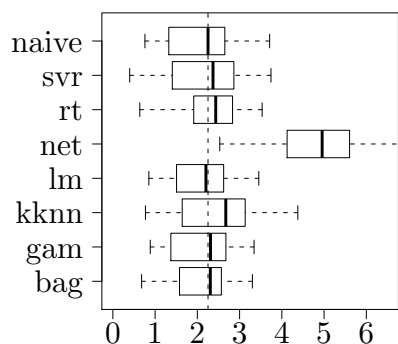
Figure B.18: F440sorte2, strategy “sensor”, regression models and spatial variable importance



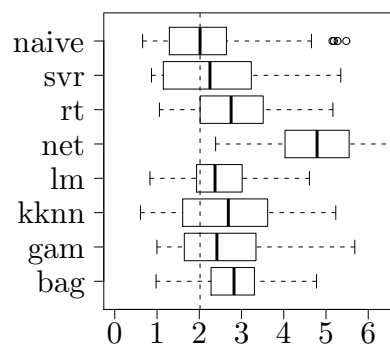
(a) F550, all strategies, rmse of models



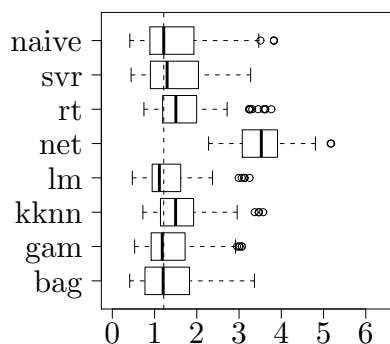
(b) F550, strategy "company"



(c) F550, strategy "mapping"



(d) F550, strategy "N-trial"



(e) F550, strategy "sensor"

Figure B.19: RMSE for F550 and its subsets (by strategy, without YIELD2003)

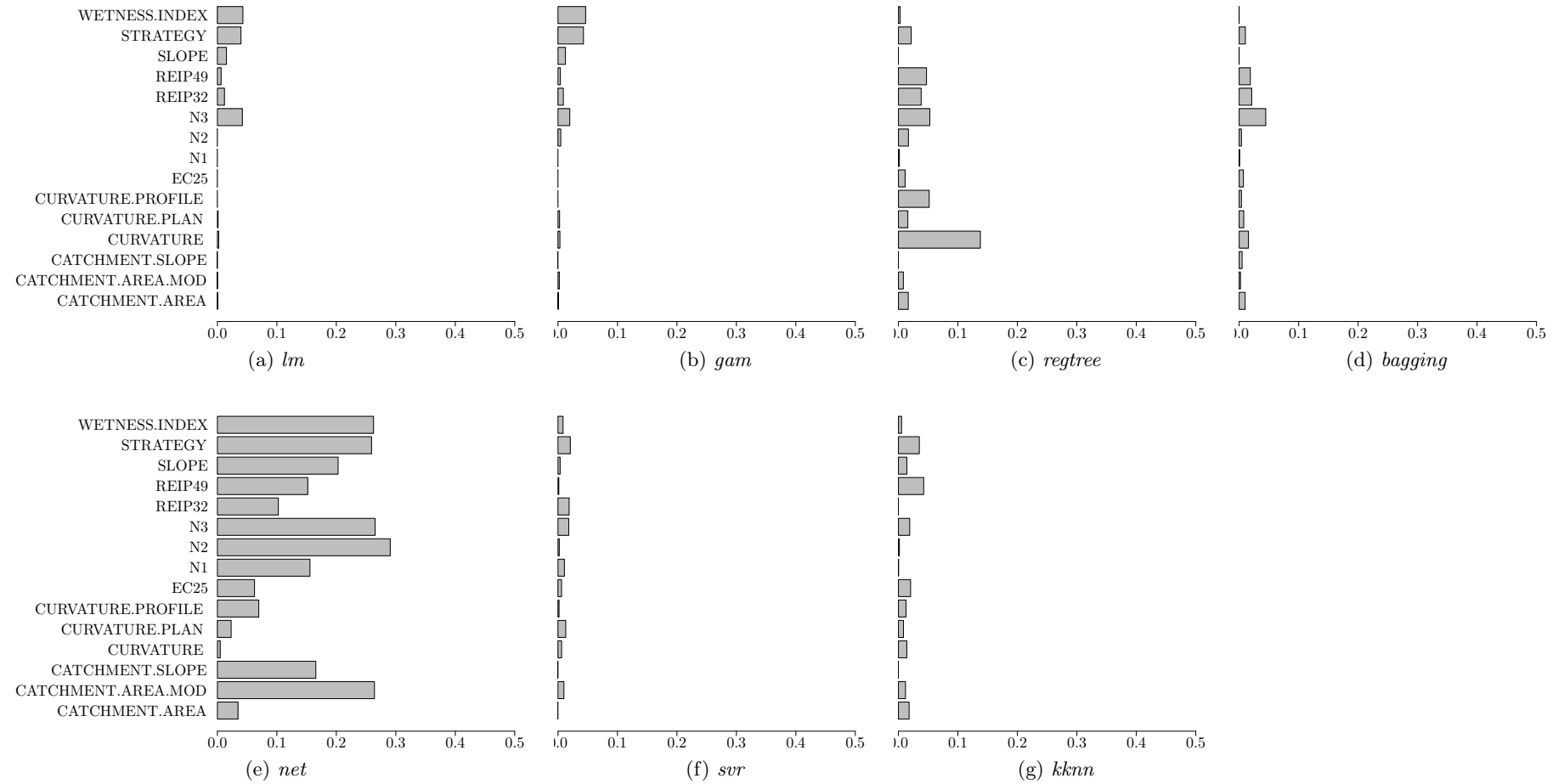


Figure B.20: F550, all strategies combined, without previous year's yield, regression models and spatial variable importance

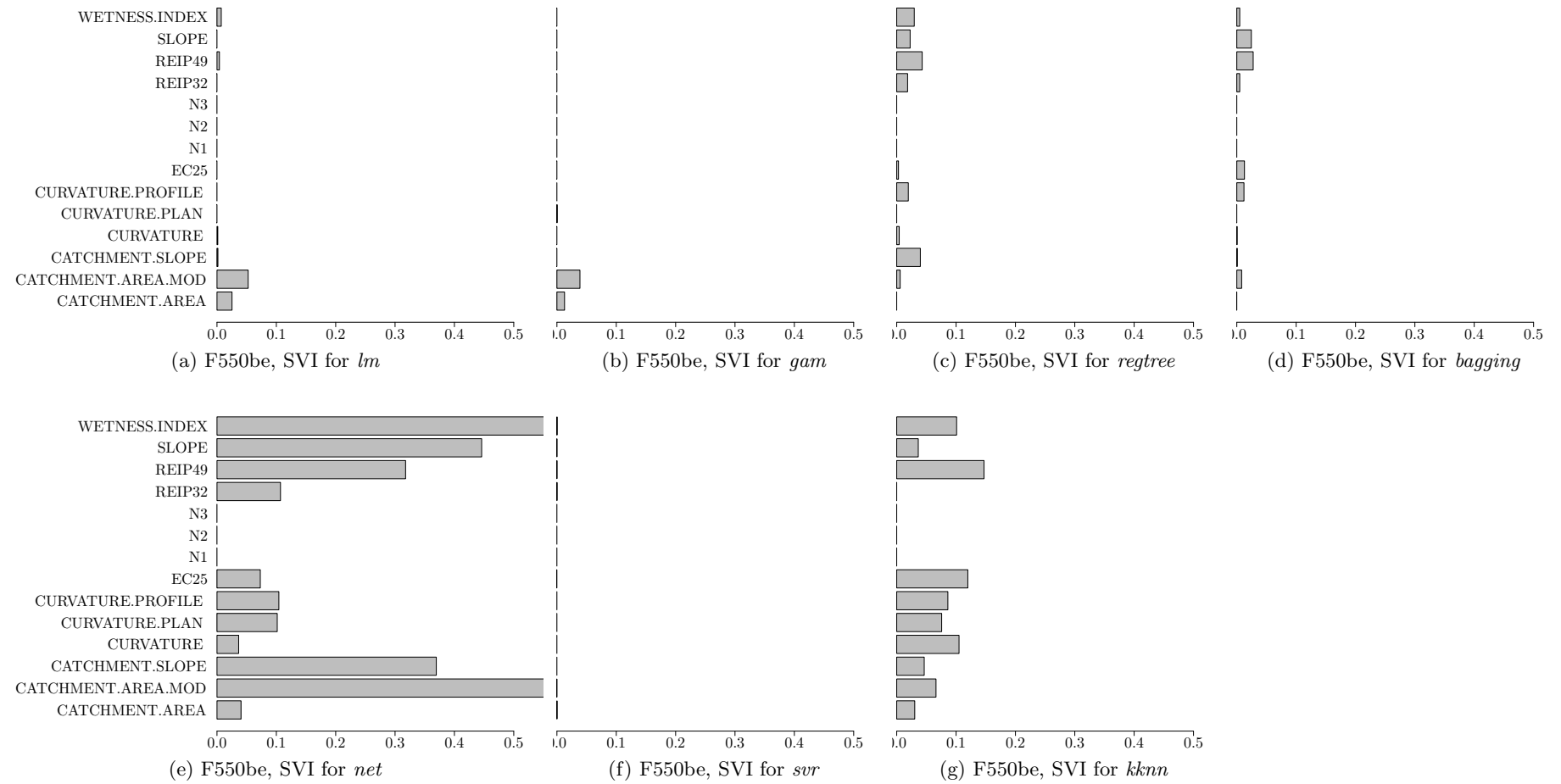


Figure B.21: F550, strategy “company”, without previous year’s yield, regression models and spatial variable importance. The neural network’s WETNESS.INDEX and CATCHMENT.AREA.MOD variables’ SVI values are 0.7 and 0.62, respectively (Figure (e)).

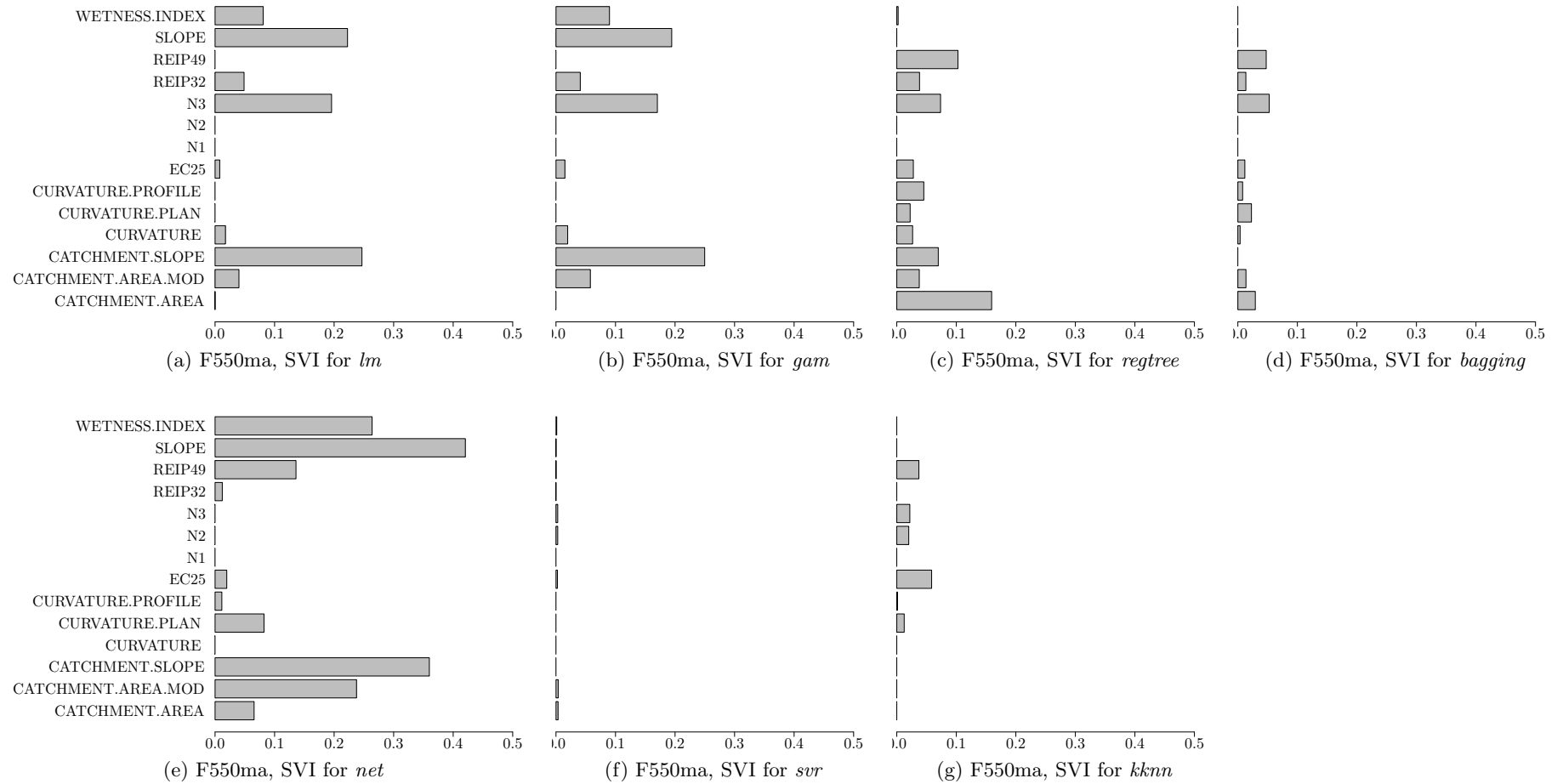


Figure B.22: F550, strategy “mapping”, without previous year’s yield, regression models and spatial variable importance

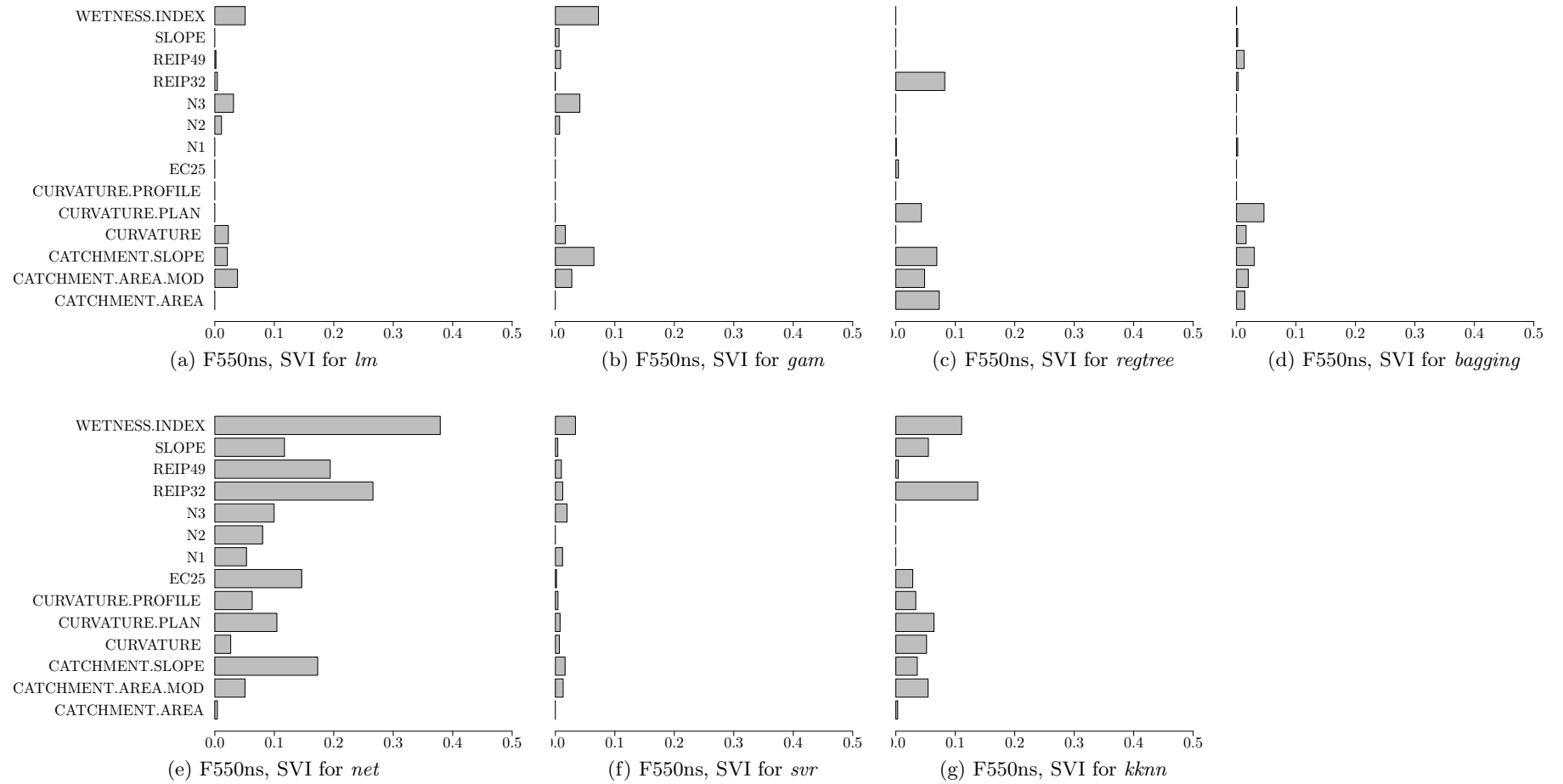


Figure B.23: F550, strategy “N-trial”, without previous year’s yield, regression models and spatial variable importance

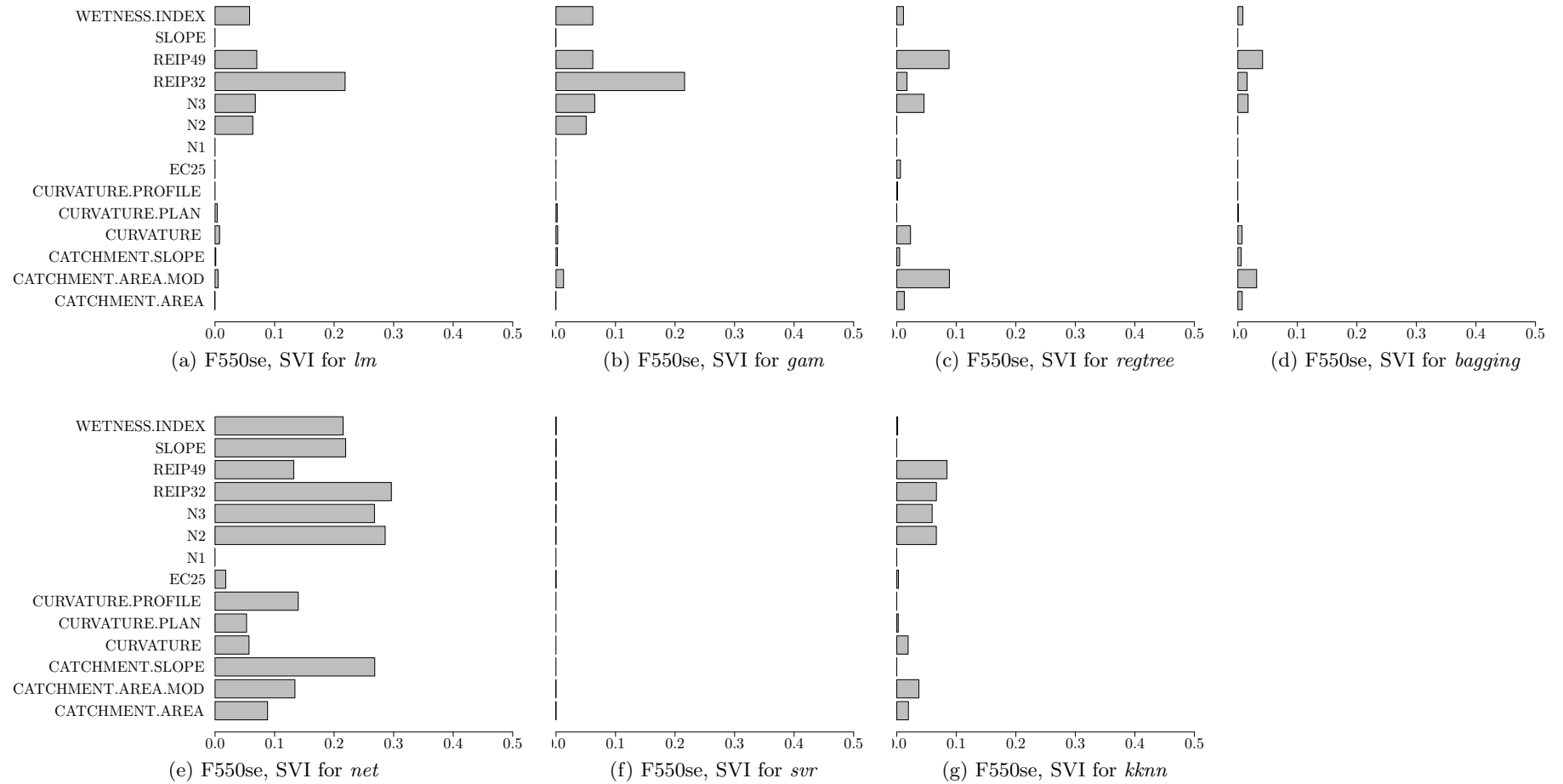
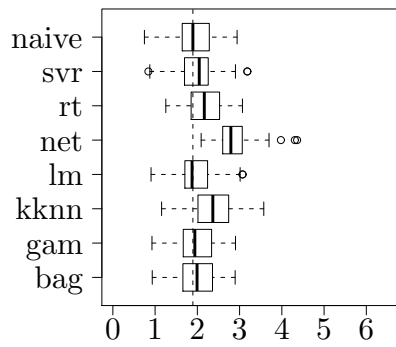
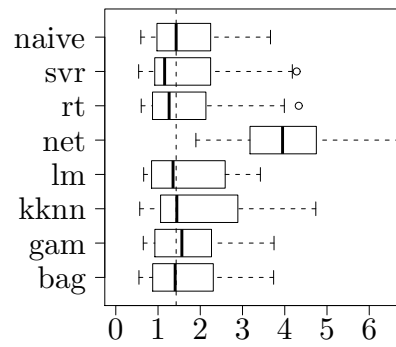


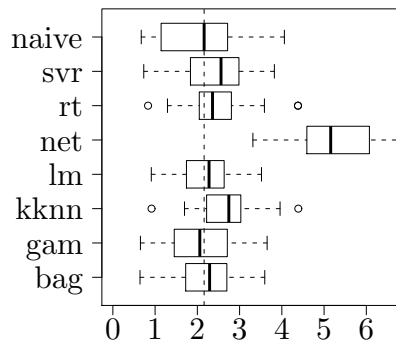
Figure B.24: F550se, strategy “sensor”, without previous year’s yield, regression models and spatial variable importance



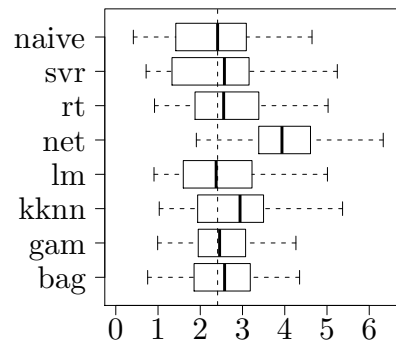
(a) F550, all strategies, rmse of models



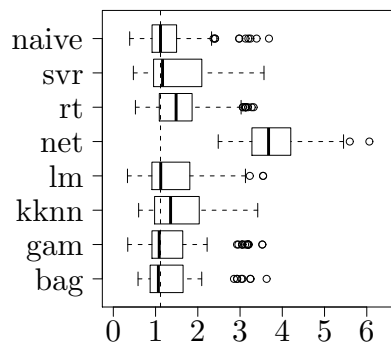
(b) F550, strategy "company"



(c) F550, strategy "mapping"



(d) F550, strategy "N-trial"



(e) F550, strategy "sensor"

Figure B.25: RMSE for F550 and its subsets (by strategy, with YIELD2003)

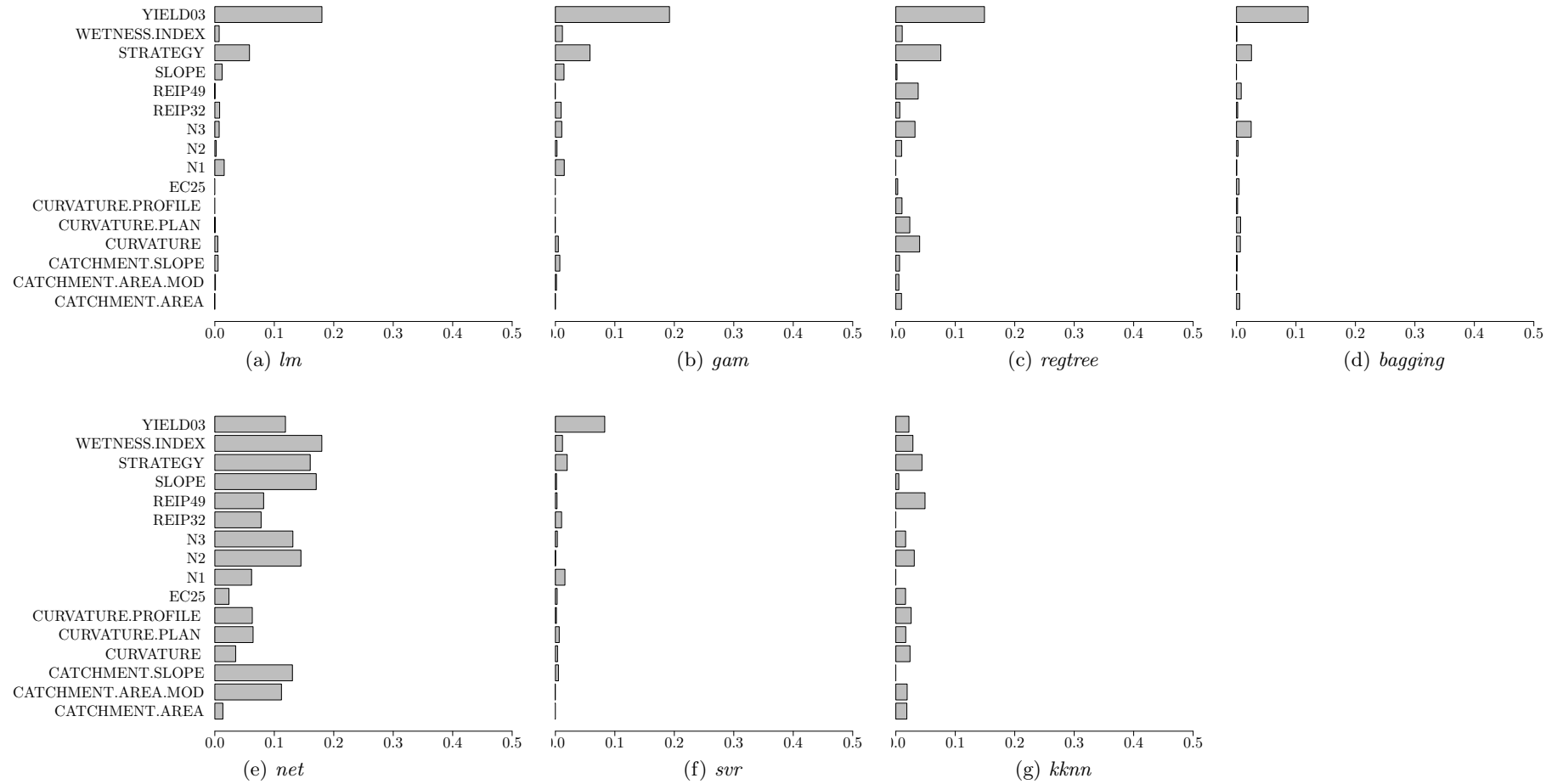


Figure B.26: F550, all strategies combined, with previous year's yield, regression models and spatial variable importance

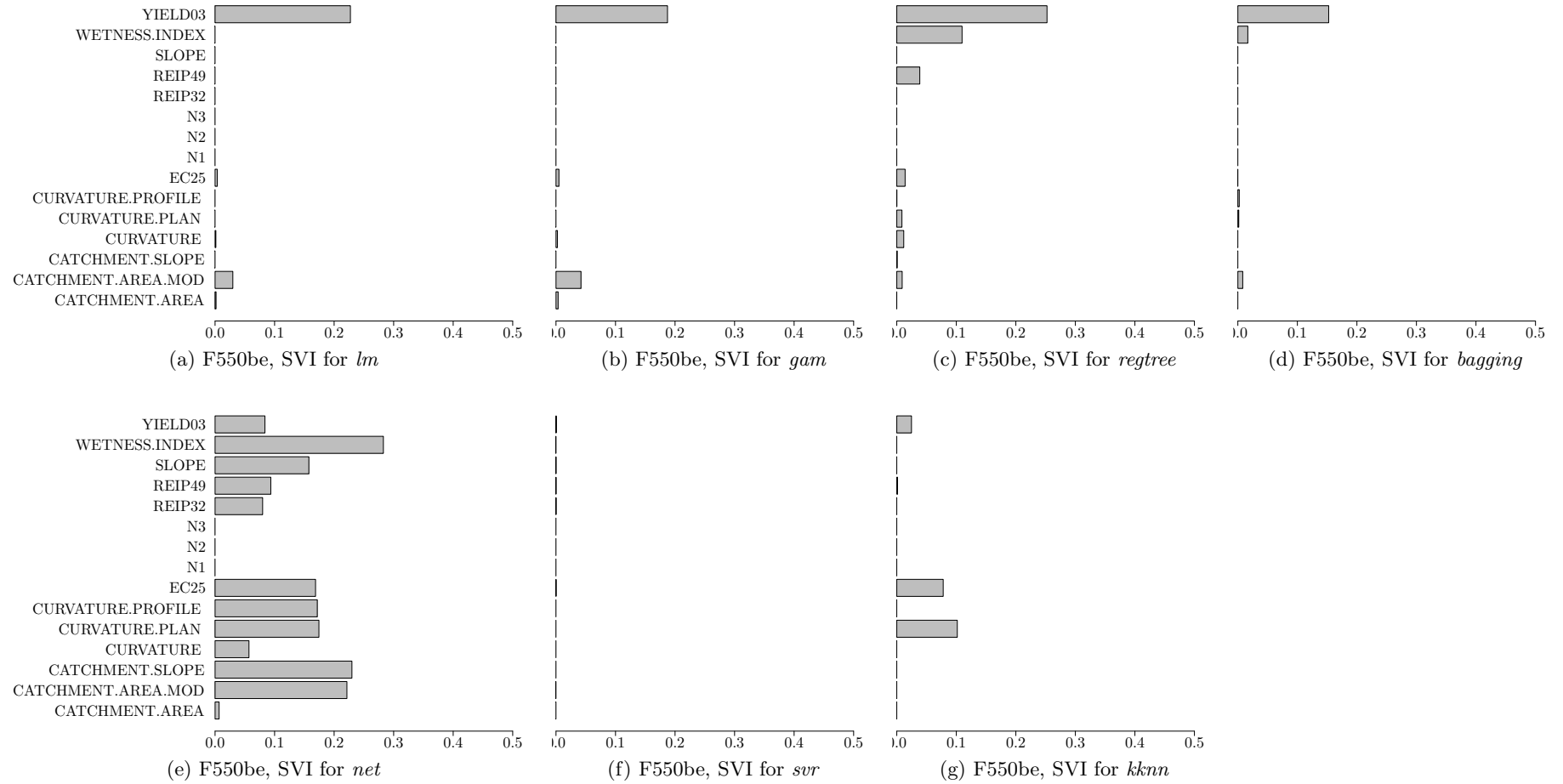


Figure B.27: F550be, strategy “company”, with previous year’s yield, regression models and spatial variable importance

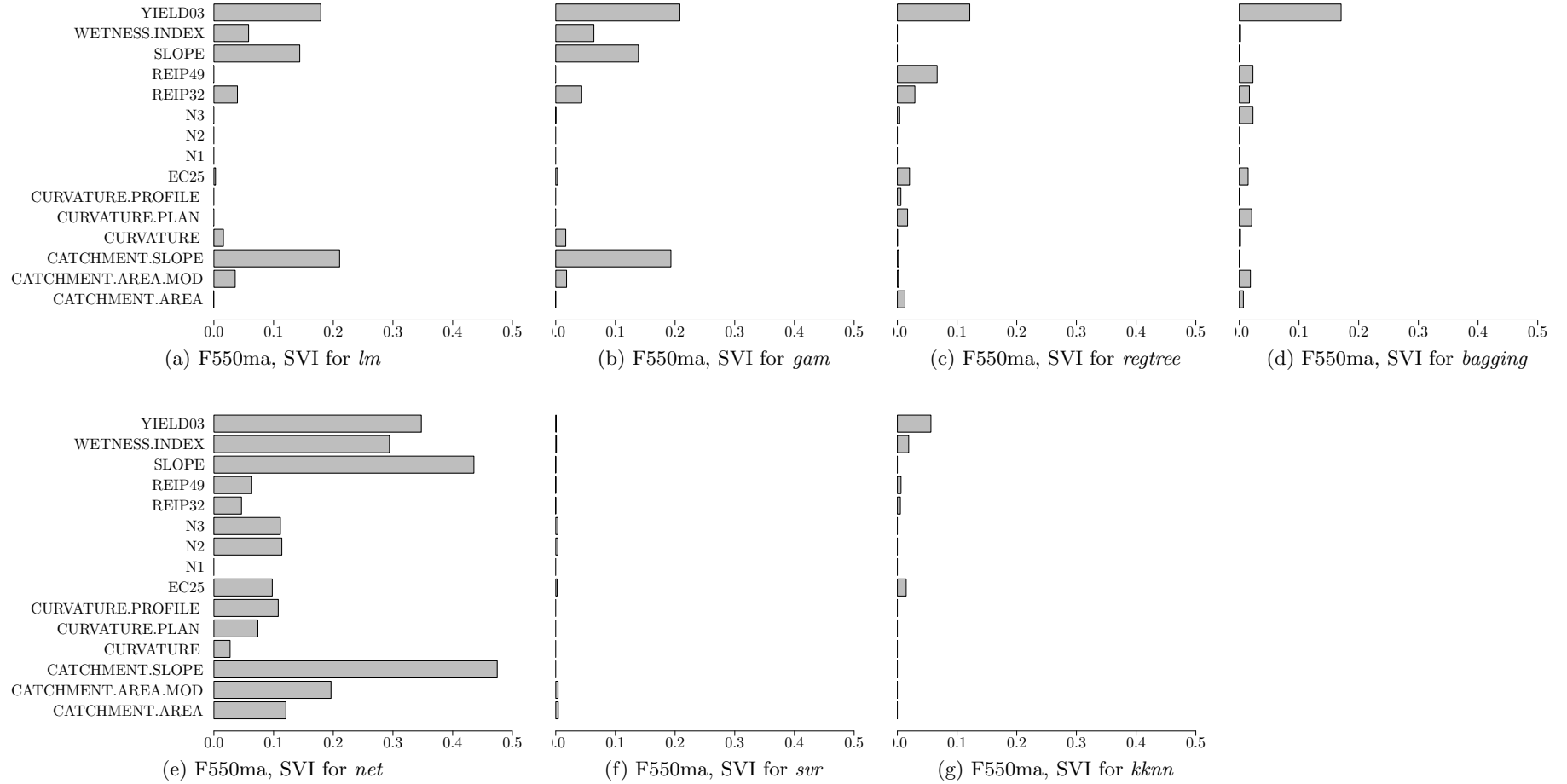


Figure B.28: F550, strategy “mapping”, with previous year’s yield, regression models and spatial variable importance

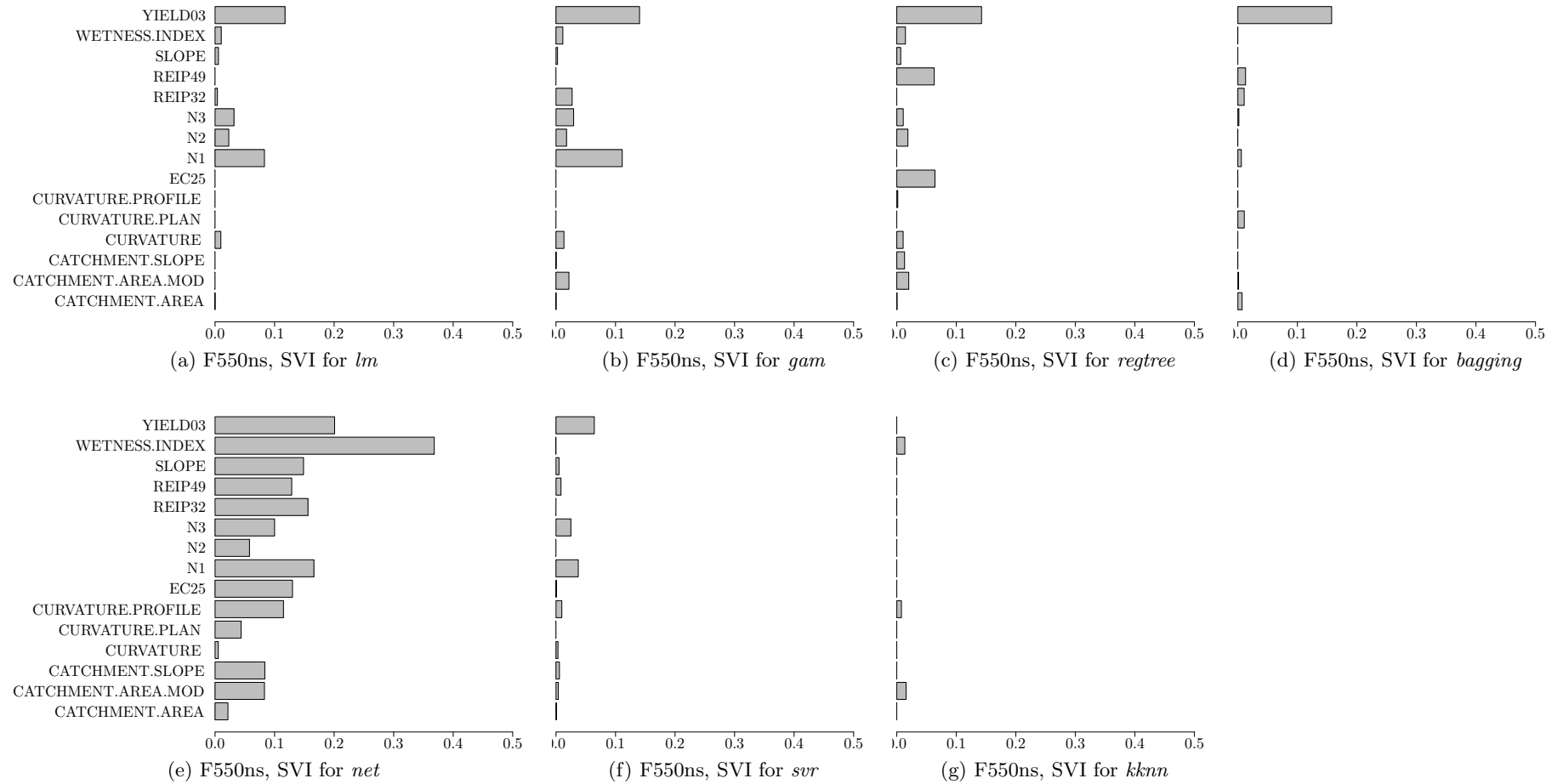


Figure B.29: F550, strategy “N-trial”, with previous year’s yield, regression models and spatial variable importance

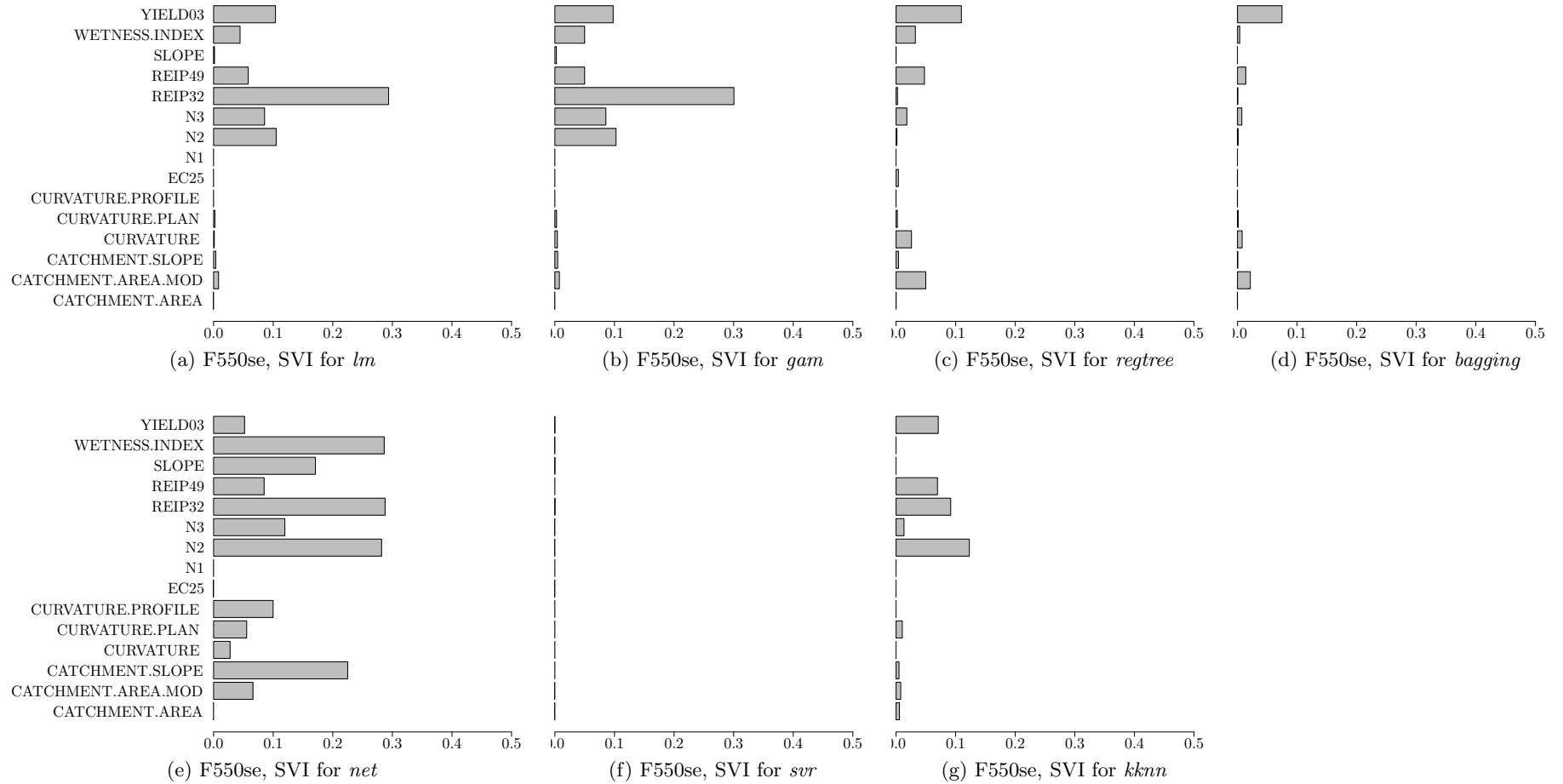
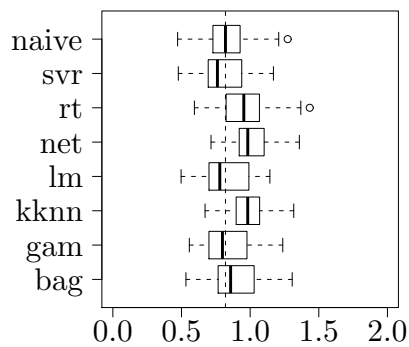
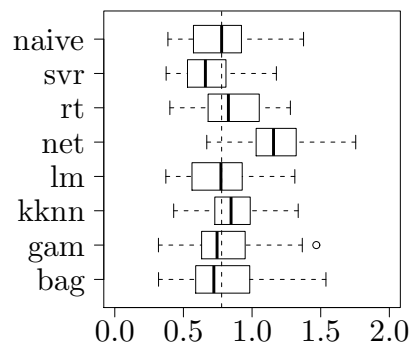


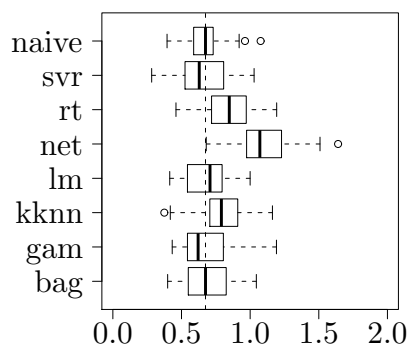
Figure B.30: F550se, strategy “sensor”, with previous year’s yield, regression models and spatial variable importance



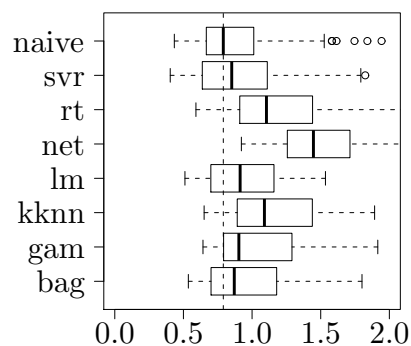
(a) F610, all strategies, rmse of models



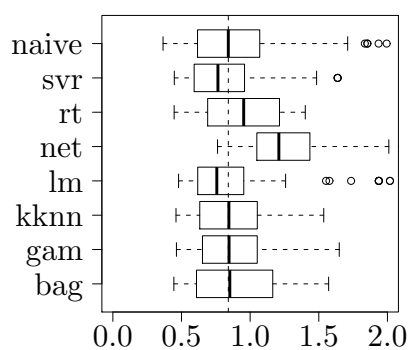
(b) F610, strategy "constant"



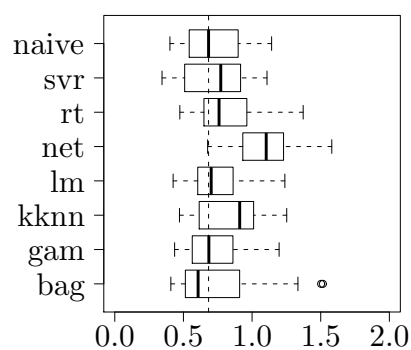
(c) F610, strategy "neural network"



(d) F610, strategy "N-Trial"



(e) F610, strategy "sensor 1"



(f) F610, strategy "sensor 2"

Figure B.31: RMSE for F610 and its subsets (by strategy)

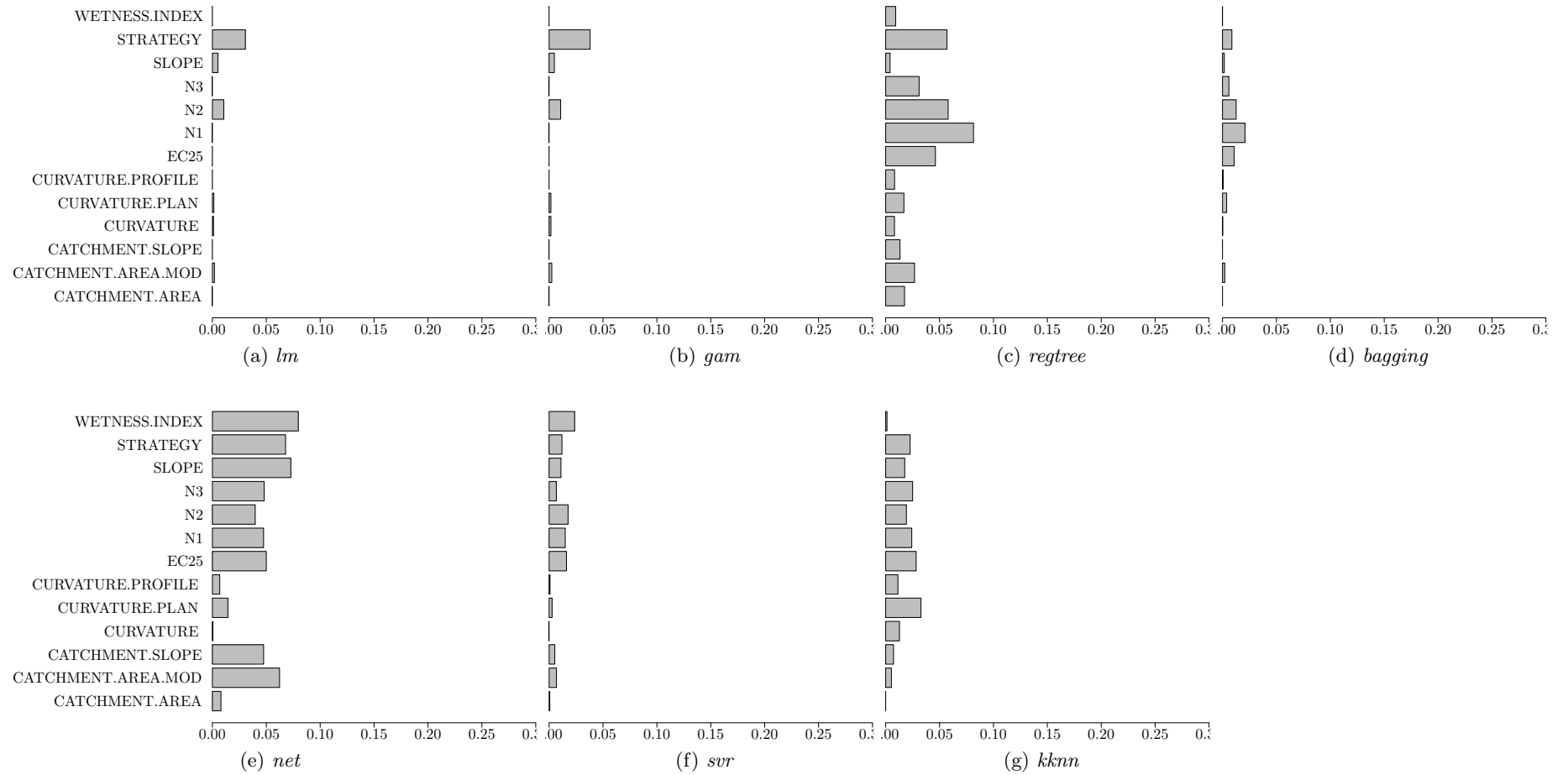


Figure B.32: F610, all strategies combined, regression models and spatial variable importance

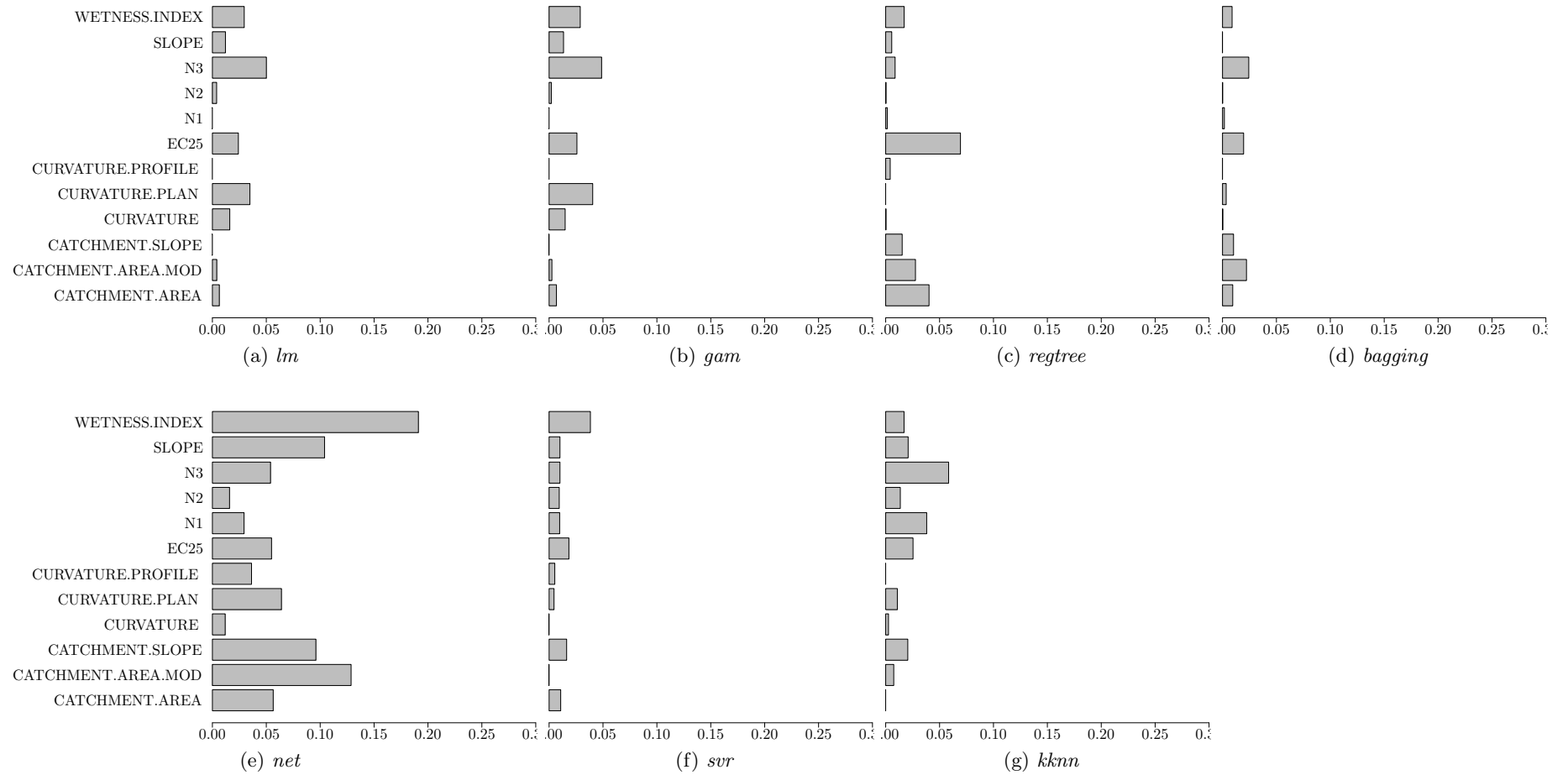


Figure B.33: F610, strategy "constant", regression models and spatial variable importance

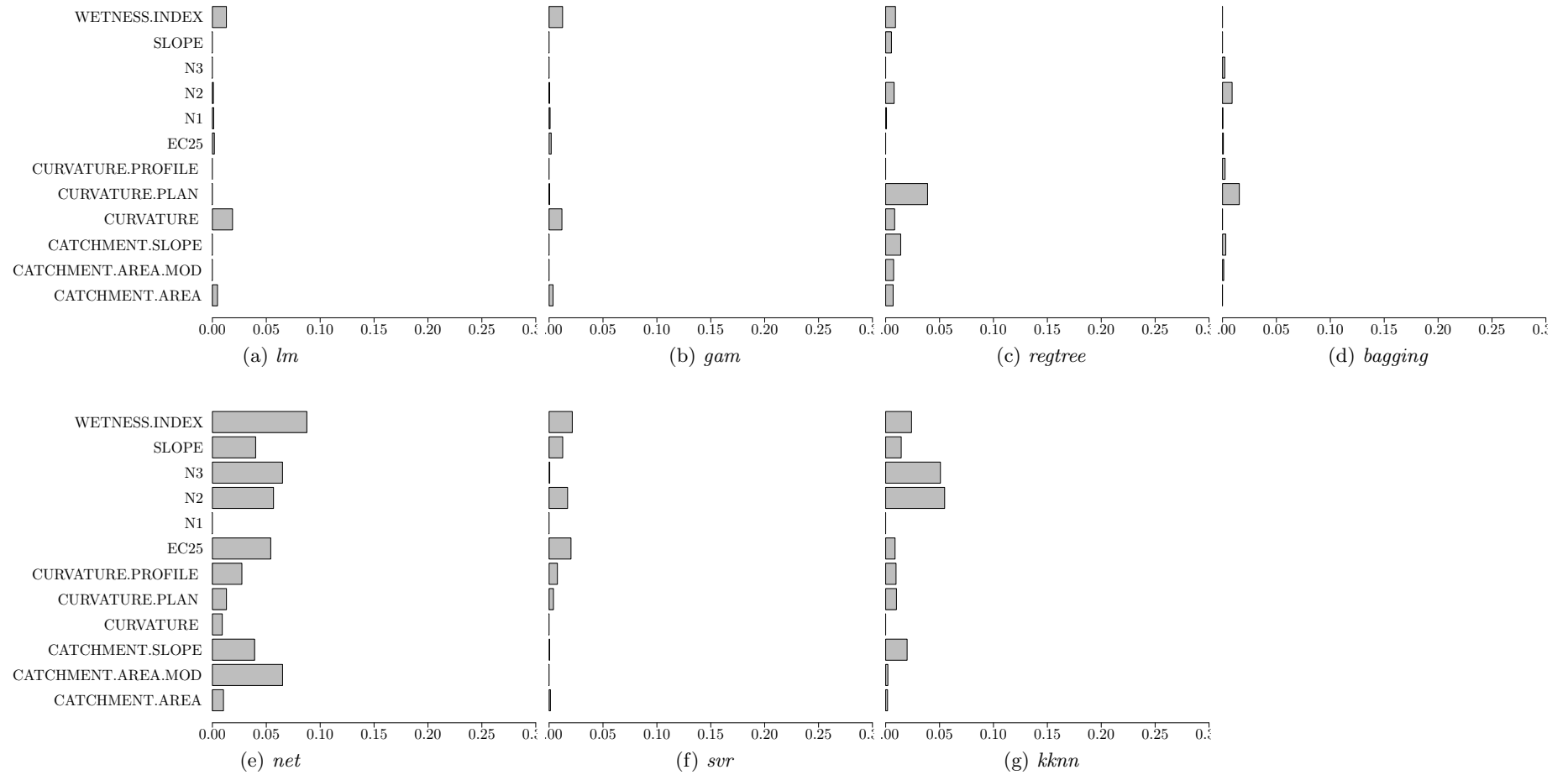


Figure B.34: F610, strategy “neural network”, regression models and spatial variable importance

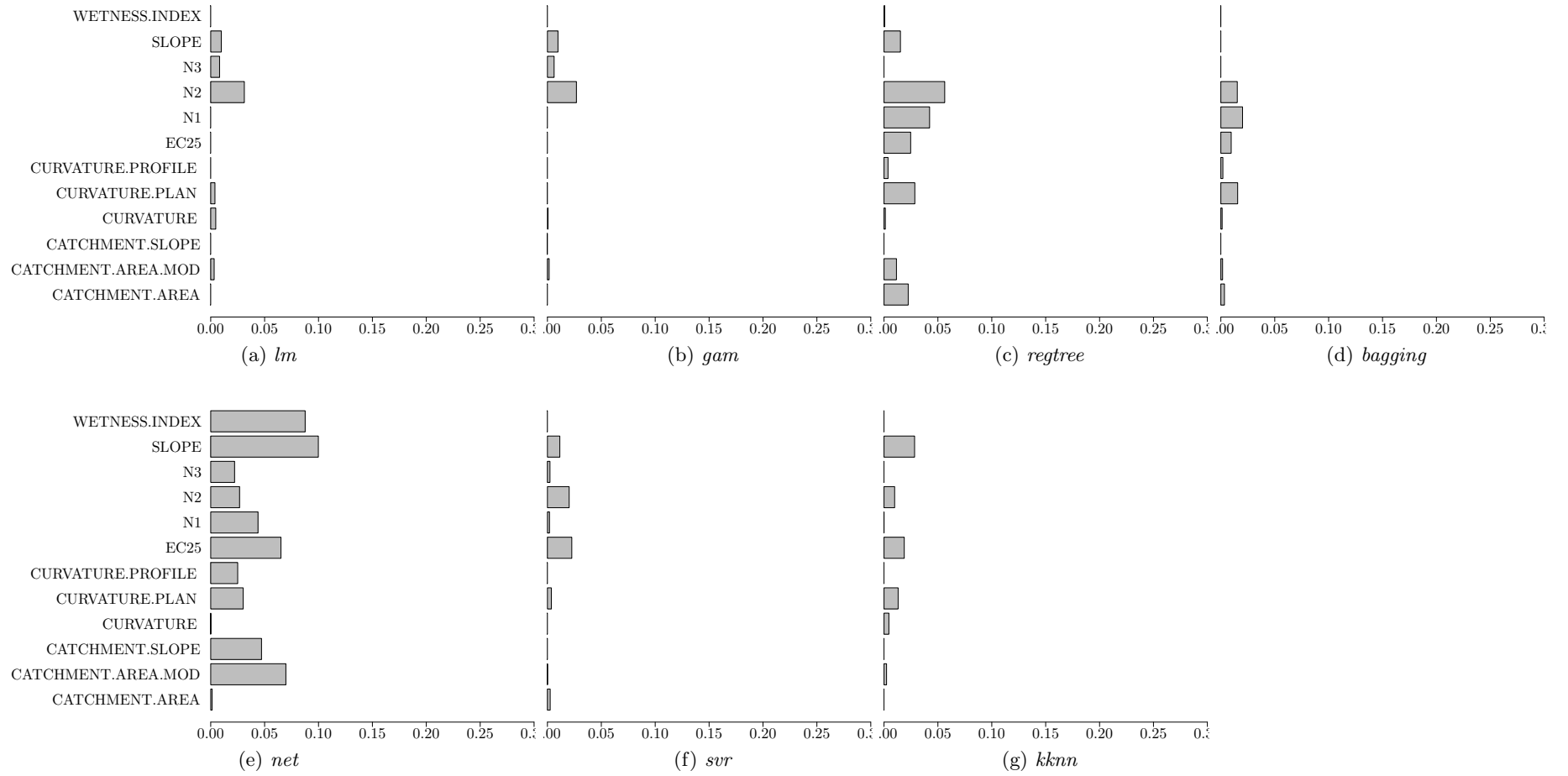


Figure B.35: F610, strategy “N-trial”, regression models and spatial variable importance

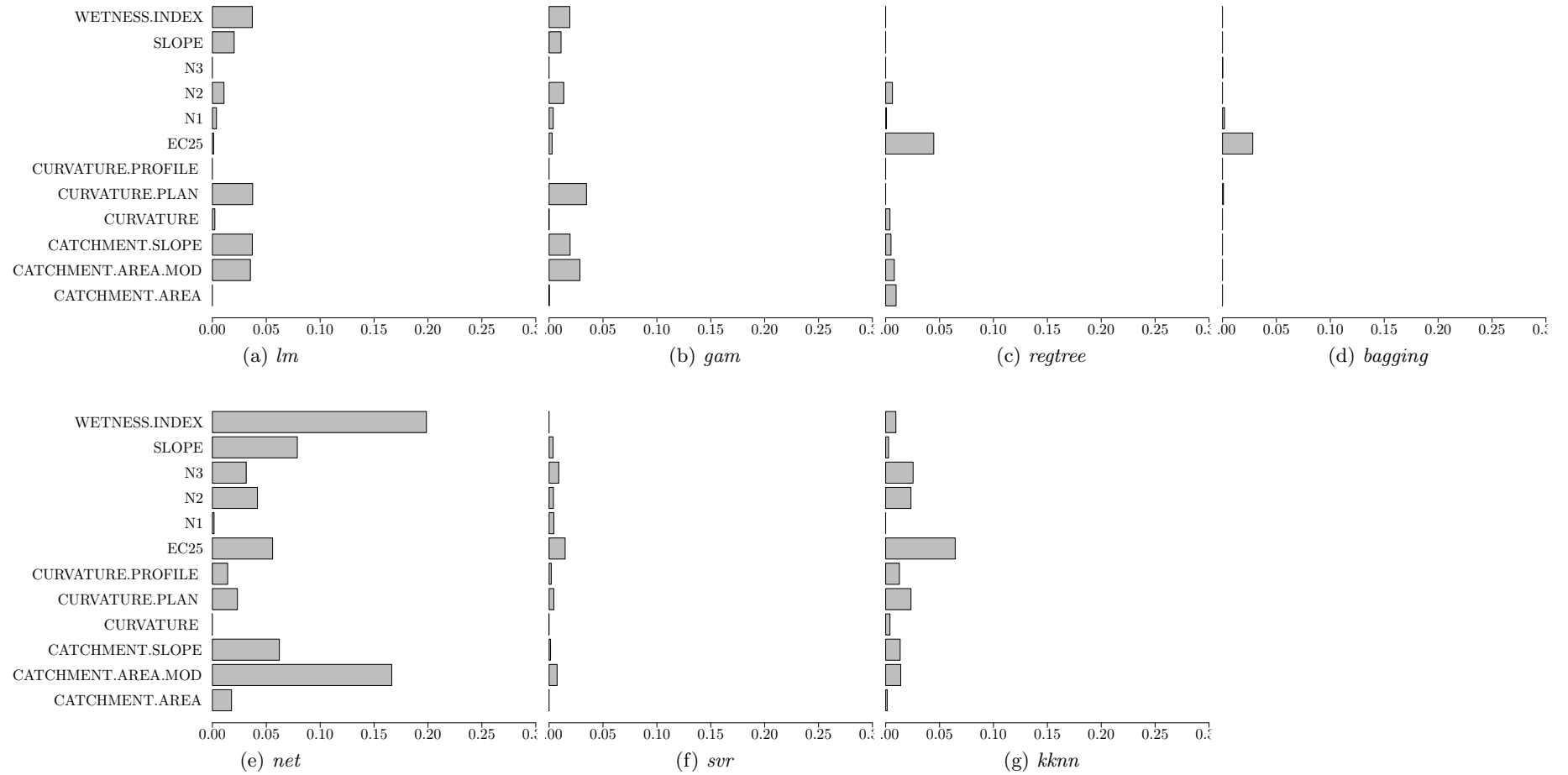


Figure B.36: F610, strategy “sensor 1”, regression models and spatial variable importance

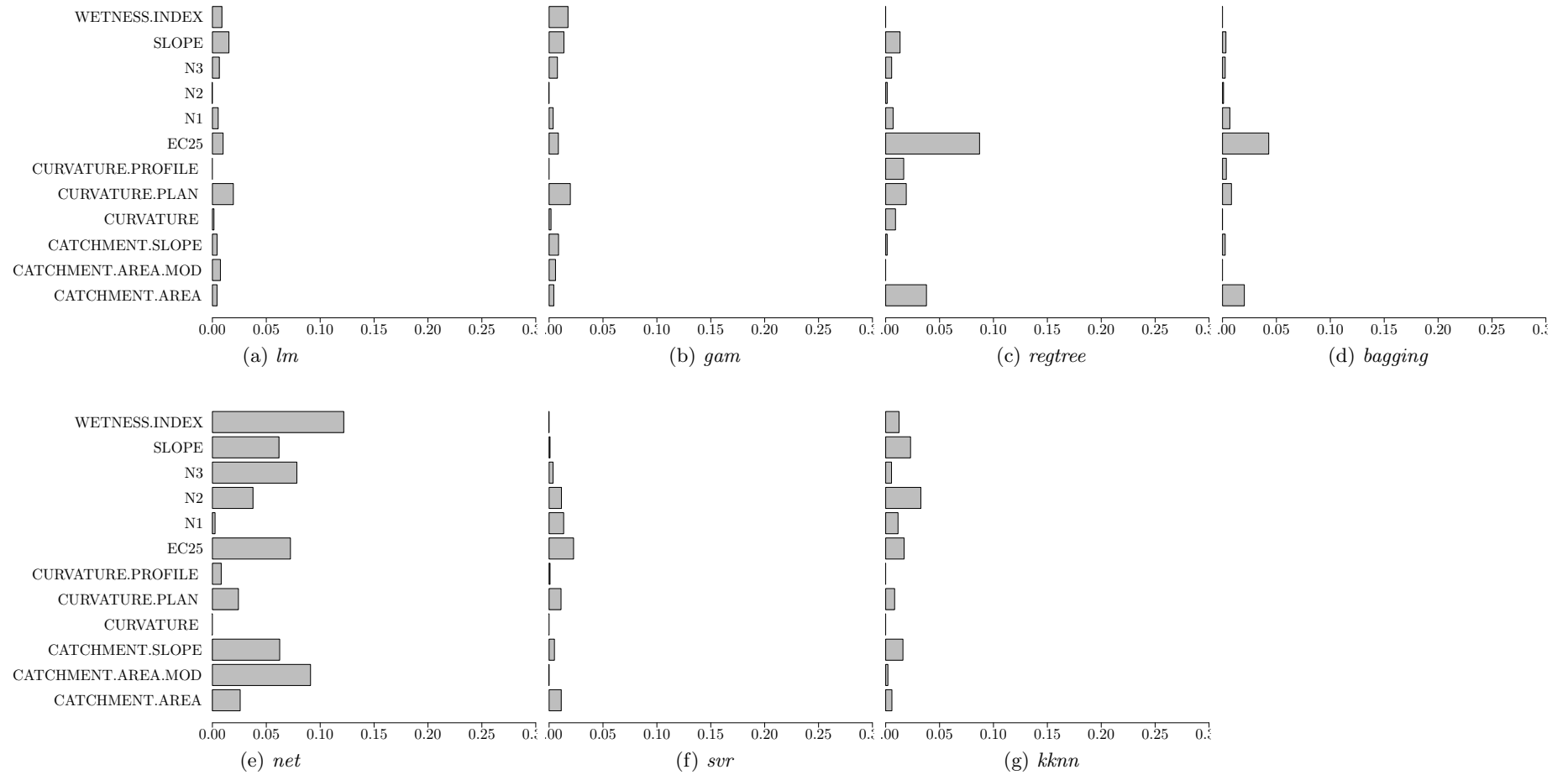
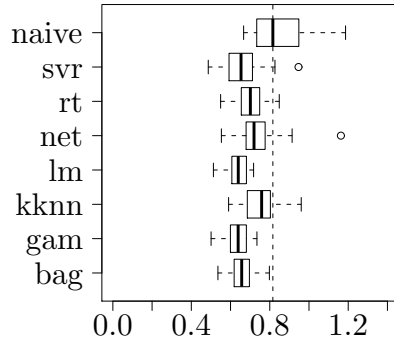
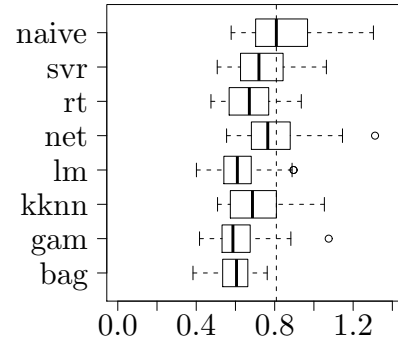


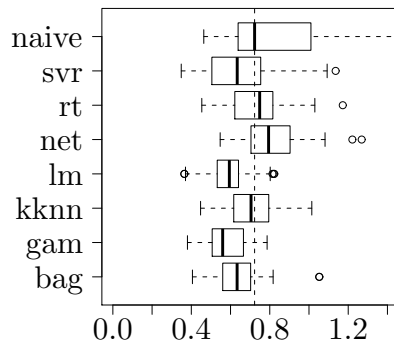
Figure B.37: F610, strategy “sensor 2”, regression models and spatial variable importance



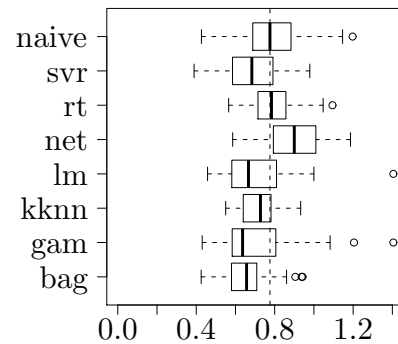
(a) F611, all strategies, rmse of models



(b) F611, strategy "constant"



(c) F611, strategy "neural network"



(d) F611, strategy "sensor"

Figure B.38: RMSE for F611 and its subsets (by strategy)

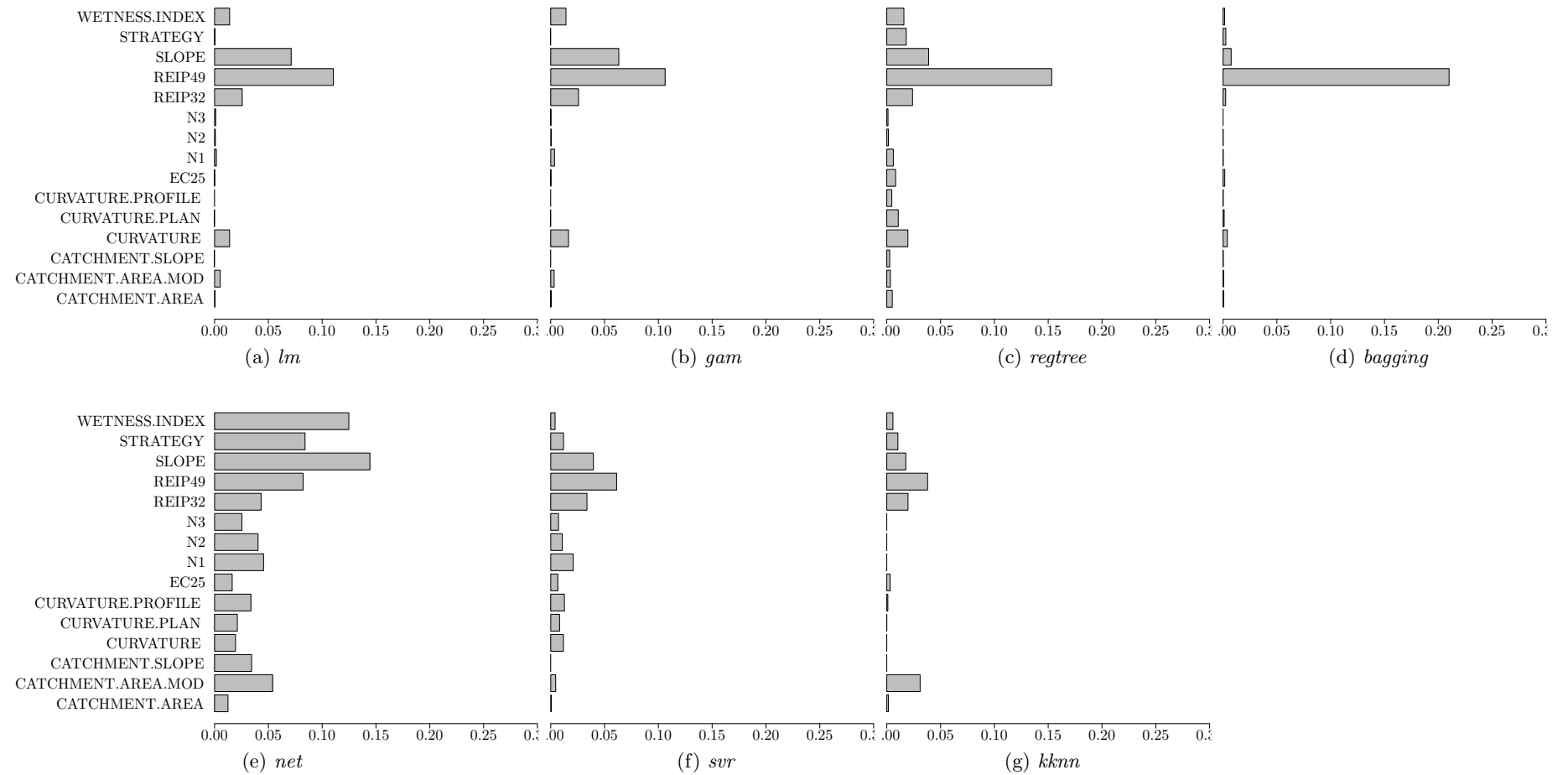


Figure B.39: F611, all strategies combined, regression models and spatial variable importance

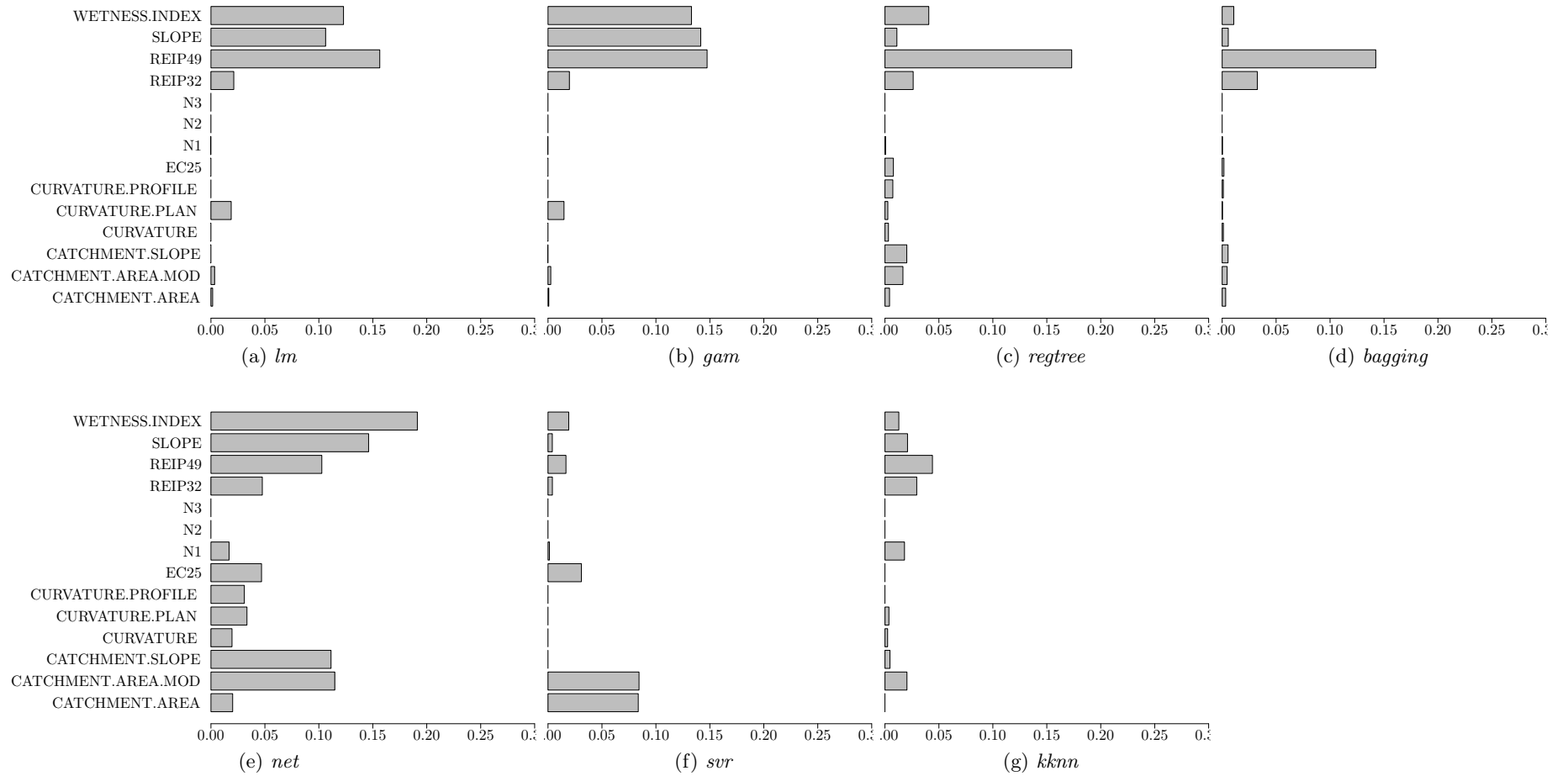


Figure B.40: F611, strategy “constant”, regression models and spatial variable importance

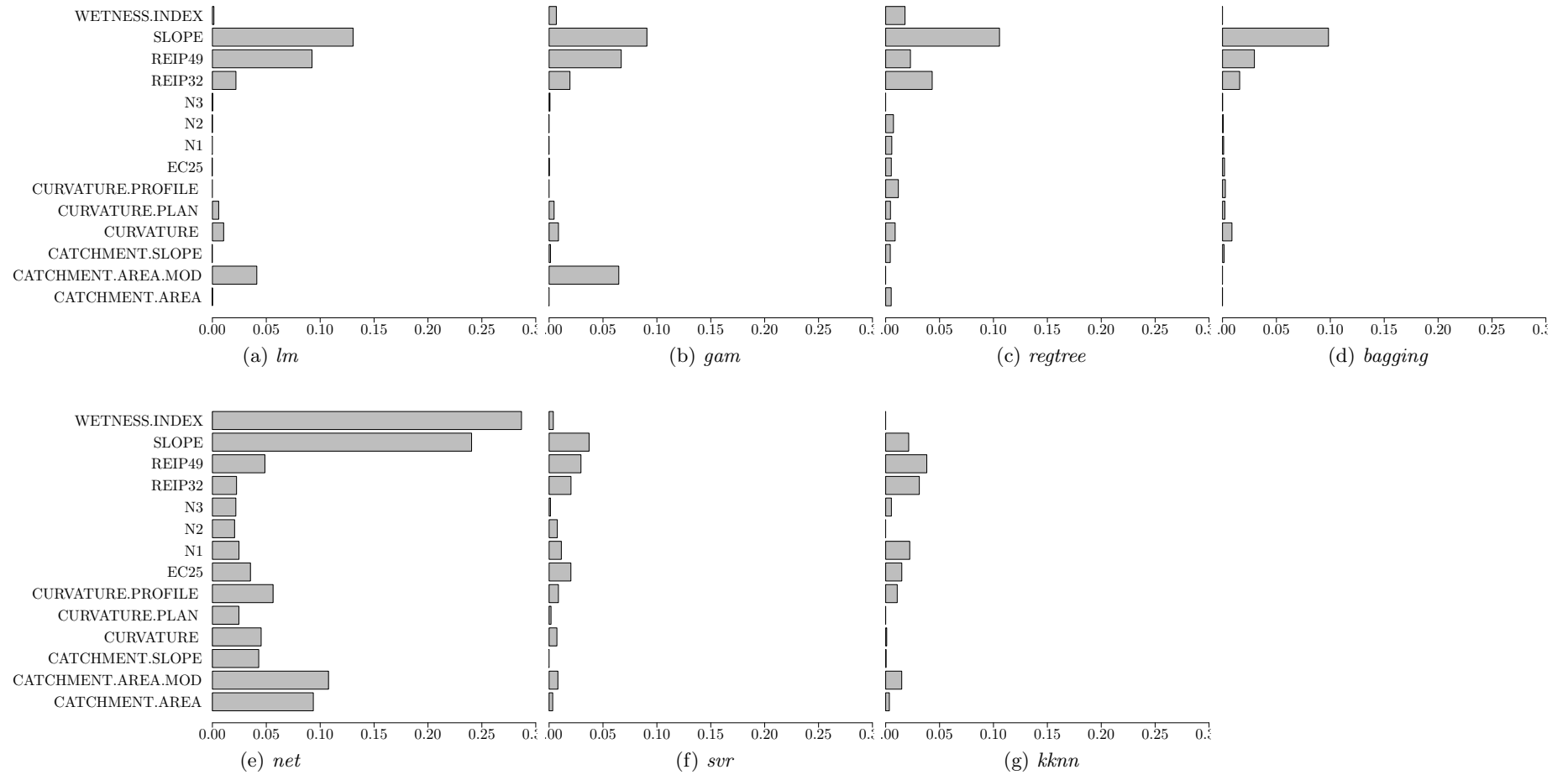


Figure B.41: F611, strategy “neural network”, regression models and spatial variable importance

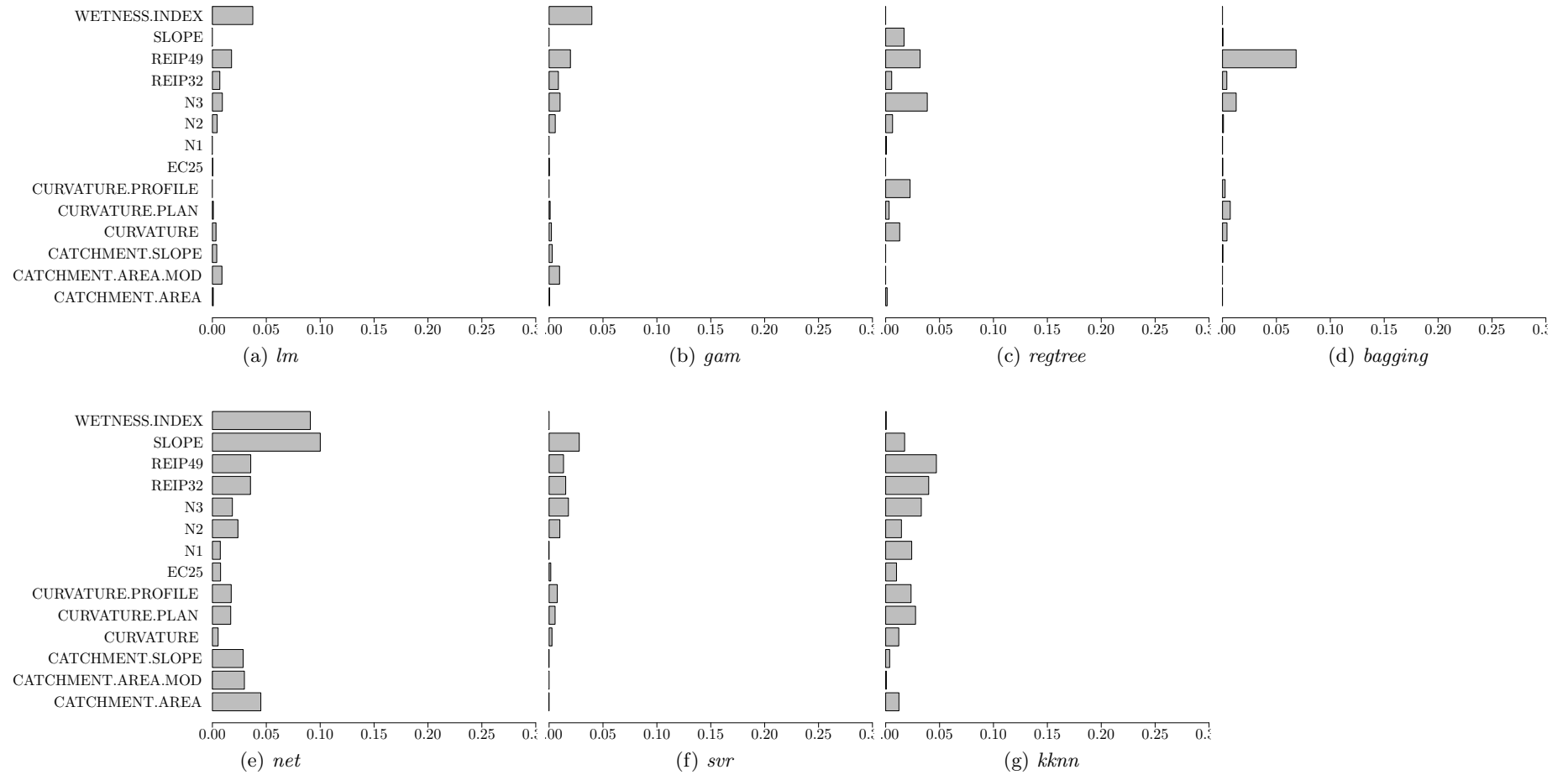


Figure B.42: F611, strategy "sensor", regression models and spatial variable importance

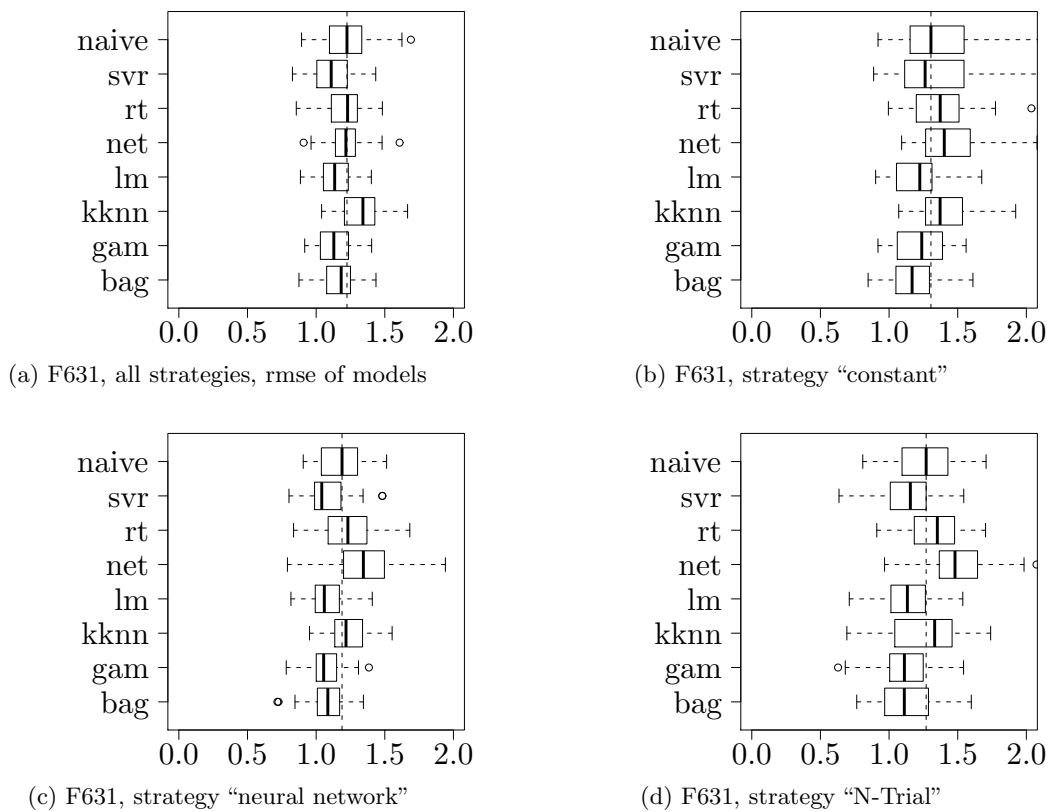


Figure B.43: RMSE for F631 and its subsets (by strategy)

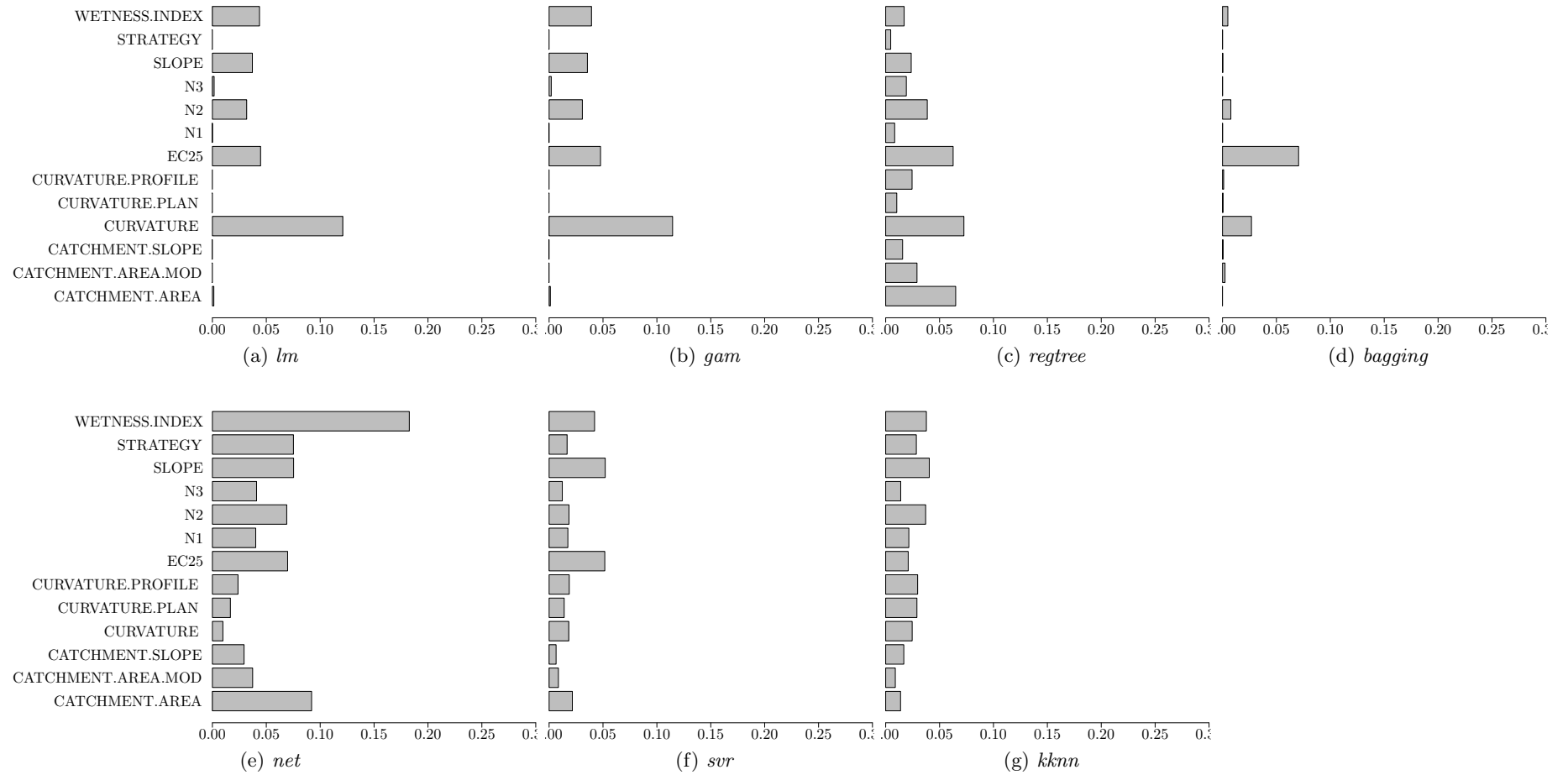


Figure B.44: F631, all strategies combined, regression models and spatial variable importance

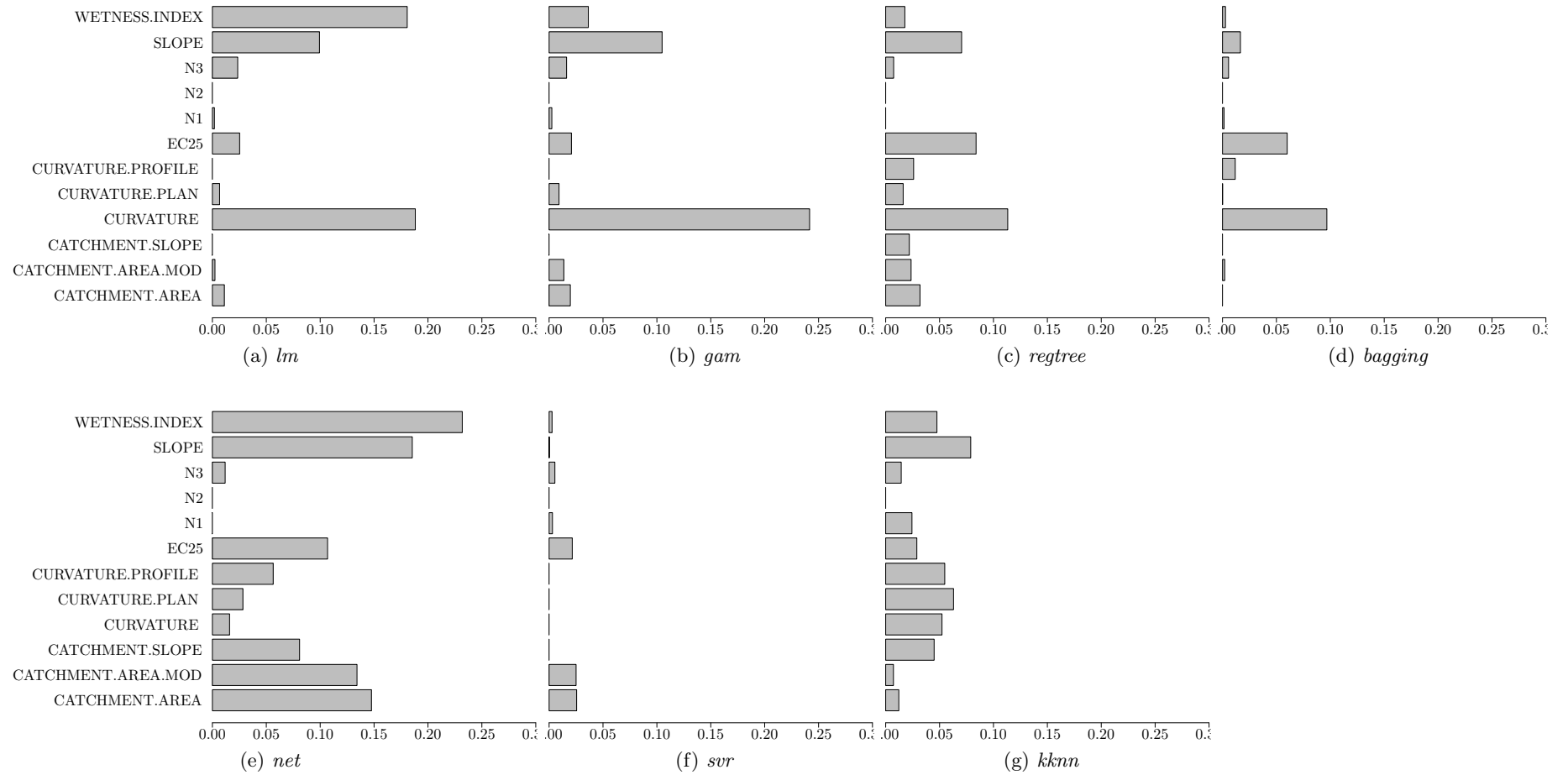


Figure B.45: F631, strategy “constant”, regression models and spatial variable importance

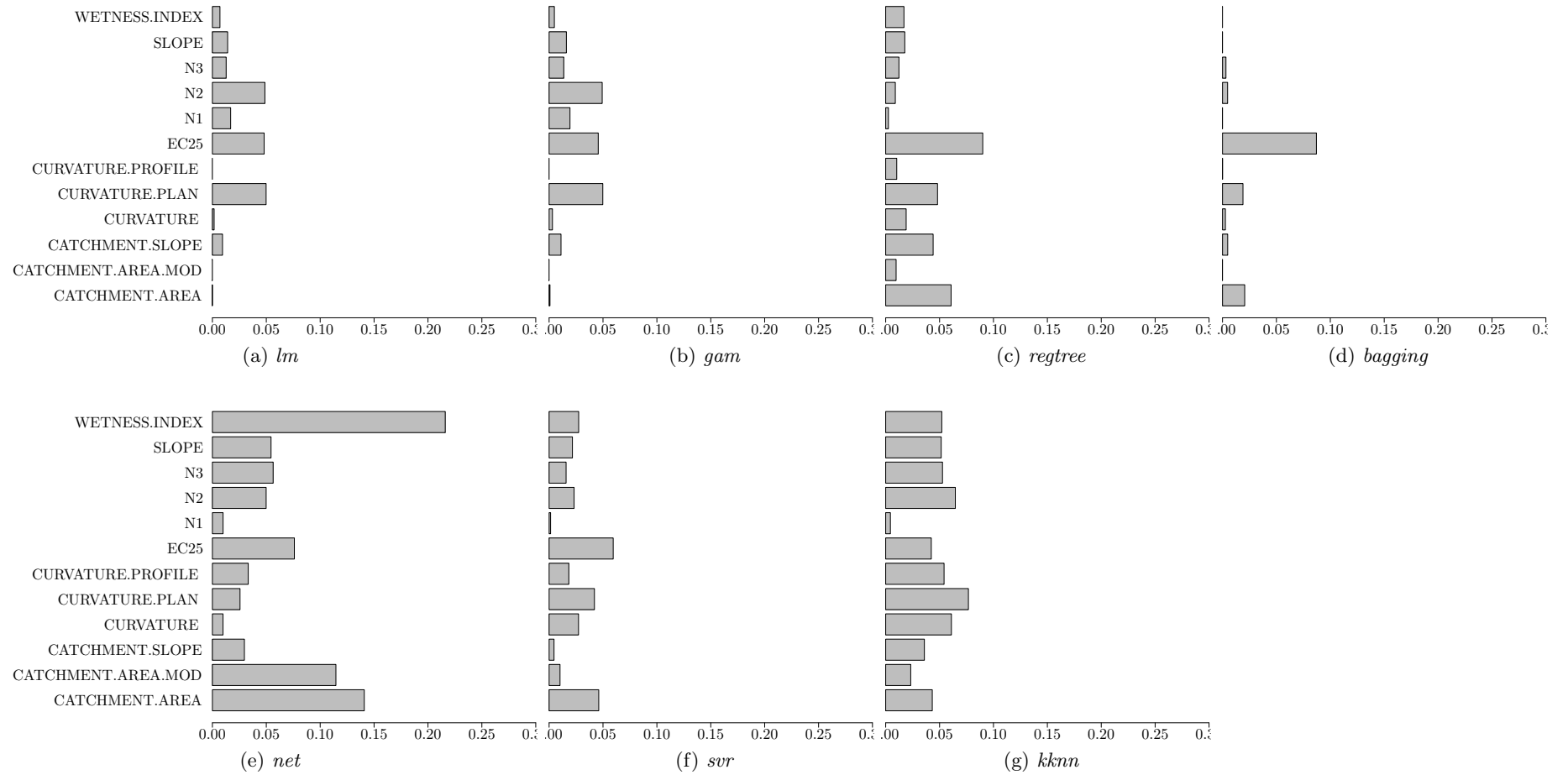


Figure B.46: F631, strategy “neural network”, regression models and spatial variable importance

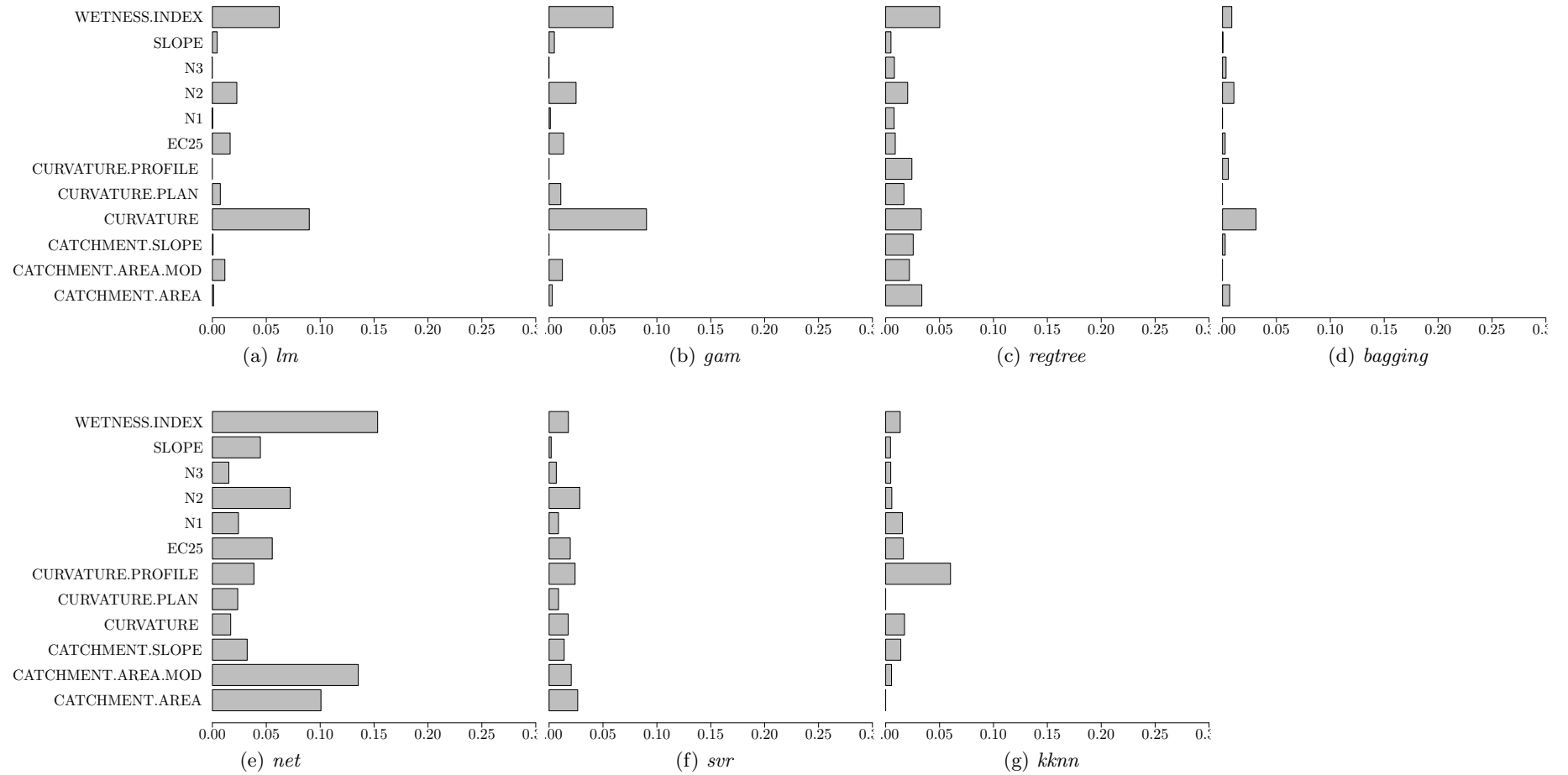


Figure B.47: F631, strategy "N-trial", regression models and spatial variable importance

Appendix C

R package details

This section provides references for the used regression models and their respective R packages/functions, as well as specific settings for those packages.

model	R function	R package	settings
<i>lm</i>	lm	stats	(no change to standard call)
<i>gam</i>	gam	gam	family = gaussian
<i>kknn</i>	kknn	kknn	k = 2, distance = 2, kernel = rectangular
<i>regtree</i>	rpart	rpart	method = anova, minsplit = 20, cp = 0.001, xval = 20
<i>net</i>	nnet	nnet	decay = 10e-3, size = 20, maxit = 1000, linout = TRUE, MaxNWts = 32768
<i>svr</i>	svm	e1071	kernel = radial
<i>bagging</i>	bagging	ipred	nbagg = 250

Table C.1: Regression models, details for the R packages