

## Crop type prediction based on farmers declarations



**Student:** Sibelina de Jong  
**E-mail:** Sibelina22@gmail.com  
**Student number:** s6001963

**Supervisor:**  
Dr. Ir. Sytze de Bruin  
**Responsible professor:**  
Dr. Ir. Ron van Lammeren

**Place of research:** RVO.nl, Assen  
**Date:** August of 2018

## Preface

Before you lies the thesis "Crop type prediction based on farmers declarations". My passion for learning new (software) skills and the fact that I have been working for the last 18 years with agricultural data, made this the perfect subject for a thesis. I have really enjoyed performing this research. With this research I hope to have aided RVO.nl in its task to implement the CAP policy in a more efficient manner.

Without the help and support of a large number of people, this thesis would not have been completed. I want to acknowledge and thank all of them here. In the first place DLG and RVO.nl, who gave me the opportunity to follow this master program. I am also grateful to RVO.nl for allowing me to perform part of this research during working hours for this helped me to complete the research in time. I want to thank the department ASB for allowing me to use data mining software without the required practical skills that are normally demanded. My special thanks goes to Henk for his unrelenting willingness to help me when I was stuck with SAS EM, and Marjolein who helped me further along when I had questions about modeling techniques. I am also very grateful to Petra who covered my work at the NCG when I was in a tight spot.

Many thanks goes to Sytze, my supervisor, for his willingness to guide me in this research, for sharing his knowledge on data mining and Agriculture, his patience in helping me understand the difference in data mining techniques and encouragement to finish this thesis. Finally, I want to thank my family and friends for all their support, love and understanding.

I am very happy that the thesis has now been completed, but I did the research and writing with a lot of pleasure. I hope you enjoy this topic as much as I do.

Assen, 8<sup>th</sup> August 2018



**"The boss wants me to create a computer algorithm that converts hindsight into foresight."**

CartoonStock.com

Figure 1 Create algorithm that converts hindsight into foresight (Bacall 2012)

## Summary

Every year the Dutch agricultural businesses submit all sorts of agricultural data like crop type and parcel geometry to RVO.nl in accordance with CAP (the Common Agricultural Policy). This annual declaration is an enormous burden for the agricultural entrepreneur. To lessen this burden, RVO.nl searched for manners in which the declaration can be simplified. The prediction of crop types cultivated in the next year based on historical declaration information was one of the efforts.

Crop type is very detailed information, in 2017 364 individual crop types could be declared in the Netherlands. Former research performed by RVO.nl in 2016 tried to predict crop types by using information about crop rotation schemas which was extracted out of historical declaration information, however, only very few rotation schemas were found. A reason for finding very few rotation schemas was the change in crop type codification during the research period. Therefore this research focused on identifying the length of crop rotation schemas in the declaration information and assessed whether this information improved crop type prediction for the declaration year of 2017.

All the declared crop type information was gathered for the research period of 2008 till 2017 and combined with additional variable data: farm type, soil type, groundwater level, climate data, user identifier and organic farming. Sixty crop types were adjusted for the change in crop type codification, corresponding crop type codes from 2017 were used instead. Furthermore, a study area with a variety of crop rotation schemas was selected. Then search patterns were used to find repetitive patterns in the crop sequences derived from the historical declaration information in order to identify the length of crop rotation schemas. This information was stored as a new variable for predicting crop types.

All the input variables contained nominal values. As a target variable was present (crop types of 2016), supervised modeling was applied. In order to predict crop types in the next year (crop types of 2017), several classification trees and neural network models were built. By fine-tuning the property settings, models were searched for that produced the lowest misclassification rates.

For 29.622 of the 185.108 sub parcels a crop rotation schema was identified. Multi-way split classification trees produced the lowest misclassification rates. The prediction accuracy was highest for sub parcels where crop rotation schemas were identified (45,8%). The prediction accuracy for sub parcels where no crop rotation schema was found was much lower (31,3%). The average prediction accuracy of the declared parcels was 42,7%. The length of crop rotation schemas was relatively important in predicting crop types. It slightly improved the number of correctly classified sub parcels with 0,9 percentage point, and was used most of all variables in the splitting process during modeling. Class confusion was present in the prediction results. Crop types that frequently formed a crop rotation schema, were often mixed up with each other.

It was found that the aggregation of crop types would lead to a higher prediction accuracy. Furthermore, the use of class weights or priors in the modeling process might increase the prediction accuracy of minority classes. Additionally it was assumed that the random forest modeling technique produces a higher prediction accuracy than a classification tree for it is less prone to overfit the training data. It was also assumed that stratifying the research area might increase the prediction accuracy based on the presumption that similar crop types grow in a small area. An area with similar crop types is assumed to be easier to predict.

The main conclusion is that the length of crop rotation schemas can be identified from historical declaration data and that the variable containing the length of crop rotation schemas was relatively important in the modeling process. By using this variable in the prediction process, the prediction accuracy was improved with 0,9 percentage point. Therefore it is concluded that this variable aids in predicting crop types in the next year. Suggestions are given to improve the current prediction accuracy of 42,7%.

## Content

Summary .....	3
Every year the Dutch.....	3
Abbreviations .....	7
1 Introduction of the prediction of crop types .....	9
1.1 Problem description .....	9
1.2 Related work .....	10
1.3 Knowledge gap .....	11
1.4 Research identification.....	11
1.5 Research questions .....	11
1.6 Innovation aimed at.....	12
1.7 Scope of the project .....	12
1.8 Area of interest.....	12
1.9 Structure of the report .....	13
2 Methodological background and important variables for predicting crop types .....	14
2.1 Identifying the length of crop rotations .....	14
2.1.1 Search patterns .....	14
2.1.2 Semi-supervised modeling.....	15
2.1.3 Missing crop type codes and inconsistent crop sequence.....	15
2.1.4 Summary of identifying the length of crop rotations.....	16
2.2 Important variables for predicting crop types .....	16
2.2.1 Variable: crop type (declaration information) .....	16
2.2.2 Variable: farm area dynamics .....	16
2.2.3 Variable: farm type .....	16
2.2.4 Variable: organic agriculture .....	17
2.2.5 Variable: altitude data .....	17
2.2.6 Variable: soil type.....	17
2.2.7 Variable: groundwater level.....	17
2.2.8 Variable: climate data.....	17
2.2.9 Variables owner and usage.....	18
2.2.10 Summary of important variables .....	19
2.3 Potentially suitable modeling techniques .....	19
2.3.1 Logistic regression .....	20
2.3.2 Classification tree modeling .....	20
2.3.3 Neural network modeling .....	20
2.3.4 Bayesian network.....	20
2.3.5 Random forest.....	20
2.3.6 Linear and kernel SVM.....	21
2.3.7 Choosing suitable modeling techniques .....	22
3 Methods.....	24
3.1 Variable: Length of crop rotation schemas .....	24
3.2 Application of the prediction methods .....	27
3.2.1 SEMMA - sample .....	29

3.2.2	SEMMA - explore .....	29
3.2.3	SEMMA – modify.....	30
3.2.4	SEMMA – model: Classification tree.....	31
3.2.5	SEMMA – model: Neural network.....	35
3.2.6	SEMMA - assess.....	37
3.3	Selection of the sub parcel as input for the modeling process .....	38
3.4	Defining a study area .....	39
4	Results .....	42
4.1	Identification of crop rotation schemas .....	42
4.2	Selection of the prediction method.....	43
4.2.1	Classification trees .....	43
4.2.2	Neural networks .....	45
4.2.3	Indication best predictive models .....	45
4.3	Assessment of the prediction accuracy .....	46
4.3.1	Prediction accuracy of the classification tree models .....	46
4.3.2	Prediction accuracy according to label for crop rotation schemas.....	47
4.3.3	Contingency tables.....	48
4.3.4	Comparison of prediction accuracy of the former research performed in 2016 and the current research .....	51
4.4	Variable importance .....	54
5	Discussion.....	57
5.1	Method for identifying the length of crop rotation schemas .....	57
5.2	Selection of the prediction method.....	57
5.3	Prediction accuracy of crop types .....	59
5.3.1	Contingency tables.....	59
5.3.2	Comparison of current prediction results with research performed in 2016 .....	60
5.4	Variable importance .....	60
6	Conclusion and recommendations.....	62
6.1	RQ I: What methods can be used to identify the length of crop rotation schemas and predict crop types?.....	62
6.2	RQ II: Which variables available at RVO can be used for predicting crop types? .....	62
6.3	RQ III: To what extent are variables available in historical data sets at RVO capable of predicting crop types? .....	62
6.4	RQ IV: To what extent does information on the length of crop rotation improve predicting crop types in the reference year? .....	63
6.5	Main Conclusion.....	63
6.6	Recommendations .....	63
	References.....	66
	Appendix A – Selections to create label crop rotation schema.....	71
	Appendix B – Example of selections to replace missing data crop types.....	73
	Appendix C - Example of a SAS Enterprise Miner diagram .....	74
	Appendix D – Class variable summary statistics .....	75
	Appendix E - Tables with misclassification rates for classification tree models .....	76
	Appendix F - Classification tree properties description.....	78

Appendix G – Splitting procedure for multi-way split classification trees.....	84
Appendix H - Autoneural network properties description.....	85
Appendix I - Variable importance description .....	91
Appendix J – Select sub parcels from combined variable data to use as input data for modeling (ArcGIS model) .....	93
Appendix K - Arable area of municipalities present in study area (CBS 2016) .....	94
Appendix L – Crop rotation schemas in the unlabeled input data set .....	95
Appendix M - Analyses of sub parcels .....	96
Appendix N – Double claims .....	98
Appendix O – Declared parcels in 2016 and 2017 .....	99

## Abbreviations

Abbreviation	Description	Meaning
ADJ	Adjustment	This abbreviation was used in the name of the columns containing the adjusted crop codes for the change in crop type codification (input data set for modeling crop types)
AOI	Area of Interest	The extent of the area used in this research
BN	Bayesian network	A supervised modeling technique
CAP	Common Agricultural Policy	European Union Policy concerning Agriculture
CBS	Centraal Bureau voor de Statistiek	
CrC	Crop codes	The number that represents a specific crop type, used in the CAP declarations
CRISP-DM	Cross-industry standard process for data mining	A data mining methodology
CRS	Crop Rotation Schemas	
DICTU	Dienst ICT Uitvoering	ICT supplier of the Ministry of Economic Affairs and Climate Policy
DSL	Dienstsleutel	A unique number to represent the user of a parcel, used in the CAP declaration and RVO.nl's correspondence with agricultural businesses
ELA	Enterprise License Agreement	A list of standardized software to be used in the work processes of the Ministry and affiliated organizations
ESRI	Environmental Systems Research Institute	A GIS software supplier
EU	European Union	
GIS	Geographical Information System	Software tooling used to analyze, adapt and display geographical information
GO	Gecombineerde Opgave	The declaration of agricultural data concerning CAP
GWC	Gewascode	Crop type code
HPSVM	High Performance Support Vector Machines	A supervised modeling technique
ICT	Information- and Communication technology	
KDD	Knowledge Discovery and Data Mining	A data mining methodology
KNMI	Koninklijk Nederlands Meteorologisch Instituut	The Dutch Weather Institute
LPIS	Land Parcel Information System	A digital parcel information system in which all information concerning the Dutch agricultural area is stored
LUS	Land-Use Successions	Crop sequences of crop rotation schemas
NAP	Normaal Amsterdams Peil	A Dutch measure for the water level
NRULES	Number of Rules	The number of times a variable was used to split node information (groups and subgroups of input data)
NSO	Nederlandse Standaardopbrengst	The Dutch Standard Output. The NSO is a standardized yield per ha or yield per animal that is achieved on an annual basis for a specific crop category or a specific animal category
PDOK	Publieke Dienstverlening op de Kaart	Dutch geoportal

Abbreviation	Description	Meaning
PMF	Pilot Monitoring Farmland	An initiative at RVO.nl to investigate farmland by applying remote sensing and data mining historical declaration data
REP	Replacement	This abbreviation was used in the name of the columns for which the missing values were replaced
ROTOR	ROTations in ORganic farming systems	A software tool that models crop rotations for organic farms
RPG	Registre Parcellaire Graphique	French translation for LPIS
RQ	Research Question	
RVO.nl	Rijksdienst Voor Ondernemend Nederland punt nl	Executive Agency of the Ministry of Economic Affairs and Climate Policy
SAS EM	Statistical Analysis Systems, Enterprise Miner	A Business Information software supplier
SEMMA	Abbreviation for 5 steps: Sample, Explore, Modify, Model and Assess	SAS EM's approach to create predictive models in 5 steps
SVM	Support Vector Machines	A supervised modeling technique
TBM	Teeltbeschermingsmaatregelen Zetmeelaardappelen	TBM is a foundation for protecting seed potatoes cultivated in the northeast of the Netherlands



# 1 Introduction of the prediction of crop types

This chapter starts with a problem description and stating the aim of this research. It discusses similar research and how this study's approach is different. Furthermore, the innovation and the scope of this project are addressed, the area of interest is displayed, and this chapter concludes with the structure of this report.

## 1.1 Problem description

The European Union (EU) is an important partnership of 28 European Member States at present. These member states work together in several general areas among which are agriculture, fisheries, consumer rights and the environment. Agriculture is an important issue in the EU: nearly 40% of the European Union budget is spent on it (EuropaNu 2017). Back in 1957 the Common Agricultural Policy (CAP) of the European Union was created. This policy was implemented to ensure that there is enough food in Europe and that agricultural products can be sold at and bought for reasonable prices.

Member states define their own agricultural policy within the limits of the CAP. In the Netherlands, RVO.nl (Rijksdienst voor Ondernemend Nederland) is responsible for the implementation of the policy by order of the Ministry of Economic Affairs and Climate Policy (RVO.nl 2017b). In accordance with the agricultural policy, farmers must take into account various interests, such as food safety, the preservation of the countryside, the environment, the living conditions of animals, and fair trade with countries outside the EU.

In order to implement this policy, it is mandatory for agricultural businesses to submit agricultural data. This is a combined submission of all sorts of data concerning the aforementioned various interests, also called GO (combined declaration), and is submitted to RVO.nl. Every year before the 15<sup>th</sup> of May agricultural businesses must present information about the usage of land, crop cultivation and specific data concerning the type of agricultural business. In this manner farmers ask for subsidies concerning the CAP (RVO.nl 2017a).

This declaration is a considerable administrative burden which repeats itself annually. To lessen this burden, the Dutch Government strives to simplify procedures. The Government only aims to ask data it does not possess or cannot derive from other available sources. However, to support agricultural entrepreneurs in diminishing the endeavor of submitting furthermore, RVO.nl is researching other manners in which the process of the declaration can be simplified. One of these efforts is predicting crop types.

This research started in 2016 and was divided into four subprojects (RVO.nl 2016):

1. Predicting owner (name) and type of ownership/usage
2. Predicting crop type based on historical data
3. Predicting crop type based on satellite data
4. Other useful information for predicting crop types

The results of the investigation were encouraging. Predicting crop types based on data mining information derived from six years of farmers declarations (also called LPIS data) and other additional variable information resulted in 69,1% correctly predicted parcels. In total 11 different crop types were present in the modeling data. Declaration data from 2008 till 2014 were used to predict crop types that were declared in 2015. Predicting crop types based on satellite data was also encouraging, although there were big differences between specific crops types. On average 64,3% of all the investigated plots were predicted correctly. Here in total 14 different crop types were present in the modeling data. Satellite images of March 2015 till July 2015 were used to predict the crop types that were declared in 2015.

Due to the limited time (only three months) and limited resources for this project, the crop types used in the prediction were aggregated (11 different crop types). The total number of actual crop types that

could be declared in 2015 was 304. Also data mining of owner (name) and type of ownership/usage could not be investigated properly.

Conclusion of the research was that there was room for improvement in predicting crop types. The research results mentioned the lack of parcels with a detectable crop rotation schema. Crop rotation schemas represent repetitive patterns present in the sequence of crop types that are cultivated on the same parcel each year. During the research the length of crop rotation schemas (i.e. two years, three years, etc.) was added as an extra variable to help in predicting future crop types (RVO.nl 2016). In order to create this variable, selections were applied to search for patterns in the sequence of declared crop types. When a repetitive pattern was found, the length of the rotation schema was deduced from the time difference between declaring the same crop type. Several additional variables were created to indicate two- and three-year rotation schemas and whether it concerned a crop rotation involving potatoes, maize, sugar beet or wheat. Unfortunately, the influence of these variables could not be investigated properly for very few rotation schemas were found: 1,4% of the declared parcels contained two-year rotation schemas and three-year rotation schemas were found on only 0,4% of the declared parcels.

However, crop rotations are a very common practice in Agriculture. Crop rotations have two main functions. They are used to control diseases, plagues and growth of weeds, and they improve soil fertility (Wijnands 2000). There are several different crop rotation schemas possible: 1:2 (two-year rotation schema), 1:3 (three-year rotation schema), 1:4 (four-year rotation schema) and even 1:6 (six-year rotation schema) (Wijnands 2000; IRS 2017). The reason mentioned for finding so few crop rotation schemas was the change in crop type codification in the data mining period (RVO.nl 2016). This means that codes used to declare which crop is cultivated, changed during the researched timeframe. For instance, the crop type code 672 that represented vegetables grown in open ground in 2014 was split up in 95 new crop type codes all representing individual vegetables in 2015. By correcting older crop types to fit the current list of crop types, more crop rotation schemas may be detected and predicting crop types may be improved. Other recommendations were to include owner (name) and type of ownership/usage in the data mining process, to use more data resources to define the data mining process, and to scrutinize the data mining process itself. As crop rotation schemas are an important farming strategy and the length of crop rotation schemas was used in the former research, in this research also a new variable for the length of crop rotation schemas will be created in order to investigate if this variable information helps in predicting crop types. Furthermore, in this research the actual crop types will be predicted instead of aggregated crop types and the research period is extended from 2008 till 2017.

## **1.2 Related work**

Research shows that information on crop rotations has been used in modelling agricultural dynamics. For example, Janssen (1994) applied available knowledge of crop rotations to improve his methodology for updating terrain object data by use of satellite imagery. This indicates that knowledge of crop rotation could also improve prediction results of crop types. This was confirmed by the research of Osman et al. (2015) that suggests to model crop rotations in order to predict crop types that will be cultivated before the agricultural season started. Study reveals that a lot of research has been done on modelling crop rotations.

CarrotAge (Le Ber et al. 2006), CropRota (Schönhart et al. 2011) and ROTOR (Bachinger et al. 2007) are such models. CarrotAge is a model that makes use of historical data on crop types (LPIS data) for which Markov models are used to extract rotation schemas from. CropRota uses agronomic criteria and crop cultivation data to create crop sequences. ROTOR is a model specifically developed for producing crop rotations for organic agricultural businesses. It differs from other models by incorporating information on crop rotations and its sensitivity to soil quality (Osman et al. 2015).

As indicated above, the use of LPIS data to determine crop rotations is not new. Leteinturier et al. (2006) used LPIS to create a crop sequence indicator by combining several successive years of LPIS

data in a GIS. The land data used in CarrotAge and CropRota was also LPIS data (Le Ber et al. 2006; Schönhart et al. 2011) and Steinmann et al. (2013) used LPIS data to identify crop sequence and crop rotations in the northern region of Germany.

In France, the RPG explorer was developed, also to investigate the agricultural landscape dynamics using successive years of LPIS data. The land cover information was aggregated however, 28 crop groups have been fabricated and the information was addressed at farmer block level (not at field level). A farmer block can contain more than one land cover. Expert knowledge was inserted into the model to identify types of farms and to identify crop rotations. The CropRota model was added to the RPG explorer. This tool can identify all the crop sequences in the researched area and identify the main crop rotations among others. However, this tool is not perfect. Using farmers blocks with multiple land covers lacking the precise geometrical boundaries of the crops that were cultivated, raises issues concerning the correct successive order of land cover (Levvasseur et al. 2016).

Xiao et al. (2014) and Osman et al. (2015) suggested to use Markov models for modelling crop rotations in order to identify crop rotations for a large area. Xiao et al. (2014) proposed data mining based on a second order hidden Markov model (HMM2) in combination with unsupervised clustering techniques. With HMM2, time events can be modelled more accurate because the crop type cultivated in year T is dependent on the crop type grown in the previous year (T-1) and the crop type grown in the year before that (T-2). According to Osman et al. (2015) Markovian properties are often used in data mining when a specific order of data is important in the modeling process. Mostly Bayesian Networks (BN) are used as Markovian models. Markovian models and Bayesian networks are both graphs. However, Bayesian network graphs are directed and non-cyclic which is different from Markovian networks. These networks are non-directional and possibly cyclic. Both Markovian networks and Bayesian networks are suitable models to predict crop types.

### **1.3 Knowledge gap**

In contrast to the RPG explorer that focusses on farmers blocks (Levvasseur et al. 2016), this research aims to identify rotation schemas at individual field level. Another difference is that in this research actual declared crop types were used to identify crop rotation schemas and predict crop types. This is different from the approach of Osman et al. (2015) and (Levvasseur et al. 2016) where aggregated crop types were used to identify crop rotations and then predict crop types. However, all the mentioned models searched for crop rotation information to either calculate crop yield, or search for the most optimal crop rotation sequence that yielded the most, or targets a specific agricultural group (i.e. organic farms), and all of them used the actual crop sequences of rotation schemas (i.e. wheat – sugar beet - potato). This research differed from that approach and aimed to identify the length of crop rotations. This information (i.e. two-year rotation schema, three-year rotation schema, etc.) was then used as an input variable for predicting crop types to investigate if this information improved the prediction modelling of crop types.

### **1.4 Research identification**

This research aims to identify the length of pertinent crop rotation based on historical data of farmers declarations and assess whether this length improves the prediction of the crop type cultivated in the next year.

### **1.5 Research questions**

The following research questions will be answered in order to determine the main objective of this research:

- (I) What methods can be used to identify the length of crop rotation schemas and predict crop types?
- (II) Which variables available at RVO can be used for predicting crop types?

- (III) To what extent are variables available in historical data sets at RVO capable of predicting crop types?
- (IV) To what extent does information on the length of crop rotation improve predicting crop types in the reference year?

### **1.6 Innovation aimed at**

The need for predicting crop types was never an issue for RVO.nl until 2010. Through public announcement the Dutch Government announced that it was important to simplify, deregulate and reduce the administrative burden in order to stimulate (agricultural) entrepreneurship, and to reduce the implementation and control burden for the Government itself (Bleker 2010). For RVO.nl it is new to predict crop types and with this research it gains knowledge on the feasibility of predicting (this) information. In relation to the project done in 2016, the innovation of predicting crop types is aimed at expanding the historical data set with more data to train the model and identifying more rotation schemas by adjusting the crop types for the change in crop type codification in order to create a better predictive model. Also the actual declared crop types are used for the prediction, not aggregated crop types.

At present a pilot is executed at RVO where farmland is monitored by use of remote sensing and satellite data, Pilot Monitoring Farmland (PMF). This research is also part of the pilot for one of the goals in this pilot is to determine as early as possible what crop type is cultivated in the season to come. The innovation is aimed at combining the results of predicting crop types based on historical declaration data with the results of the remote sensing process in order to determine which crop type is cultivated. By combining this information it is assumed that a more informed and accurate prediction result is produced.

### **1.7 Scope of the project**

The aim of this project is to improve the prediction method used in RVO.nl's research of 2016 by focusing on the identification of more crop rotation systems. In scope is adding the length of crop rotation schemas as a variable to the input data set for modeling crop types, however, out of scope is the research into the accuracy of the prediction results if the crop sequences of crop rotation schemas would have been used in creating models like the related work suggests.

Another aim of this research is investigating suitable modeling techniques for predicting crop types. In scope is which alternative modeling techniques there are to predict crop types, however out of scope are modeling techniques that are not available within the current version of RVO.nl's data mining software.

### **1.8 Area of interest**

As of 2005, every European Member State is required to have a digital parcel information system in which information about the Dutch agricultural area is stored, also called LPIS (Land Parcel Information System). This is the area for which farmers can declare subsidy. The quality and the accuracy of this geographical information is updated on a three-year basis. All parcels are checked once every three years and updated according to the most recent (summer) areal image. As mentioned, LPIS is used by farmers to submit agricultural data however, RVO.nl also uses LPIS to verify the validity of these declarations. LPIS contains a little over 500.000 objects and the area is about 1.900.000 ha large. Compared to the Dutch land surface 54% is agricultural area. About 40% of the LPIS area is arable farmland where crops are cultivated. This is the area where crop rotation schemas are expected to be found, see Figure 2.

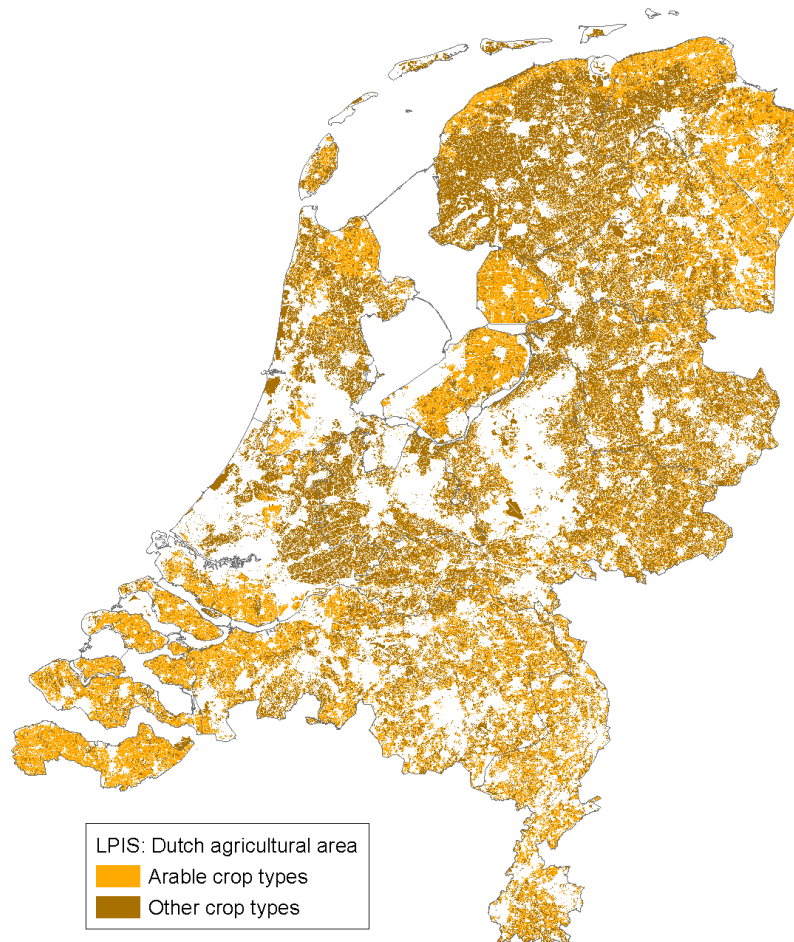


Figure 2 Agricultural area 2017

During the research period of this study, the number of crop type codes that could be declared changed. Table 1 provides an overview of the number of crop type codes that could be declared in the Netherlands for each year of the research period.

Table 1 number of crop type codes that can be declared per research year

	Research period									
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Number of crop type codes	103	98	114	96	89	90	90	304	361	364

## 1.9 Structure of the report

In chapter 2 a methodological background is provided. Here the method to identify the length of rotation schemas is described, which variables are important in the modeling process and a list of suitable modeling techniques to predict crop types is provided. Chapter 3 describes the applied methods. In Chapter 4, results are provided for the identification of the length of crop rotation schemas, misclassification rates of the created models, accuracy results of predictions, contingency tables, and comparing research results of this research and the research performed in 2016. The chapter concludes with the results of variable importance. In chapter 5, the results are discussed while in chapter 6 all research questions are answered. Finally, a conclusion is drawn concerning the main question of the research and recommendations are provided for further research. At the end of this report a reference list is provided and additional information can be found in appendices.

## 2 Methodological background and important variables for predicting crop types

This methodological background addresses methods for identifying the length of crop rotation schemas, variables influencing the prediction of crop types and modeling techniques that are suitable for predicting crop type.

### 2.1 Identifying the length of crop rotations

Research has indicated that information on crop rotation schemas can aid in predicting crop types. Janssen (1994) applied available knowledge of crop rotations to improve the method to update terrain object data by use of satellite imagery and Osman et al. (2015) suggested modelling crop rotations in order to predict crop types prior to the actual agricultural season. Research performed in this area of expertise focused mainly on identifying the actual crop sequence (i.e. maize-sugar beet-potato) and not on the length of the crop rotation (i.e. 2-year crop rotation, 3-year crop rotation, etc.).

In the previous research of 2016 (RVO.nl 2016), the length of crop rotation schemas (i.e. 2 years, 3 years, etc.) was used in predicting future crop types. Unfortunately, very few crop rotation schemas were found and the influence of the length of crop rotation schemas could not be investigated in detail. A reason for not finding crop rotation schemas was the greening measure imposed by the EU. The Greening measure was implemented to preserve the environment, increase the biodiversity of the agricultural area and to reduce climate change. One of the practices of greening was diversifying crops. Farmers with a substantial area of arable land are required to grow 2 or 3 different crop types per year in order to get payment based on greening (European Commission 2017). Therefore more distinct crop types codes were necessary that could be used in the declaration process. Consequently the number of declarable crop types codes increased from 90 crop codes in 2014 to 304 in 2015 and 361 in 2016. Therefore, crop types need to be corrected for the change in crop type codification before crop rotation schemas can be found. The approach for this correction is described in Paragraph 3.1.

#### 2.1.1 Search patterns

When the crop types have been adjusted for the change in crop type codification, search patterns can determine the length of a crop rotation. Therefore, a sequence of crop types must first be created by combining all the declared crop types. Such sequences can be searched for crop types that are declared frequently and checked whether the time between the cultivating of the same crop types is the same. The time between the cultivation of the same crop types represents the length of that crop rotation. Because the length of the crop rotation is searched for in this research, it is not important which crop types form the rotation schema, only that there is a repetitive pattern that remains the same. Mari et al. (2010) and Xiao et al. (2014) applied search patterns in their research. They also use search patterns for finding LUS's (Land-Use Successions) in combination with a mayor land-cover category, see search pattern in Table 2.

Table 2 Search pattern used for finding crop rotations (Mari et al. 2010)

Search pattern for extracting all 4-year long LUS involving one of the main land-use categories			
1st Year	2nd Year	3rd Year	4th Year
X	?	?	?
?	X	?	?
?	?	X	?
?	?	?	X

X, the current main land-use category; ?, any land-use category

### 2.1.2 Semi-supervised modeling

In case very few crop rotations are found, semi-supervised modeling is optional to find more crop rotation schemas. With semi-supervised modeling, labeled data (the length of crop rotations is known) is used to find crop rotations in the unlabeled data (the length of crop rotations is unknown) (Han et al. 2011). When semi-supervised modeling is applied, supervised modeling techniques and clustering are combined (Hall et al. 2014). Self-training and co-training are considered semi-supervised modeling approaches (Han et al. 2011). Supervised modeling techniques use labeled data to train models. The target values (length of crop rotation schemas) are compared with the prediction results of the created model. When the predicted value is not the same as the target value of the model, the model can be adjusted in order to produce a more accurate prediction result. Supervised modeling techniques are described in Paragraph 2.3.

#### *Self-training*

With self-training a model will be created by using the labeled data. To create the model a modeling technique must be chosen, i.e. one of the appropriate supervised modeling techniques. This model will then be used to label the unlabeled data. The record with the most confident label prediction is added to the labeled data and then a new model will be created based on the new set of labeled data. This method is easy to understand however errors may be reinforced (Han et al. 2011).

#### *Co-training*

With co-training two (or more) models are created based on non-overlapping labeled feature sets. Also to create the model a modeling technique must be chosen, i.e. one of the appropriate supervised modeling techniques. Again, the record with the most confident label prediction is added to the labeled data, only not to the set of labeled data that was used to predict the label. It is added to the labeled data of the other (or another) model. Then new models will be created. In this manner each model teaches the other model, resulting in models that are less sensitive to errors although it may be difficult to split the features into independent non-overlapping sets (Han et al. 2011).

Clusters of labeled data are formed based on the length of the crop rotation schemas. Crop sequences of these clustered data can be used to find similar crop sequences in the unlabeled data. The length for crop rotation schemas of the clustered data can then be added to unlabeled data of which the crop sequences resemble the crop sequences in the clustered data.

Semi-supervised modeling should be applied when only a few crop rotation schemas are found. Unfortunately, no research was found on how many detectable, repetitive crop rotation schemas exist in the Netherlands. It can be conceived that every (arable) farm has a crop rotation schema, for crop rotation schema are used to control diseases, plagues, the growth of weeds, and to improve soil fertility (Wijnands 2000). However, rotation schemas can change over time owing to changes of user, economic reasons, pest problems or weather conditions (SARE 2012). Therefore, a decision to apply semi-supervised modeling can only be based on the number of labels found relative to the total population and the area that is involved. In the present study, semi-supervised modeling would be applied if for less than 10% of the sub parcels a crop rotation schema were found.

### 2.1.3 Missing crop type codes and inconsistent crop sequence

Another reason for failing to find crop rotations are missing crop type codes in the crop sequences. When a parcel was not declared in every year of the data mining period, gaps will appear in the crop sequence when this declaration information of the data mining period is combined. Reasons for gaps in the crop sequences are multiple:

- a farmer does not use a parcel for agricultural purposes anymore,
- a building or barn has been built on a (former) agricultural plot,
- infrastructure changes (i.e. a road was constructed),
- silage is stored in the same spot of a parcel every year
- change in agricultural policy excludes types of land from the agricultural area,
- change in agricultural policy adds parcels to the agricultural area.

These gaps are missing values and influence the identification of crop rotations.

Another reason for failing to find the length of crop rotations might be unclear crop type sequences. The most straightforward and clear pattern is one where no values are missing and a consistent repetitive pattern is found: for instance A-B-C-A-B-C. However, values are missing and patterns are not always clear. In the pattern A-B-C-D-A-X-C-Y, crop type A and C are consistent and make this a four-year crop rotation schema, while crop type B, D, X and Y are inconsistent and therefore more difficult to predict accurately.

#### **2.1.4 Summary of identifying the length of crop rotations**

Search patterns and semi-supervised modeling are described as methods to identify the length of crop rotation schemas. First, search patterns are used and when necessary, semi-supervised modeling is applied. Although it can be conceived that every arable farm uses crop rotation schema, repetitive patterns may remain undetected by pattern searching. Alternatively, semi-supervised learning may be used. An assumption was made to apply semi-supervised modeling if for less than 10% of the parcels a crop rotation schema was found.

## **2.2 Important variables for predicting crop types**

The length of a crop rotation schema is not the only variable influencing crop cultivation. Before a model can be created, an inventory of other predictive variables must be made. There are several variables that play a role in crop cultivation and are therefore important for predicting crop types. Levavasseur et al. (2016) and Bouty et al. (2015) mentioned the influence of farm territory dynamics on the change of crop sequences. Xiao et al. (2014) wrote about the link between farm types and change in crop sequence patterns. Osman et al. (2015) mentioned the inclusion of digital elevation models, climatic data and soil type maps and Wijnands (2000) argued that different rules apply for crop sequences on organic farms which leads to longer rotation schemas compared to non-organic farms. All these variables are linked to identifying crop rotations and predicting crop types.

### **2.2.1 Variable: crop type (declaration information)**

Crop rotation schemas are derived from historical declaration records. Every year it is mandatory for farmers to submit agricultural data. Among the declaration information is the geometry of the parcel and the crop type code that is cultivated that year.

Crop type codes are subject to change. Every year RVO.nl provides a complete list of declarable crop types and the change in crop type codification is compared with the previous year. The implementation of the greening measures caused a huge increase in the number of different crop type codes in 2015 and 2016. However, change in crop type codes occurred more often. From 2009 till 2012 also some of crop type codes changed or new crop type codes were added to the declaration list, however, the difference in added or changed crop type codes was never larger than 18 codes, see Table 1 on Page 13.

### **2.2.2 Variable: farm area dynamics**

The research of Bouty et al. (2015) investigated whether farm area dynamics affected the change in crop sequence. Farm area dynamics is specified here as an increase of area belonging to one agricultural business. The results were inconclusive. Changes in crop sequences were found for both stable farms (no area increase) and for growing farms. This indicates that the change in farm territory does not affect the change in crop sequences and crop rotation schemas.

### **2.2.3 Variable: farm type**

For future studies Xiao et al. (2014) recommend to research the effect of difference in farm types on changes in crop sequences. They mention the economically based EU community typology for agricultural businesses. This commission regulation (EUR-lex 2008) is the guideline for the Dutch farm typology called the NSO typology (Dutch Standard Output) (van Everdingen et al. 2016). The difference with the farm types mentioned in the commission regulation is that the NSO typology only includes farm types that are present in the Netherlands. In the development of the classification, the



European characterization is followed as best as possible. The NSO farm typology is used by RVO.nl to deduct the farm type from the information declared by the farmers.

Farm type is also a variable that is linked to crop rotation schemas and which crop types are cultivated. It is easily conceived that a change in farm type will have an impact on rotation schemas. A logical example is that a dairy farm with grazing livestock has (permanent) grassland while an arable farm in all likelihood has a diverse set of possible crop types placed in a rotation schema.

#### **2.2.4 Variable: organic agriculture**

Furthermore, organic farming was mentioned to take into account as a variable. In the NSO typology no distinction was made for organic farming, however, the information of organic farming is available in the declaration information since 2015. Although this variable does not influence the specific crop type that is cultivated, it is assumed to have an influence on the length of crop rotation schemas. Where non-organic farms have crop rotation schemas of two or three or four-year cycles, the best rotation length for organic farms is at least six years. In this manner diseases and pests and the use of (organic) pesticides are kept to a minimum. Two other factors that are important in creating a crop sequence for organic farms are alternating soil-structure enhancing crop types with soil-structure deteriorating crop types, and the sequence of crops that need more nitrate and crops that need less nitrate (Wijnands 2000). To be able to create a sustainable rotation schema, these three factors need to be addressed.

#### **2.2.5 Variable: altitude data**

Osman et al. (2015) suggested including altitude data when crop rotation schemas are researched. It is easily conceived that farming strategies and crop rotation schemas differ when farming at sea level or farming at high altitude. At high altitude average temperatures are expected to be lower than at sea level, and crop types are expected to be different when both sites are compared. However, height differences in the Netherlands are non-consequential regarding the cultivated crops. The highest point is about 322 meters above NAP (Normaal Amsterdams Peil) while the lowest point in the province Zuid-Holland is almost 7 meters below sea level.

#### **2.2.6 Variable: soil type**

Soil type has also been mentioned as an influential variable for crop types and crop sequences (Osman et al. 2015). Soil texture, physiochemical and biological properties differ among soil types. Dependent on specific requirements, a crop will grow better on particular soil types (Quan et al. 2017). It stands to reason that agricultural businesses consider soil type in connection with possible yield results when choosing what crop to cultivate.

#### **2.2.7 Variable: groundwater level**

Finally, groundwater level is also mentioned as a variable that influences crop cultivation. Crop growth is influenced by how deep the groundwater level is (Wesseling 1978). The deeper the groundwater level is, the less water is retained by the soil, therefore the groundwater level might influence which crop is cultivated. Soil type and rooting depth play a role here, too. Clay harbors more water in its pores than sand and the deeper the roots go, the more water can be drawn from the ground.

#### **2.2.8 Variable: climate data**

Studies have also shown that farming is affected by climate and variations in climate, and that climate is partly responsible for variations in annual crop yields (Wang et al. 2001; Howden et al. 2007; Lobell et al. 2011). Climate change might force a farmer to adapt its farming strategies and search for other crop types that can withstand (large) differences in temperature and water supply (Howden et al. 2007). Therefore, regional differences in climate may affect crop type choice and result in different crop rotation schemas.



Figure 3 Long-term average spring temperature

### Temperature

The Netherlands is a relatively small country. Nevertheless, there are regional differences in average temperature and precipitations (KNMI 2015b). In summer, the temperature in the southeast of the Netherlands is higher than in the northwest. On cold winter days, the coastal areas are warmer than the inland. The Dutch meteorological institute (KNMI) has long-term average temperature data available for every season. The long-term average temperature is important for crops are affected by a change in temperature during the growing season (Grise 2013), see Figure 3 for the springtime average of the KNMI. However not only average temperature differences are important, also extremely low temperatures are important for they impact harvest yields (Kabat et al. 2000). The degree of occurrence of extreme low temperatures might also affect crop type choice.

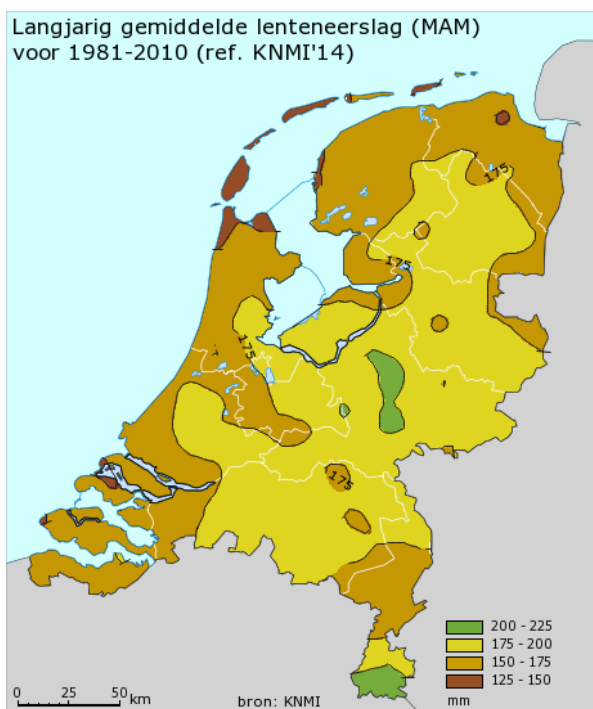


Figure 4 Long-term average spring precipitation

### Precipitation

Studies have indicated that the amount of precipitation affects the choice of crop type and that farmers adapt their crop choice according to climate conditions (Seo et al. 2008; Wang et al. 2010). Particularly precipitation in the spring season is important and can influence crop choice (Grise 2013; Prairie farmer 2017; Southern States 2018). Due to too much precipitation in the spring crops might not be sown in time, or crops do not grow well due to a (too) high groundwater level and/or nutrient leaching. KNMI has long-term average precipitation data available for springtime (KNMI 2015a), see Figure 4. Regional differences in precipitation are at most 100 mm, where on average the coast is drier than the East of the Netherlands

## 2.2.9 Variables owner and usage

Finally, sequences in crop types are most easily predicted if the owner or user of the parcel remains the same. If ownership or usage of a parcel changes, it is likely that other farming strategies are applied and different crops are cultivated. When an owner of a parcel leases it to another user, the latter user is required to declare this parcel. Information about which farmer declares which area, can be retrieved from the variable 'ZAAKID' in the declaration information.

### 2.2.10 Summary of important variables

Summarizing the described research, the following variables will be used in this research:

- Crop type
- Farm type
- Climate data
  - o Long-term average temperature (growing season)
  - o Long-term average precipitation (growing season)
- Soil type
- Organic agriculture
- Owner data/Usage data
- Groundwater level

Most of the referred data are available. Data on crop type, organic agriculture and usage data were retrieved from the declaration information. Farm type data were present for the years 2017, 2016, 2014 and 2013. Farm types for 2015 were derived from this data for farm types are linked to the user's unique identifier which is available in the declaration information. The information on farm types was manufactured by the CBS specifically for RVO.nl. Data sets for soil type and groundwater level were present in RVO's reference database. The information for groundwater level is based on the ranges of groundwater levels, for the groundwater level is dynamic information. This information was available in the feature containing soil types (Dutch soilmap 1:50.000 from 2006). To create the ranges of groundwater levels the highest groundwater levels in winter were used and the lowest groundwater levels in summer (Wageningen University & Research 2018). Although the information is not entirely up-to-date anymore, it was used for there was no more recent information available. The information on climate data was partly available at the KNMI (<http://www.klimaatscenarios.nl/getallen/overzicht.php>). Long-term average data on temperature and precipitation for the growing season (spring) were available in the KNMI's climate database. However, no result was found for (extreme) minimum temperatures during springtime at the site of the KNMI or in the Dutch geo-register (<http://www.nationaalgeoregister.nl/>). Also variable information on farm territory dynamics was not included in this research for it was indicated that the change in farm territory did not cause a change in crop sequences. Altitude information was also not used in this research for the height differences in the Netherlands are not significant enough.

### 2.3 Potentially suitable modeling techniques

An important aspect of predicting crop types is choosing an appropriate modeling technique. In order to decide which modeling techniques are suitable, it is necessary to explore the characteristics of the data that are to be predicted a bit more. The target variable in this project is crop type. It is a nominal variable, which means that although crop type codes are presented in numbers, they cannot be added or multiplied or ranked. There is no order among this information (Agresti 1996). Crop type is a multinomial variable (Řezanková et al. 2009) representing the 364 different crop type codes that could be declared in 2017 in the Netherlands.

In broad terms, modeling techniques that are appropriate for this research can be labeled as supervised modeling, unsupervised modeling or semi-supervised modeling. With supervised modeling techniques models are trained using reference target values. By comparing the prediction results with these target values the model is adjusted to improve the prediction accuracy. With unsupervised modeling no target values are present. The model's algorithm must search for a relation in the input data itself to cluster data. Semi-supervised modeling can be applied when reference target values are only partly present. Then based on the known target values, the model's algorithm searches for similar input values to create more target values (Hall et al. 2014). As the crop types of 2016 are known and can be used to train a model to predict the crop types that were declared in 2017, supervised modeling can be applied. In this paragraph supervised modeling techniques are described that can model multinomial variables.

### **2.3.1 Logistic regression**

Logistic regression searches for a linear function in the relationship between input and target variables. Logistic regression modeling is applied when the target variable contains categorical values. The values of the input variables can be continuous or categorical (Matignon 2007). In this case the appropriate model to predict crop types is a multinomial logistic regression model (Upton 2016) for the target variable is categorical and has multiple values.

Even though logistic regression can handle multinomial values, the range of categorical values of both the target variable and the explanatory variables is very large (i.e. 364 crop type codes in 2017 for the whole of the Netherlands). Predicting crop types with this many values is too complex to solve with logistic regression modeling and will end in failing to create a model.

### **2.3.2 Classification tree modeling**

Classification tree models are used when a nonlinear relationship exists between input and target variables (SAS 2011). This modeling technique splits the target values into smaller, more homogeneous groups. To perform a split, the values of input variables are used in an if-then decision rule to create the subgroups of target values. This splitting process is repeated multiple times where subgroups are split up into even smaller groups in order to increase the homogeneity of target values even more. When the splitting process is visualized, the diagram looks like an upside-down tree with branches (Matignon 2007). Both automatic and interactive training of classification trees is possible (SAS 2011). An advantage of interactively training a classification tree is that the splitting process can be investigated during the creation of the tree.

### **2.3.3 Neural network modeling**

Neural network models also models the nonlinear relationship of input and target variables. Building a neural network model involves two steps. In the first step a specific architecture is chosen. In the second step the model is trained iteratively (SAS 2011) by specifying the activation functions, the target layer error function and the number of hidden layers to determine adequate network weights and create a predictive model (Matignon 2007). Most neural network models are not easy to interpret. However, these models can still be used when prediction results are more important than knowing how the model works (Zhao 2015).

### **2.3.4 Bayesian network**

A Bayesian network is a directed non-cyclic graph where variables are presented as nodes and the relationship (or dependency) of connected nodes as arcs. It computes the probability of outcomes (Osman et al. 2015), which in this research is the likelihood that a crop type changes into another one. According to Osman et al. (2015) Bayesian networks are suitable models to predict crop types.

### **2.3.5 Random forest**

A random forest is a collection of classification or regression trees. For each tree grown in a random forest a different subset of the input data is used, and for each node a different subset of input variables is used to split the input data present in the node. The group of input data which is used in the first split of a tree, is called the root node. The data in the root node are split up into subgroups of data. These groups of data are also called (sub) nodes. One variable is chosen per split from the subset of input variables to divide the input data in such a manner that the values in the end nodes of the classification tree represent individual target values as much as possible (Coursera 2018). The end nodes contain the last groups the input data are split into. When predicting a result based on new input values (also called observations), an output result is created for all the trees grown in the random forest. The output result that is produced the most, is the predicted result for the new input data (StatQuest 2018). Figure 5 shows a representation of a random forest consisting of binary classification trees. The square in the middle of the figure represents the input data from which every time a classification tree is created a different random sample is taken. The question mark represents the variable chosen from a subset of variables to split the input data in the node.

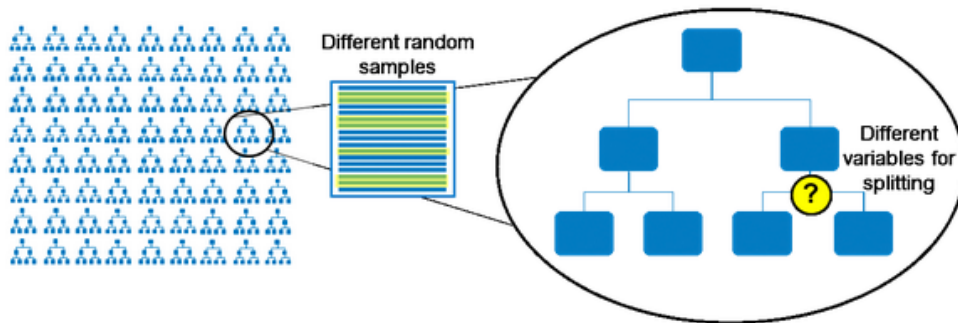


Figure 5 Representation of a random forest (SAS Communities Library 2017).

### 2.3.6 Linear and kernel SVM

SVM (support vector machines) is a binary classification technique. For example, a line is used to split the mapped input data into two classes in a two-dimensional space (Abbey et al. 2017). The line is placed exactly between the values of the two classes to maximize the distinction between the two classes as much as possible. When the classes cannot be separated by a straight line (linear SVM), a kernel trick can be applied. By using a kernel trick, nonlinear functions can be displayed as linear functions in a higher plane (or dimension). Figure 6 shows a representation of a kernel trick. On the left the classes in the data set are displayed in a 2D setting. The classes are mixed up and cannot be separated by a straight line. By mapping the data into a higher plane (changing from a 2D view into a 3D view represented by the data in the cube on the right), the input data can be separated in two distinctive classes. Now a plane is used to split the data set instead of a straight line (Bambrick 2016; SAS 2017).

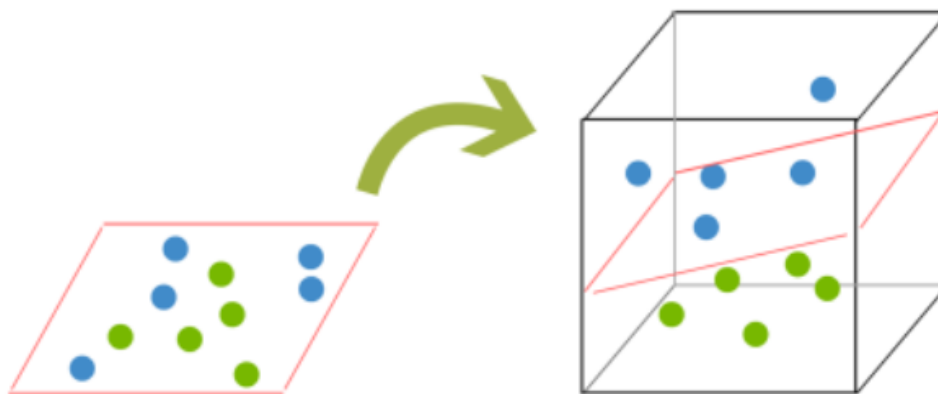


Figure 6 Representation of a kernel trick used to split input data into two distinctive classes (Bambrick 2016)

However, in this research the target variable that needs to be predicted is multinomial. Abbey et al. (2017) mention two approaches in which SVM can perform a multinomial classification: one-versus-one and one-versus-all. In the one-versus-all approach as many SVMs are trained as there are target classes. In the one-versus-one approach SVMs are trained for all the pairs of target classes. When there are  $k$  target classes, the formula for calculating the number of SVMs to be trained is  $k*(k-1)/2$  (Abbey et al. 2017). When the one-versus-all option would be applied for modeling all the crop types declared in 2017 in the Netherlands, 364 binary (dummy) target variables would need to be created, followed by the training of 364 models. Subsequently the results of these models would need to be combined and assessed which of the individual results is the most reliable. In case of the one-versus-one option a staggering number of 66066 model would need to be trained and assessed  $((364*363) / 2)$ .

### 2.3.7 Choosing suitable modeling techniques

The question is which of the described techniques is best suited to predict crop types. There seems to be no particular preference for a specific type of the mentioned modeling techniques. When for instance classification trees are compared with neural networks, advantages of classification trees are that results are faster computed once the model is complete, they are easier to learn how to build, and are very interpretable (the if-then decision rules are straightforward and easy to understand) (Jacobsen et al. 1999; Stack Overflow 2012) Understanding how a modeling technique works and what decisions were made to split the input data when a classification tree is created, may be preferable to a black box modeling technique where the applied steps of a technique are not clear such as a neural network with hidden layers. Knowing how a technique works, provides control of the output result but also confidence in the accuracy of the output result. However the predicted result of a neural networks might prove to be more accurate than that of a classification tree.

Missing values (i.e. empty cells) in the input data set might cause additional problems for several modeling techniques. Logistic regression, neural networks and SVM's do not train models based on input data with missing values. The records with missing values will be discarded, reducing the input data set and increasing the probability of a less accurate predictive model. However, missing data can be replaced. By replacing missing values with appropriate information extracted from the input data set, errors already present in the input data may create further errors due to copying in the replacement process. Classification trees and random forest networks do not have problems with missing data.

Still, to create a more definitive distinction among the modeling techniques which were to be used in this research, the following parameters were used:

- Software version: is the modeling technique available in SAS Enterprise miner version 12.1?
- Performance: can the modeling technique handle many multinomial (target) values?

#### *Software version*

The reason to use SAS Enterprise Miner is mostly an organizational one. RVO.nl is an executive organization of the Ministry of Economic Affairs and Climate Policy. The ICT supplier of the Ministry is DICTU. To uphold a good work quality standard, both parties agreed to a set of standardized software to be used in the work processes of the Ministry and affiliated organizations. The standardized software is documented in an ELA (Enterprise License Agreement) and SAS software is listed in this ELA. An additional advantage is that knowledge is available within the organization for this software package. SAS enterprise miner (SAS EM) creates predictive models based on a collection of specific data related to the topic to be predicted and was used in this research. Although all modeling techniques mentioned here are available in the newest version of SAS EM (version 14.1), not all are available in the version used by RVO.nl. SAS EM version 12.1 lacks the following modeling techniques mentioned:

- Bayesian networks
- Random forest
- HPSVM (can handle multinomial target values)

#### *Performance*

As was mentioned in Paragraph 2.3.1 logistic regression cannot be applied for some of the explanatory variables and also the target variable contains too many categorical values to create a regression model. Therefore logistic regression is not used in this research.

Another modeling technique that would pose performance problems, is support vector machines (SVM). Although there is a modeling tool for SVM in version 12.1, this tool supports only binary target variables. In theory a predictive model could be created with this modeling tool, however, this is very laborious, see Paragraph 2.3.6. Literature does not specify a maximum number of classes that can be modeled with this technique. In blogs this question was answered, however, the answer is very vague. Performance issues or memory issues are mentioned when more than 30 or more than 100

target values need to be modeled (ResearchGate 2014; MathWorks 2018). Although this is not an exact answer, it is safe to say that when 364 crop type codes for 2017 need to be modeled with SVM, it will be a very laborious task and the probability of performance and memory issues are highly likely. This makes other modeling approaches more preferable and also more likely to succeed.

When the restrictions from the software version and the performance issues are compared, the following modeling techniques remain to be used in this research, see Table 3:

Table 3 Selection of supervised modeling techniques to investigate

Supervised modeling techniques	Modeling technique is present in SAS EM 12.1	Handle multinomial values	Use in research
multinomial logistic regression	y	n	n
classification tree	y	y	y
neural network	y	y	y
Bayesian network	n	y	n
random forest	n	y	n
linear and kernel SVM	n/y	n	n

In order to find the modelling technique that was best suited for predicting crop types, the remaining modelling techniques classification trees and neural network models were tested.

### 3 Methods

The goal of this project was to predict crop types based on historical declaration information. As stated in the introduction, in 2016 RVO.nl performed research into this subject for the first time. Owing to a limited research period, some aspects such as the influence of the length of crop type rotations on predicting crop types and scrutinizing the modeling process could not be studied in sufficient detail. Problems with identifying crop rotation schemas were attributed to changes in crop codes. Therefore, this research focused on identifying the length of crop rotations by adjusting for the change in crop type codification. This information was then used as an extra variable in the modeling process to predict crop types in order to assess if this information helps predict crop types. In this chapter is first described how the variable for the length of crop rotation systems was created, and then all the steps of creating models to predict crop types and the selection process. The choices for the property setting of the modeling techniques were described and how the prediction results were assessed. Also the selection process of the input data for modeling was described and how the study area was created.

#### 3.1 Variable: Length of crop rotation schemas

Every year in the declaration farmers have to state for each parcel they use which crop is cultivated. The declaration forms contain a crop type table in which crop types are described with a corresponding crop type code. A farmer needs to use these crop type codes (CrC) to declare the crop types that will be cultivated that year. The crop codes that changed during the research period (2008 till 2017) were adapted to the 2017 crop type codes. To produce this reference code for the altered crop types, all crop type tables of the declarations were gathered and combined based on the crop type codes. For the crop type codes that ceased to exist during the research period a corresponding crop type code for 2017 was traced for in the combined data. This resulted in a table containing information of the original crop type codes and the reference codes they changed into.

However not all crop codes referred to one corresponding code in 2017. When several crop type codes merged into a new crop type code, there was no problem adapting codes. Still, there were several codes that were split up into a range of new codes:

- 60 old crop type codes were merged into 20 new crop type codes
- 26 old crop types codes were split into 197 new crop type codes

The codes that were split up, were not altered (for there was no right choice) and the assumption was made that this information could still be used in deriving crop rotation schemas from the declaration information. The created table was used to change the crop type codes that ceased to exist during the data mining period into the corresponding crop type code of 2017.

However to use this table to adjust crop types, the input (variable) data needed to be combined into one feature file first. Therefore the available variable data were combined in ArcGIS. Because the aim of this research was to predict crop types for 2017, the declared parcels of 2017 were given a unique identifier (column UNIEK was added to declaration information). Next all the declaration information containing the crop type codes of 2008 till 2017 were combined based on a spatial overlay. The ArcGIS tooling 'union' was used which resulted in the intermediate feature file BBR\_PCR. Only the combined information within the boundaries of the declared parcels of 2017 was kept. See Figure 7 for a flowchart of the data preparation on Page 27.

However, due to differences in the geometry of parcels over the years, parcels were split up in multiple sub parcels when the declaration data were combined. The aim was to split up parcels as little as possible to keep the runtime of the data mining process to a minimum, yet also to keep pollution (slivers) out of the data. Therefore one value per declared parcel (2017) was determined for the variables soil type, groundwater level, long-term average springtime temperature and long-term average springtime precipitation. Largest area-percentage was used to establish the values for each parcel. Farm type data (source: CBS, farm type is based on NSO-typology) were added to the feature



file BBR\_PCR based on the entrepreneur's unique identifier (dienstleutel) and caused no further intersections of the parcels. No information was available of farm types in 2015. Therefore the farm type information of 2014 and 2016 was used to create this variable.

Once the combining of the variable information was complete, crop types were adjusted for the change in crop type codification. The table containing the corresponding crop type codes for 2017 was joined with the feature file BBR\_PCR based on crop codes, and for every year of the data mining period a new column was added to the feature file table containing the adjusted crop types codes. The adjusted crop type codes were stored in new columns, GWC\_ADJ\_2008 till GWC\_ADJ\_2016), see also Table 4 for the variables used in this research and the description of the variables.

The cleaned data were searched for crop rotation schemas. Search patterns were applied as was discussed in Paragraph 2.1.1. The patterns that were searched for differed in length, ranging from one-year crop rotation schemas (which meant that the same crop type was cultivated every year) till six-year crop rotation schemas. During the pattern search, it was important to note that there were different sequences within a rotation year with a specific length. For instance a four-year rotation can have the following crop sequence: A – B – C – D. Other patterns belonging to that same rotation are B-C-D-A, C-D-A-B and D-A-B-C. This is important when crop rotation schemas contain crops that are not used on a regular basis. In the schema A-B-C-D-A-X-C-Y-A, crop type A and C are consistent and make this a four-year crop rotation schema, while crop type B, D, X and Y are inconsistent and therefore more difficult to predict accurately.

The length of the crop rotation cycle was determined by creating models in ArcGIS (one model for every rotation schema length) in which selections were made based on corresponding crop type codes in the different columns that were adjusted for the change in crop type codification (columns GWC\_ADJ\_2008 till GWC\_ADJ\_2016). The length of the crop rotation schema was recorded in a new column LAB\_O\_CRS. Then patterns were searched for by comparing if the adjusted crop types columns had the same crop type code. For example the following selections were made to identify three-year crop rotation schemas (first selection, consistent crop sequence):

```
LAB_O_CRS IS NULL AND  
GWC_ADJ_2016 = GWC_ADJ_2013 AND  
GWC_ADJ_2015 = GWC_ADJ_2012 AND  
GWC_ADJ_2014 = GWC_ADJ_2011 AND  
GWC_ADJ_2013 = GWC_ADJ_2010
```

However, patterns were not always consistent. As explained in Paragraph 2.1.3 missing crop types occurred due to reasons like change in infrastructure or a plot was used for other (non-agricultural) purposes, etc. Also crop rotation schemas existed where only part of the crop types form a pattern. It was conceived that when one year was missing or one crop type is divergent, crop rotation schemas could still be detected. Therefore in the second selection one crop type was inconsistent or one year was missing:

```
(LAB_O_CRS IS NULL AND  
GWC_ADJ_2016 = GWC_ADJ_2013 AND  
GWC_ADJ_2015 = GWC_ADJ_2012 AND  
GWC_ADJ_2014 = GWC_ADJ_2011 AND  
GWC_ADJ_2012 = GWC_ADJ_2009) OR
```

```
(LAB_O_CRS IS NULL AND  
GWC_ADJ_2016 = GWC_ADJ_2013 AND  
GWC_ADJ_2014 = GWC_ADJ_2011 AND  
GWC_ADJ_2013 = GWC_ADJ_2010 AND  
GWC_ADJ_2012 = GWC_ADJ_2009) OR
```

```
(LAB_O_CRS IS NULL AND  
GWC_ADJ_2016 = GWC_ADJ_2013 AND  
GWC_ADJ_2015 = GWC_ADJ_2012 AND  
GWC_ADJ_2013 = GWC_ADJ_2010 AND  
GWC_ADJ_2012 = GWC_ADJ_2009) OR
```

```
(LAB_O_CRS IS NULL AND  
GWC_ADJ_2015 = GWC_ADJ_2012 AND  
GWC_ADJ_2014 = GWC_ADJ_2011 AND  
GWC_ADJ_2013 = GWC_ADJ_2010 AND  
GWC_ADJ_2012 = GWC_ADJ_2009)
```

Values were recorded for one-year crop rotation schemas up till six-year crop rotation schemas. The models and a description of the selections used to identify and record the length of crop rotation schemas are listed in Appendix A. No semi-supervised modeling was applied.

To assess how well the selections found patterns representing crop rotation schemas, crop sequences were created based on the declaration information. The crop sequence is represented by a series of declared crop type codes, which start with the codes declared in 2016 followed by the codes declared in 2015 and so on. The codes in the sequence were separated by a dash. This information was stored in a column in the combined variable data (BBR\_PCR file). The length of a crop rotation schema was compared with the crop sequence to find anomalies. For the input data where no label was present for a crop rotation schema, it was assessed if there were undetected crop rotation schemas present. In order to find undetected crop rotation schemas, the crop sequences were searched for a repetitive crop code patterns.

However, as was mentioned in Paragraph 2.1.3, not all parcels were declared every year of the data mining period. Missing crop type codes appeared in the crop sequences and neural network modeling cannot handle missing values. Records of the input data with missing crop type codes would be excluded from the modeling process, leading to a smaller input data set which might affect the quality and accuracy of the predictive model negatively. To prevent the exclusion of some of these records, the length of crop rotation schemas was used to replace missing crop type codes. In this manner the pattern of a crop rotation schema was reflected in the crop sequence, when missing values were replaced. For example when a missing value occurred in 2013 and the length of the crop rotation schema was 2, a corresponding crop type code was searched for in the declared crop types for 2015 first. When no values were found in 2015, values were searched for in the declared crop types of 2011. The model builder in ArcGIS was used to create selections and replace missing crop type values where possible. For an example of selection criteria to replace missing crop types in the declaration years 2008 and 2009, see Appendix B. In Figure 7 a flowchart is provided of combining the variable information, the identification of the length of crop rotation schemas and replacement of missing crop type codes.

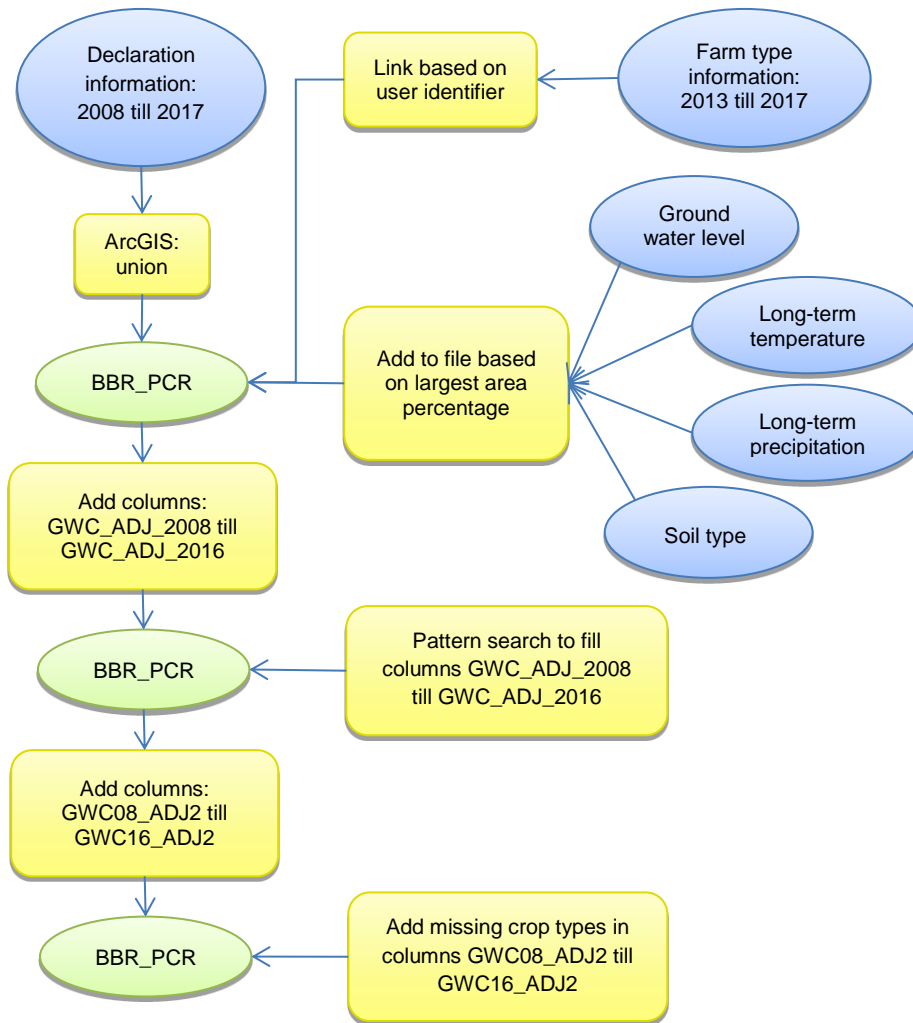


Figure 7 Flowchart of combining the variable information, appliance of search patterns and replacing missing crop type codes

### 3.2 Application of the prediction methods

Two modeling techniques remained to be investigated with respect to their suitability for predicting crop types: classification tree and neural network. The software package SAS enterprise miner (SAS EM) was used to create models. In order to perform a data mining project, SAS developed their own data mining approach SEMMA meaning: Sample, Explore, Modify, Model and Assess (SAS 2011).

SEMMA contains the following steps:

- Sample: in this first step the data are sampled by dividing the input data into more data sets for training, validation and/or testing the model, depending on the modeling technique that will be applied.
- Explore: in the second step the input data are investigated in order to understand the characteristics of the input data. Statistical techniques are applied to find relationships or anomalies in the input data set.
- Modify: in the third step the input data can be modified. New variables can be created and selected, and existing variables can be adjusted. When the input data contains missing values, tools are available to create input for these missing values.
- Model: in the fourth step a model is created. Several modeling tools are present like classification trees and neural network models.
- Assess: In the final step the accuracy and applicability of the model is investigated.

Before using the SEMMA approach, the input data set was selected and a diagram chart was created in SAS EM, for an example of a diagram chart where classification trees are created, see Appendix C. The following variables were used to create the models, see Table 4:

Table 4 All input and target variables used in this research to create predictive models

Variable name	Description	Role	Number of distinct values
GWC_ADJ_2008	Adjusted crop type codes declared in 2008	input	68
GWC_ADJ_2009	Adjusted crop type codes declared in 2009	input	64
GWC_ADJ_2010	Adjusted crop type codes declared in 2010	input	63
GWC_ADJ_2011	Adjusted crop type codes declared in 2011	input	63
GWC_ADJ_2012	Adjusted crop type codes declared in 2012	input	65
GWC_ADJ_2013	Adjusted crop type codes declared in 2013	input	63
GWC_ADJ_2014	Adjusted crop type codes declared in 2014	input	61
GWC_ADJ_2015	Adjusted crop type codes declared in 2015	input	130
GWC_ADJ_2016	Adjusted crop type codes declared in 2016	target	143
GWC08_ADJ2	Adjusted crop type codes declared in 2008, missing values replaced	input	70
GWC09_ADJ2	Adjusted crop type codes declared in 2009, missing values replaced	input	65
GWC10_ADJ2	Adjusted crop type codes declared in 2010, missing values replaced	input	64
GWC11_ADJ2	Adjusted crop type codes declared in 2011, missing values replaced	input	67
GWC12_ADJ2	Adjusted crop type codes declared in 2012, missing values replaced	input	65
GWC13_ADJ2	Adjusted crop type codes declared in 2013, missing values replaced	input	63
GWC14_ADJ2	Adjusted crop type codes declared in 2014, missing values replaced	input	61
GWC15_ADJ2	Adjusted crop type codes declared in 2015, missing values replaced	input	132
GWC16_ADJ2	Adjusted crop type codes declared in 2016, missing values replaced	target	145
BIO_2015	Organic farms	input	2
BTH_2013	Main farm type in 2013	input	8
BTH_2014	Main farm type in 2014	input	8
BTH_2015	Main farm type in 2015	input	16
BTS_2013	Sub farm type in 2013	input	34
BTS_2014	Sub farm type in 2014	input	36
BTS_2015	Sub farm type in 2015	input	59
LG_neerslag	Long-term average (spring) precipitation	input	4
LG_temp	Long-term average (spring) temperature	input	5
BOD_EENVOU	Main soil type	input	16
EERSTE_BOD	Sub soil type (more distinctive)	input	191
DSL_2008	Unique identifier owner/user 2008	input	4156
DSL_2009	Unique identifier owner/user 2009	input	4028
DSL_2010	Unique identifier owner/user 2010	input	3950
DSL_2011	Unique identifier owner/user 2011	input	3877
DSL_2012	Unique identifier owner/user 2012	input	3834
DSL_2013	Unique identifier owner/user 2013	input	3805
DSL_2014	Unique identifier owner/user 2014	input	3757
DSL_2015	Unique identifier owner/user 2015	input	3713
GWT	Groundwater level	input	15
EERSTE_GWT	Groundwater level (highest levels)	input	17
LAB_O_CRS	Length (in years) of crop rotation schema	input	6

All the variables were used to train the models, with GWC\_ADJ\_2016 and GWC16\_ADJ2 as the target variables. The crop type code series GWC\_ADJ\_2008 till GWC\_ADJ\_2016 were used to create classification tree models for these variables contain missing values and this modeling technique could handle missing values. The crop type code series GWC08\_ADJ2 till GWC16\_ADJ2 were used to create neural network models for these variables were adjusted for missing crop type codes as much as possible. Neural network models are unable to handle missing values. The number of distinct variable values is based on the study area that was used for this research, see Paragraph 3.4.

### **3.2.1 SEMMA - sample**

The input data set was split up in two parts: 70% of the data set was used for training the models (the training data set) and 30% of the data was used for a final assessment of the created model (the validation data set). The 70/30 rule division was applied in this research and is a common rule in machine learning research (ResearchGate 2016).

### **3.2.2 SEMMA - explore**

To explore the data, the 'StatExplore' tool of SAS EM was used. The summary statistics table which was produced by the tool reported missing data (i.e. empty cells in the input data) for several variables, see the column 'missing' in Table 5. Replacing missing values was important for neural networks models could not handle missing values. The input data with missing values would be excluded when a model was created, resulting in a smaller subset used to create a model which would influence the quality and accuracy of the predictive model.

All the farm type variables (variable names BTH\_2013 till BTH\_2015 and BTS\_2013 till BTS\_2015) and the variable for organic farming (variable name BIO\_2015) contained missing values.

The table was also an indication that the values for organic farming needed to be adjusted, see Appendix D for the complete class variable summary statistics table. The intention of this research was to differentiate between organic farms and non-organic farms. The column 'number of levels' in the class variable summary statistics table in Appendix D showed there were 3 values present. When the input declaration information was checked, there were two values present; 1 and 2. The third number of level represents the empty cells (null).

Table 5 Number of missing variable values from the class variable summary statistics table produced by the StatExplore tool

Class Variable Summary Statistics  
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Missing
TRAIN	BIO_2015	28005
TRAIN	BOD_EENVOU	0
TRAIN	BTH_2013	244
TRAIN	BTH_2014	1588
TRAIN	BTH_2015	567
TRAIN	BTS_2013	244
TRAIN	BTS_2014	1588
TRAIN	BTS_2015	567
TRAIN	DSL_2008	0
TRAIN	DSL_2009	0
TRAIN	DSL_2010	0
TRAIN	DSL_2011	0
TRAIN	DSL_2012	0
TRAIN	DSL_2013	0
TRAIN	DSL_2014	0
TRAIN	DSL_2015	2
TRAIN	EERSTE_BOD	0
TRAIN	EERSTE_GWT	0
TRAIN	GWC_ADJ_2008	1327
TRAIN	GWC_ADJ_2009	707
TRAIN	GWC_ADJ_2010	533
TRAIN	GWC_ADJ_2011	505
TRAIN	GWC_ADJ_2012	253
TRAIN	GWC_ADJ_2013	146
TRAIN	GWC_ADJ_2014	75
TRAIN	GWC_ADJ_2015	51
TRAIN	GWT	0
TRAIN	LAB_0_CRS	0
TRAIN	LG_neersl	0
TRAIN	LG_temp	0
TRAIN	GWC_ADJ_2016	129

### 3.2.3 SEMMA – modify

In the modify step the data were adjusted. Here missing values found during the exploration phase were replaced. To modify the input variables, the replacement tool of SAS EM was used. This tool provides a replacement editor to change input values of variables, see Table 6.

Variable	Formatted Value	Replacement Value	Frequency Count
BIO_2015		<b>NO_BIO</b>	28005
BIO_2015	01	<b>BIO</b>	1425
BIO_2015	02	<b>BIO</b>	192
BIO_2015	_UNKNOWN_	<b>_DEFAULT_</b>	.
BTH_2013	Akkerbouwbedrijven		15542
BTH_2013	Graasdierbedrijven		8126
BTH_2013	Gewas/veecombinaties		1906
BTH_2013	Hokdierbedrijven		1617
BTH_2013	Gewascombinaties		1141
BTH_2013	Tuinbouwbedrijven		711
BTH_2013		<b>ONBEKEND</b>	244
BTH_2013	Veeteeltcombinaties		233
BTH_2013	Blijvendeteeltbedrijven		102
BTH_2013	_UNKNOWN_	<b>_DEFAULT_</b>	.
BTH_2014	Akkerbouwbedrijven		15393
BTH_2014	Graasdierbedrijven		7484
BTH_2014	Gewas/veecombinaties		1665
BTH_2014		<b>ONBEKEND</b>	1588
BTH_2014	Hokdierbedrijven		1416
BTH_2014	Tuinbouwbedrijven		904
BOD_FEMWOL	_UNKNOWN_	<b>_DEFAULT_</b>	.

Table 6 Part of the replacement editor table

In the replacement editor individual variable values were replaced. For example, the values listed for organic farming (null, 1 and 2 in column 'Formatted Value') were replaced by the word 'BIO' (column 'Replacement Value') and the empty values (null) by the word 'NO\_BIO'. All the missing values of the farm type variables were replaced with the word 'ONBEKEND'. When variables were adjusted with the replacement tool, the adjusted values were stored in a new variable. All new variables got the same name as the original column, preceded by the abbreviation REP (representing the replacement action).

### 3.2.4 SEMMA – model: Classification tree

In this fourth step of the SEMMA approach, classification trees were built first.

When a classification tree is built, the input data are split up into smaller subgroups based on decision rules (if-then statements). Subgroups are also called nodes. Aim of dividing the input data into smaller subgroups is to increase the homogeneity of values within the created subgroups. Another aim of the division is that when values of the different subgroups on the same depth-level of the tree are compared, these are very different. Repeating this process of splitting up the subgroups into even smaller subgroups increases the homogeneity of values within the smaller subgroups and the heterogeneity among subgroups even more. By splitting the input data into subgroups (and subgroups of subgroups), branches are formed (De Ville et al. 2013).

For this research the Gini impurity index was used (De Ville et al. 2013). It measures the variability or purity of categorical variables. It is a "measure of node impurity where  $p_1, p_2, \dots, p_r$  are the relative frequencies of each class in a node" (De Ville et al. 2013). In this research  $r$  represents the crop type codes of 2016 (the target variable):

$$\text{Gini impurity} = 1 - \sum_j^r p_j^2$$

Groups with a Gini index near to 0 represent pure nodes (the values within a group after a split are very similar). In groups with a Gini index that is nearer to 1, the values within the group are very different which means the node is not very pure. As the Gini index searches for impurity, the splitting criterion (based on a decision rule of the variable) will be used that produces the most pure nodes.

The process of splitting nodes into sub nodes is repeated until a stopping rule (like maximum depth) prevents the growing of that particular subtree any further or to prevent the tree from overfitting the training data. To prevent overfitting, validation of the model was applied. A tree is built based on the training data and this model is tested on the validation data. Then the prediction results (represented by the misclassification rates) of both trees are compared branch by branch. The more a tree grows, the more the misclassification rate differs when the branches of the training data and the validation data are compared. As soon as the misclassification rate for the validation data does not decrease any further, or even increases again, that is the sign that the prediction ability of the model decreases. The development of that particular subtree will then be stopped (De Ville et al. 2013).

Cross validation was used to apply validation. In predictive modeling 10-fold cross validation is often applied (Refaeilzadeh et al. 2009), it is an accepted standard procedure to assess the performance of models. By applying 10-fold cross validation, the training data set was split up at random into 10 subsets. Nine subsets were then used to train a model and the tenth subset was used to validate the model. This procedure was repeated 10 times so that every subset was used as validation data and 10 results were produced. The model that predicted crop types the most accurate and where the misclassification rate of the subsets did not differ too much was chosen. The advantage of cross validation was that the training data set would not be reduced by reserving a part of this data for the validation of the model (De Ville et al. 2013).

As multiple classification trees were built during this research, the misclassification rate was set as the assessment method to select the best predictive tree. It was assumed that the lower the misclassification rate was, the more accurate the model would predict crop types.

To create a classification tree model, the 'decision tree' tool was applied in SAS EM and the following settings were used to create classification trees, see Table 7:

Table 7 Table containing the setting used to create classification trees

Decision Tree Node Train Properties	Settings	Default
Variables		-
Interactive		-
Use Frozen Tree	No (default)	y
Use Multiple Targets	No (default)	y
Precision	4	y



Decision Tree Node Train Properties: Splitting Rule	Settings	Default
Interval Criterion	-	-
Nominal Criterion	Gini	n
Ordinal Criterion	-	-
Significance Level	-	-
Missing Values	Most correlated branch	n
Use Input Once	default	y
Maximum Branch	trial and error	n
Maximum Depth	trial and error	n
Minimum Categorical Size	trial and error	n
Split Precision	-	-
Decision Tree Node Train Properties: Node		
Leaf Size	5 (default)	y
Number of Rules	5 (default)	y
Number of Surrogate Rules	0 (default)	y
Split Size	10 (default)	y
Decision Tree Node Train Properties: Split Search		
Use Decisions	No	y
Use Priors	No	y
Exhaustive	5000 (default)	y
Node Sample	20000 (default)	y
Decision Tree Node Train Properties: Subtree		
Method	Assessment	n
Number of Leaves	-	-
Assessment Measure	Misclassification	n
Assessment Fraction	-	-
Decision Tree Node Train Properties: Cross Validation		
Perform Cross Validation	Yes	n
Number of Subsets	10	n
Number of Repeats	1	n
Seed	12345	y
Decision Tree Node Train Properties: Observation-Based Importance		
Observation Based Importance	No	y
Number Single Var Importance	-	-
Decision Tree Node Train Properties: P-Value Adjustment		
Bonferroni Adjustment	Yes	y
Time of Kass Adjustment	Before	y
Inputs	No	y
Number of Inputs	-	-
Split Adjustment	Yes	y
Decision Tree Node Train Properties: Output Variables		
Leaf Variable	No	y
Performance	Disk	y
Decision Tree Node Score Properties		
Variable Selection	Yes	y
Leaf Role	Segment	y

Many settings were left to the default, however, some settings were adjusted. The nominal criterion was set to Gini because this setting produced lower misclassification rates when models were created, see Appendix E for the comparison of the misclassification rates when Gini and entropy were used. The splitting rule for missing values was set to 'most correlated branch'. In this research the only remaining missing values were crop type codes in all the declaration years. Classification trees are able to handle missing data, however, how missing values should be treated is clarified through the setting for this property. Instead of using surrogate rules to handle missing values, missing values were replaced (De Ville et al. 2013). Research performed by Feelders (1999) indicated that better results are produced when missing values are replaced instead of handled via surrogate rules. The theory was that a value of any field could be seen as a function of the values in other fields for the same record. By using this function as a predictive equation in the classification tree, a valid value was predicted for all the missing values. Based on the predicted values, the corresponding records were assigned to the branch containing the most similar values (i.e. most correlated branch).

The splitting rule 'use input once' was left default (No). It was conceived that a variable might be used more than one time to split the input data because the target variable contained many distinctive values. If variables were used only once, the splitting performance would probably not produce an optimal modeling result. All further remaining properties were left to the default for these properties were not relevant due to the nature of the input data (nominal variables) or no appropriate settings could be derived for the input data (for instance the leaf size and the split size). For a full description of the SAS help section on the decision tree tool, see Appendix F.

To find the best result for a classification tree, different settings were used for the properties maximum branch, maximum depth and minimum categorical size. The property maximum branch specifies the maximum number of groups in which the input data are divided during a split. The property maximum depth specifies the maximum number of times a splitting procedure may be applied. The property minimum categorical size specifies the number of variable values that must be present in the input data before it can be used in a decision rule. Through trial and error a model was created with a misclassification rate as small as possible. Two types of classification tree model were created: models with different settings using only maximum branch = 2, and models with different settings using a maximum branch that was larger than 2. When the property maximum branch = 2 was applied, the input variable data was split into two sub data sets during the splitting process. This is also called a two-way split classification tree. Using a larger setting for this property meant that the input data was split up into more than two subgroups corresponding with the number of the maximum branch property. This is also called a multi-way split classification tree. Reason for creating a two-way classification tree was interpretability of the splitting process and the results.

For both the two-way split classification trees and the multi-way split classification trees the Gini index tries to maximize purity in all the subgroups that are created during a split. There is no difference in how the index is applied for these two types of classification trees. For two-way split trees the input data was always split into two groups. This is in accordance with the CART procedure of Breiman (De Ville et al. 2013). For multi-way split trees the input data was split up in a maximum number of subgroups defined by the value for the property 'maximum branch'. This meant that any number of subgroups could be formed, restricted by the value for the property 'maximum branch', see correspondence with SAS Tech support in Appendix G.

During the creation of two-way split classification trees the maximum value for the property maximum depth was used (maximum depth = 50). To investigate if the misclassification rate would decrease even more, some of the explanatory variables were left out of the modeling process. By leaving out different combinations of the variables for user identifiers 2008 till 2013 and the distinct soil type information (variable EERSTE\_BOD), classification trees with even lower misclassification rates were built.

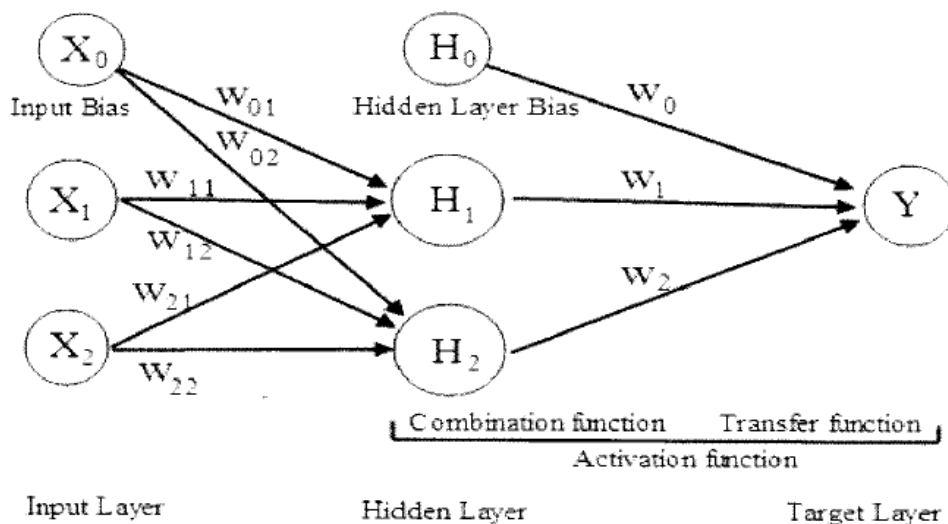
Once the modeling of crop types was finished, all created models were compared using the model comparison tool. The model with the smallest misclassification rate for the validation data set was flagged. See Appendix E for the applied settings and models that were created.

### 3.2.5 SEMMA – model: Neural network

The second modeling technique that was applied, was a neural network model. The autoneural tool was used to create neural network models. Although still several properties needed to be set, this modeling technique was more automated. Again input variables and a target variable (crop types of 2016) were used. This modeling technique searched for the optimum number of hidden layers and the optimum order of functions in the network in order to create a model that predicts crop types the most accurate (Matignon 2007).

A neural network consists of input layers, hidden layers and an output layer, see Figure 8. In Figure 8 all individual input variables (input layers X0, X1 and X2) are connected to all the neurons in the hidden layers (H1 and H2) and all hidden layers are again connected to the target variable (output layer Y). The lines connecting the input data with the neurons in the hidden layers and the output data represent weight vectors (W01, W02, W11, W12, W21, W22, W0, W1 and W2) which are optimized during the fitting of the model. Iterative nonlinear solutions are used to optimize the neural network, also called activation functions. When activation functions are used, two steps are performed. In the first step all the input variables are multiplied with their corresponding weight. The output results are then accumulated into a single value. This is also called the combination function. In the second step (also called the transfer function) a non-linear transformation is applied to the results of the first step. In Figure 8 tanh is used as transfer function. When multiple activation functions are used in a neural network, the interpretability of the model is lost. Therefore validation data are used to verify if the model does not overfit the training data (Matignon 2007):

Figure 8 Representation of a (feed-forward) multi-layered neural network with two input layers (X1 and X2) and two hidden layers (H1 and H2) (Matignon 2007)



Neural Network Model:  $Y = w_0 + w_1 \cdot H_1 + w_2 \cdot H_2 + e$   
 where  $H_1 = \tanh(w_{01} + w_{11} \cdot X_1 + w_{21} \cdot X_2)$   
 and  $H_2 = \tanh(w_{02} + w_{12} \cdot X_1 + w_{22} \cdot X_2)$

There are several activation functions available however which ones to use depends on the (target layer) error function. The error function represents the relationship between the neurons' output results (predicted values) and the target values. A proper function must be chosen to describe this relationship. Only Cauchy and multiple Bernoulli were appropriate in this research for these error

functions apply to nominal values (Matignon 2007). Finally, only the error function ‘multiple Bernoulli’ was used, for initial experiment with the error function ‘Cauchy’ produced very high misclassification rates compared to ‘multiple Bernoulli’ and ‘Cauchy’ was not investigated further. As the target variable for this research was nominal, the activation function that was used in combination with the error function ‘multiple Bernoulli’ was ‘softmax’. The softmax activation function is similar to a multiple logistic function, and produces either a ‘0’ or a ‘1’ (Matignon 2007).

Next to choosing the right error function and activation function, also a network architecture had to be selected. There are four feed-forward architectures to choose from: a single hidden layer network, a funnel network, a block network and a cascade network. The manner in which the hidden neurons are added to the network differs for all the architectures. In a single layer network architecture neurons are added one by one, in a funnel network architecture the added neurons form a funnel pattern and in a block network architecture each added neuron is a new layer. In a cascade network architecture the added neurons connect with other neurons and form multiple connections with neurons and target values (source: help section SAS EM).

To find the most optimal (auto)neural network, the following settings were used, see Table 8:

Table 8 Table containing the setting used to create neural network models

AutoNeural Node Train Properties: Model Options	
Architecture	trial and error
Termination	Overfitting
Train Action	Search
Target Layer Error Function	MBernoulli
Maximum Iterations	8
Number of Hidden Units	2
Tolerance	Medium
Total time	Two hours
AutoNeural Node Train Properties: Increment and Search	
Adjust Iterations	Yes
Freeze Connections	No
Total Number of Hidden Units	75
Final Training	Yes
Final Iterations	5
AutoNeural Node Train Properties: Activation Functions	
Direct	No
Exponential	No
Identity	No
Logistic	No
Normal	No
Reciprocal	No
Sine	No
Softmax	Yes
Square	No
Tanh	No
AutoNeural Node Score Properties	
Hidden Units	No
Residuals	Yes
Standardization	No

For the architectures single hidden layer, funnel network and block network models were built in order to find the best prediction model based on the lowest misclassification rate for the validation data set.

During the modeling process, the model building was terminated as soon as the misclassification rate for the validation data no longer decreased or even increased again. The other reason for ending the model building was that the error in the training data set did not decrease enough anymore. By setting the training action property to search, neurons were added to the network based on the selected architecture. The software itself searched for the best performing network (with the lowest misclassification rate), which was retained as the final model. The property for the target layer error function was set to 'multiple Bernoulli'. Also the settings for the number of maximum iterations and number of hidden units was left to default. Due to the settings for the increment and search properties, the number of maximum iterations was not a restriction. By setting the property for adjust iterations to yes, the software could deduce itself if more than 8 iterations were necessary to create a better predictive model. All other properties were left to default except for total time, this setting was changed from one hour maximum runtime into two hours maximum runtime. For a full description of the SAS help section on the properties of the autoneural tool, see Appendix H.

When autoneural network models were built, the variables for identifying the users (DSL\_2008 up till DSL\_2015) and sub soil type (EERSTE\_BOD) were excluded. These variables contained much more distinctive categorical values than other variables (see Table 4 on Page 28) which complicated the mathematical functions of the model. Too many pseudo variables needed to be formed and the system would have run out of memory very fast. A solution for this problem would have been aggregating variable values. However, user identifiers represent individual agricultural businesses and aggregating this information would have made no sense. Aggregating the variable containing the distinctive soil type (variable EERSTE\_BOD) was unnecessary for aggregated soil type information was already present in the input data set (variable BOD\_EENVOU). Therefore, to still be able to create neural network models, the variables with user identifier information and EERSTE\_BOD were left out of the modeling process.

### **3.2.6 SEMMA - assess**

In the final step of the SEMMA approach the models were assessed. The misclassification rate of the validation data set produced during the modeling process was used to assess the accuracy of the models. The misclassification rate indicated how often the model predicted the target variable wrong. To compare the results of all the created models, the comparison tool of SAS EM was used. By specifying the misclassification rate of the validation data set as a comparison parameter, one model was flagged that produced the lowest misclassification rate. As the differences between misclassification rates of the prediction models were very small, it was decided to predict crop types for 2017 for five models with the lowest misclassification rate in order to assess their predictive abilities.

The score tool of SAS EM was used to predict the crop types for 2017. To be able to perform a prediction, the content of the variables of the input data file had to shift one year. By renaming the all the variables for the declared crop type codes, user identifier, organic farming and farm type (i.e. variable name GWC\_ADJ\_2016 was changed into GWC\_ADL\_2015, GWC\_ADJ\_2015 was changed into GWC\_ADJ\_2014, etc.) and leaving the variable for declared crop types in 2016 empty, the declared crop types for 2017 were predicted. The table result of the prediction was imported in ArcGIS and connected to the corresponding (sub) parcel features. As the information about the actual declared crop types in 2017 was already present in the table information, the predicted result could easily be compared with the actual declaration information.

To explore the quality of the crop prediction, a confusion matrix was created for the model that predicted crop types the best (Osman et al. 2015). A confusion matrix or contingency table has double entries: the actual declared crop types (producer's accuracy) and the predicted crop types

(user's accuracy). The cells in the table contain the number of times a crop type was predicted related to the crop type that was actually declared. The number of accurate predicted crop types is found where the actual declared crop type code is aligned with the same predicted crop type code (in the diagonal element of the table), the other numbers represent inaccurate predicted crop types. The information in the contingency tables shows the prediction accuracy of individual crop types. The crop type description of 2017 was added to the tables and the (aggregated) crop type description used in the research of 2016. Because the complete contingency tables were too large, the top 20 declared crop types of 2017 were used to create the table.

However, the prediction accuracy was created for sub parcels. During the declaration of agricultural information only one crop type is declared per parcel, therefore all the prediction results of the sub parcels were dissolved for the declared parcels of 2017. The models that correctly classified most sub parcels were used. The dissolved result that had the largest area percentage for a declared parcel, was chosen as the prediction result. As incomplete declared parcels were present in the input data sets, only those declared parcels received a prediction result where the largest area percentage represented more than half of the declared parcel's surface. The prediction accuracy of the declared parcels of 2017 was also compared with the prediction accuracy of the research performed in 2016. Therefore the prediction results of the research performed in 2016 were added to the prediction results of the declared parcels of 2017 based on largest area percentage.

The importance of the variable containing the length of crop rotation schemas was also researched. To investigate the influence of this variable, a new classification tree was created based on the same property settings of the tree that correctly classified the most sub parcel. The same variables were used in the modeling process except the variable containing the length of the crop rotation schemas, this variable was left out of the modeling process. For this new tree the crop types of 2017 were also predicted and the prediction results of both trees were compared.

The relative importance of all variables used in the modeling process was also researched. When a variable was used in the splitting procedure, Gini was applied to measure the purity of the splitting result. The more pure the split result was, the more important the variable used in the decision rule to split the data was compared to other variables. By measuring the extent in which the purity increased during all the splits for all the variables, and assigning these measures to the responsible variables, the relative importance of all variables could be determined. The results are stored in the variable importance table of a created model. For a full description of the SAS EM help section on how variable importance is measured, see Appendix I.

### 3.3 Selection of the sub parcel as input for the modeling process

As was mentioned in Paragraph 3.1 the geometry of parcels declared during the research period varied. Due to intersecting the declaration data, the declared parcels were often split up in multiple sub parcels representing input information for modeling, however, some of those sub parcels represented errors, also called slivers. Slivers are typically narrow and long-shaped features which are very small in size. Slivers pollute the variable input information for they contain no relevant or false information and slow the performance of the modeling process. Therefore slivers need to be removed from the input file for the modeling process.

A ratio was used to distinguish between slivers and sub parcels. Therefore the features perimeter was divided by the square root of its area (Nakos 2001):

$$k = \frac{L}{\sqrt{A}}$$

$k$  = numerical expression to describe if the shape is a sliver or not  
 $L$  = length of the shape perimeter  
 $A$  = shape area

The following selection was applied to select slivers:

A < 50 OR  
 ( k > 7 AND A >= 50 AND A < 500 ) OR  
 ( k > 19 AND A >= 500 AND A < 1000 ) OR  
 ( k > 27 AND A >= 1000 )

All features selected by this selection were removed from the input data file for modeling. All features smaller than 50 m<sup>2</sup> were not considered as declaration parcels by RVO.nl. Another rule applied by RVO.nl is that parcels with an initial size smaller than 1000 m<sup>2</sup> are not split up (applied rule when mapping the Dutch agricultural area for permanent grassland and arable farmland). When these parcels were split up due to the spatial overlay of the declaration data, the initial boundary was restored and again the largest area-percentage principle was applied to fill the feature file table with the appropriate information. To execute these actions a model was built in ESRI ArcGIS, see Appendix J for the description.

### 3.4 Defining a study area

A study area was chosen where many crop rotation schemas were expected to be found. Therefore, the study area had to contain mostly arable farming land, see Figure 9. A combination was made of CBS information about municipalities with a large area of arable farmland (Boerenbusiness 2016) and the variables soil type, long-term average temperature, long-term average precipitation and organic farm type, for the CBS table information see Appendix K. For the municipality boundaries the feature file containing municipality boundaries for the year 2016 from RVO's reference database was used (source: Kadaster). The study area was chosen from the group of municipalities with more than 7000 ha of arable farmland.

The only exception to this selection criteria were the municipalities Eijsden-Margraten and Gulpen-Wittem due to the appearance of specific variable value for soil type (loam) and long-term average precipitation in spring (200 – 225 mm). This resulted in the selection of eight municipalities with a large variety in distinctive values of the mentioned variables. Also the pilot area of the RVO project 'Pilot Monitoring Farmland' was added to the study area and it contains all parcels linked to organic farms, see Table 9:

Table 9 Overview of variable values or variable ranges for the municipalities within the study area

Municipality	Area (hec)	Main soil type	Longterm average temperature in spring (°C)	Longterm average precipitation in spring (mm)
Oldambt	10553	Zeekleigronden en moerige gronden	8,4 - 8,7	150 - 175
Aa en Hunze	7230	Humuspodzolen en zandgronden	8,7 - 9,0	150 - 200
Emmen	13763	Moerige gronden, humuspodzolen en veengronden	8,7 - 9,3	150 - 175
Noordoostpolder	27044	Zeekleigronden en zandgrond	9,0 - 9,3	150 - 200
Dronten	17967	Zeekleigronden	9,0 - 9,3	150 - 200
Schouwen-Duiveland	9396	Zeekleigronden	9,0 - 9,6	125 - 175
Eijsden-Margraten	1391	Brikgronden en leemgronden	9,6 - 9,9	200 - 225
Gulpen-Wittem	1365	Brikgronden en leemgronden	9,6 - 9,9	200 - 225
PMF pilot area	14083	Rivierkleigronden en zeekleigronden	9,3 - 9,6	175 - 200
BBR2017 biological farms	19085	Zeekleigronden en humuspodzolen	8,4 - 9,6	125 - 225

Only those parcels were selected that had an arable crop type code in 2017. The study area covered 186.692,0 ha which represented 9,9% of the total declared area of 2017.



Figure 9 Study area

To cope with the memory problem when running SAS EM, the input data were split into two groups: a group where a label was found for crop rotation schemas ( $LAB\_O\_CRS > 0$ ) and a group where no rotation schemas were found ( $LAB\_O\_CRS = 0$ ). The size of the group without crop rotation labels was reduced by manually picking parcels from the initial study area. The reduction of the study area resulted in a group of sub parcels with a total area of 60239,0 ha. This area represented 3,2% of the total declared area of 2017. For a detail of the adjusted study area, see Figure 10.



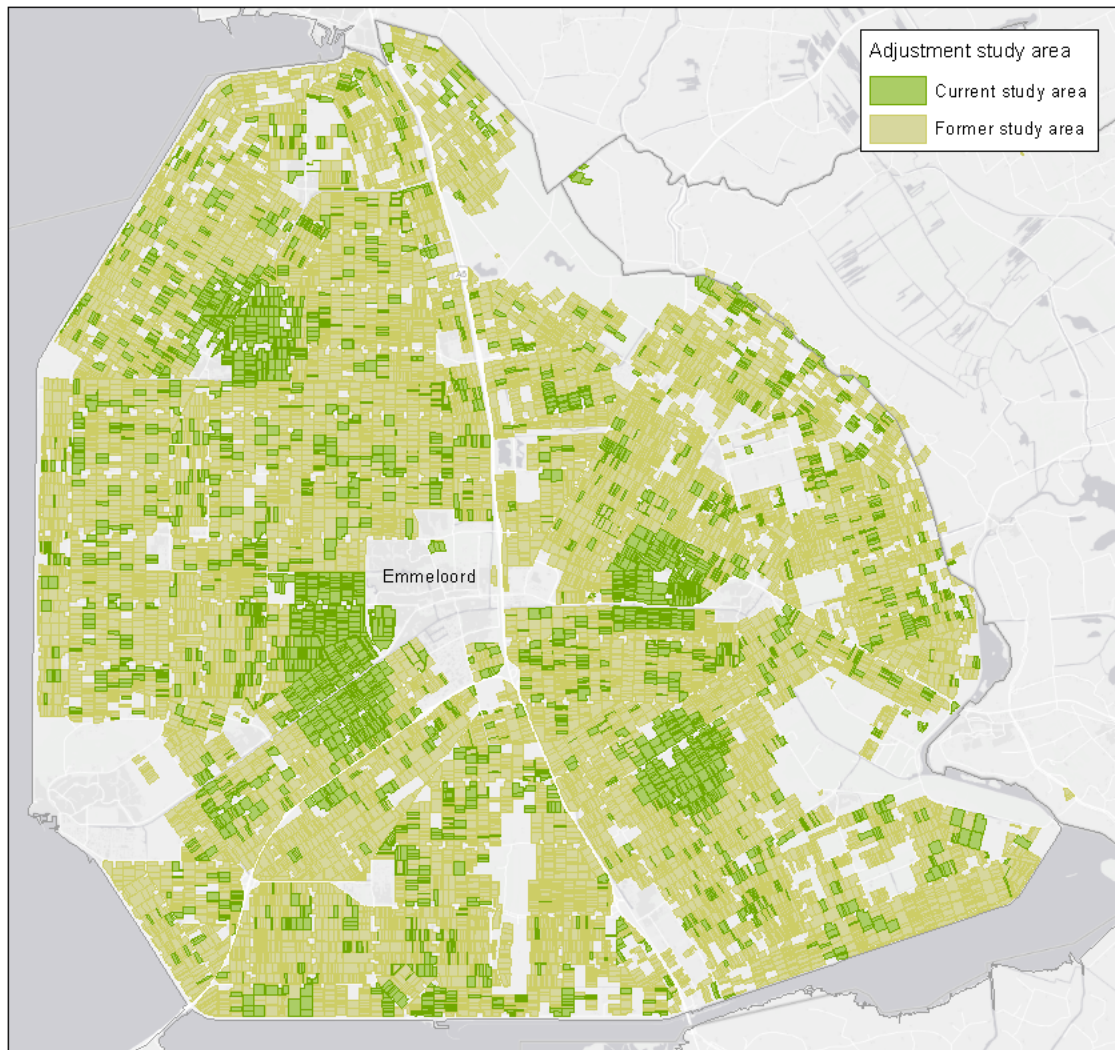


Figure 10 Detail of the adjusted study area

## 4 Results

In this chapter, results are presented of the predictive modeling process. First an overview is presented of the identification of crop rotation schemas. Next, the results are presented of the sliver removal script, followed by the results of the model selection. Then results of the models' prediction accuracy are provided and this chapter concludes with the results of the importance of variables in the prediction process.

### 4.1 Identification of crop rotation schemas

Rotation schemas were found ranging from one year to six years. In total 29.622 rotation schemas were found in the input data set representing the study area. Also a distinction was made in the selections for the presence of missing data in the sequence of crop rotation schemas. The results show that many more rotation systems were found when missing values were allowed in the crop sequence, see Table 10:

Table 10 Number of crop rotation schemas found per length

Length of the crop rotation schema	Total	Total (no missing values in crop sequence)	Total (missing values in crop sequence)
1	11.827	7.231	4.596
2	1.491	270	1.221
3	3.887	1.177	2.710
4	5.192	1.372	3.820
5	4.186	491	3.695
6	3.039	0	3.039
Total	29.622	10.541	19.081

Also monocultures were found (length of the rotation schema is 1). Most often they concern grass, corn or (winter-) wheat.

The study area contained in total 185.108 sub parcels in which the variable information for rotation schemas was searched, see Table 11 :

Table 11 Total number of crop rotation schemas found in the population (study area)

	Total number of sub parcels	Total percentage	Area sub parcels (ha)	Total percentage
Variable present	29.622	16%	42.052	23%
No variable present	155.486	84%	144.639	77%
Total	185.108	100%	186.691	100%

For 16% of the total population's sub parcels a rotation schema was found. This group represents almost a quarter of the study area's surface, therefore semi-supervised modeling was not applied for finding more rotation schemas.

The sequence of crops based on the declaration years was used to determine the length of the rotation schema. In the data set where labels were present for crop rotation schemas 5810 distinct crop sequences were present. In the data set where no labels were found for crop rotation schemas 9525 distinct crop sequences were present. The crop sequences were compared with the length of the crop rotation schemas. Several results were not in line with expectations. Table 12 shows results of crop sequences from two-year crop rotation schemas that might also represent one-year rotation schemas or even four-year rotation schemas.

Table 12 Crop sequences of two-year rotation schema for which the rotation length is debatable

Crop sequence	Frequency of occurrences
233 - 233 - 233 - 234 - 233 - 234	3
233 - 233 - 233 - 256 - 233 - 256	11
234 - 234 - 234 - 238 - 234 - 238	4
234 - 234 - 234 - 256 - 234 - 256	6
236 - 236 - 236 - 2016 - 236 - 2016	3
237 - 237 - 237 - 238 - 237 - 238	4
238 - 238 - 238 - 234 - 238 - 234	1
258 - 258 - 258 - 238 - 258 - 238	4
259 - 259 - 259 - 2014 - 259 - 2014	5
259 - 259 - 259 - 2016 - 259 - 2016	20
259 - 259 - 259 - 233 - 259 - 233	15
259 - 259 - 259 - 238 - 259 - 238	2
259 - 259 - 259 - 266 - 259 - 266	1
266 - 266 - 266 - 265 - 266 - 265	4
266 - 266 - 266 - 672 - 266 - 672	1
266 - 266 - 266 - 814 - 266 - 814	1

The input data where no labels were found for crop rotation schemas were also verified for the presence of crop rotation schemas. Crop sequences were found where a detectable pattern for a rotation schema was present, see Appendix L for a list of the sequences containing a detectable pattern. Because 9525 different crop sequences were present in the input data set, only those sequences were checked that occurred more than 10 times in the data set (95 results in total). Among these 95 crop sequences, 56 results contained a detectable pattern for a rotation schema. Another 854 sub parcels would have been labeled with a length for a crop rotation schema on top of the 29.622 existing schemas.

## 4.2 Selection of the prediction method

Two different types of classification trees were built (two-way split trees and multi-way split trees) and two different input data sets were used (crop rotation schema present in the data set or not). This resulted in four tables with misclassification rates for the classification trees and two tables with misclassification rates for the neural network models:

### 4.2.1 Classification trees

Several classification tree models were created. In tables 13 till 16 only the top 5 results are displayed based on the lowest misclassification rate for the validation data set. The complete tables with results of all models created can be found in Appendix E. A description of the variable names mentioned under the heading 'variable usage' in the tables is provided in Table 4 on Page 28. Under the heading 'Misclassification rate' the misclassification rates are provided for the training data set and the validation data set. During the creation of the two-way split classification trees some of the variables were left out of the modeling process. Under the heading 'Variable usage' is stated if a variable was used or not in the modeling process. The created trees are ranked according to the lowest misclassification rate for the validation data set:

Table 13 Two-way split tree: misclassification rates of data models where labels for crop rotation schemas are present

Data set two-way split classification tree: labels for crop rotation schemas are present, maximum branch = 2, maximum depth = 50											
Misclassification rate				Variable usage							Parameter settings
Nr.	Tree name	Train	Validate	DSL 2008	DSL 2009	DSL 2010	DSL 2011	DSL 2012	DSL 2013	EERSTE BOD	Minimum categorical size
1	Tree15	0,1128277	0,1493995	No	No	No	No	No	Yes	No	5
2	Tree24	0,1098827	0,149624	No	No	No	Yes	Yes	Yes	No	3
3	Tree3	0,1146623	0,1511954	No	No	No	No	No	No	No	5
	Tree22	0,1138898	0,1520934	No	No	No	Yes	Yes	Yes	No	4
	Tree12	0,1179935	0,1520934	No	No	No	Yes	Yes	Yes	No	5

Table 14 Two-way split tree: misclassification rates of data models where no labels for crop rotation schemas are present

Data set multi-way split classification tree: labels for crop rotation schemas are not present, maximum branch = 2, maximum depth = 50											
Misclassification rate				Variable usage							Parameter settings
Nr.	Tree name	Train	Validate	DSL 2008	DSL 2009	DSL 2010	DSL 2011	DSL 2012	DSL 2013	EERSTE BOD	Minimum categorical size
1	Tree25	0,1612189	0,24853	No	No	Yes	Yes	Yes	Yes	Yes	3
2	Tree20	0,1618148	0,254214	No	No	No	Yes	Yes	Yes	Yes	3
3	Tree15	0,1885427	0,2687181	No	No	Yes	Yes	Yes	Yes	Yes	5
	Tree21	0,182414	0,2687181	No	No	Yes	Yes	Yes	Yes	Yes	4
	Tree17	0,1916071	0,2700902	No	No	No	Yes	Yes	Yes	Yes	5

Table 15 Multi-way split tree: misclassification rates of data models where labels for crop rotation schemas are present

Data set Classification tree: labels for crop rotation schemas are present, maximum branch > 2						
Misclassification rate				Parameter settings		
Nr.	Tree name	Train	Validate	Maximum branch	Maximum depth	Minimum categorical size
1	Tree37	0,0681829	0,1420198	7	11	1
2	Tree38	0,0929635	0,1457394	7	11	2
3	Tree35	0,1119732	0,1532913	7	11	3
	Tree39	0,1254061	0,159716	8	11	5
	Tree28	0,1248242	0,1630974	7	11	5

Table 16 Multi-way split tree: misclassification rates of data models where no labels for crop rotation schemas are present

Data set Classification tree: labels for crop rotation schemas are not present, maximum branch > 2						
Misclassification rate				Parameter settings		
Nr.	Tree name	Train	Validate	Maximum branch	Maximum depth	Minimum categorical size
1	Tree15	0,1181478	0,2191298	9	10	1
2	Tree16	0,147174	0,244022	9	10	2
3	Tree2	0,1737317	0,25343	9	10	3
	Tree19	0,1869254	0,2659741	7	10	3
	Tree3	0,1905005	0,2679341	9	10	4

For Table 13 and Table 15 the sub data set was used where labels for the length of crop rotation schemas were present. For Table 14 and Table 16 the sub data set was used where no labels were present for crop rotation schemas.

Results show that the models where a label for crop rotation systems was present, had a substantially lower misclassification rate for the validation data set than the models where no label was present. The lowest misclassification rates of the validation data set for tree models where a label was present for crop rotation schemas in the input data set, hardly differed; the two-way split tree had a slightly higher misclassification rate than the multi-way split tree, see Table 13 and Table 15. The lowest misclassification rate of the validation data set for trees where no label was present for crop rotation schemas, differed more: the multi-way split tree had a lower misclassification rate than the binary tree, see Table 14 and Table 16.

#### 4.2.2 Neural networks

Also several neural network models were created. In the tables the misclassification rates of the models are displayed for which the two sub data sets were used during the modeling, see table 17 and Table 18:

Table 17 Neural Network: misclassification rates of data models where labels for crop rotation schemas are present

		Data set Neural Network: labels for crop rotation schemas are present				
		Misclassification rate		Parameter settings		
Nr.	Neural Network name	Train	Validate	Architecture	Target layer error function	Activation function
1	AutoNeural	0,4568628	0,4661578	Single layer	MBernoulli	Softmax
2	AutoNeural4	0,5606141	0,5644854	Funnel Layer	MBernoulli	Softmax
3	AutoNeural3	0,5629315	0,5668425	Block layer	MBernoulli	Softmax
	AutoNeural2	0,7405011	0,7412729	Single layer	Cauchy	Direct, Normal, Sine, Softmax, Tahn
	AutoNeural5	0,7944769	0,7999776	Single layer	MBernoulli	All activation functions

Table 18 Neural Network: misclassification rates of data models where no labels for crop rotation schemas are present

		Data set Neural Network: no labels for crop rotation schemas are present				
		Misclassification rate		Parameter settings		
Nr.	Neural Network name	Train	Validate	Architecture	Target layer error function	Activation function
1	AutoNeural2	0,6864147	0,6926695	Single layer	MBernoulli	Softmax
2	AutoNeural4	0,7167177	0,7236378	Funnel Layer	MBernoulli	Softmax
3	AutoNeural3	0,7210589	0,7267738	Block layer	MBernoulli	Softmax
	AutoNeural6	0,7800477	0,7843983	Single layer	MBernoulli	all activation functions
	AutoNeural	0,7873681	0,7898863	Single layer	Cauchy	Direct, Normal, Sine, Softmax, Tahn

Results show again that the models where a label for crop rotation systems was present, had a substantially lower misclassification rate than the models where no label was present. In both cases (data set with and without a label for crop rotation schemas) the model with the single layer network architecture produced the lowest misclassification rates for the validation data set.

#### 4.2.3 Indication best predictive models

When all the created models were compared, the classification trees had a substantially lower misclassification rate than the neural network models. This indicated that classification trees would yield more accurate prediction results than neural network models. Therefore predicting the crop types for 2017 was only performed with the classification tree models containing the lowest misclassification rates.

### 4.3 Assessment of the prediction accuracy

All the trees mentioned in tables 13 till 16 were used to assess the prediction accuracy.

#### 4.3.1 Prediction accuracy of the classification tree models

For all the created trees mentioned in Tables 13 till 16 the crop types of 2017 were predicted. The tables contain the ranking of Tables 13 till 16 and the percentage correctly classified sub parcels.

Table 19 Classification accuracy for two-way tree models on data with crop rotation label

Rank	Model	Rank Table 13	Number of sub parcels classified correctly	Total population	Percentage correct
3	Tree15	1	13.171	29.622	44,46
1	Tree24	2	13.367	29.622	45,13
5	Tree3	3	13.002	29.622	43,89
4	Tree22	4	13.153	29.622	44,40
2	Tree12	5	13.201	29.622	44,56

Table 20 Classification accuracy for two-way tree models on data without crop rotation label

Rank	Model	Rank Table 14	Number of sub parcels classified correctly	Total population	Percentage correct
5	Tree25	1	4.981	17.079	29,16
4	Tree20	2	5.070	17.079	29,69
2	Tree15	3	5.296	17.079	31,01
1	Tree21	4	5.343	17.079	31,28
3	Tree17	5	5.266	17.079	30,83

Table 21 Classification accuracy for multi-way tree models on data with crop rotation label

Rank	Model	Rank Table 15	Number of sub parcels classified correctly	Total population	Percentage correct
5	Tree37	1	13.205	29.622	44,58
4	Tree38	2	13.294	29.622	44,88
3	Tree35	3	13.354	29.622	45,08
1	Tree39	4	13.560	29.622	45,78
2	Tree28	5	13.470	29.622	45,47

Table 22 Classification accuracy for multi-way tree models on data without crop rotation label

Rank	Model	Rank Table 16	Number of sub parcels classified correctly	Total population	Percentage correct
4	Tree15	1	4.915	17.079	28,78
1	Tree16	2	5.344	17.079	31,29
3	Tree2	3	5.043	17.079	29,53
5	Tree19	4	4.851	17.079	28,40
2	Tree3	5	5.225	17.079	30,59

Results show that the ranking of the misclassification rates of the validation data set in Tables 13 till 16 is different from the ranking of the prediction accuracy of 2017 (column name Nr.). For instance Tree39 in Table 21 did not have the lowest misclassification rate (rank number 4 in Table 15),

however, it predicted the crop types for 2017 the most accurate of all five trees mentioned in Table 21.

Results also show that the prediction accuracy is higher for the input data set where labels were present for crop rotation schemas (Table 19 and Table 21) than for the input data set where no labels are present (Table 20 and Table 22). Furthermore, the number of correctly classified sub parcels did not differ much for two-way split models and multi-way split tree models (comparison of Table 19 and Table 21, and comparison of Table 20 and Table 22). The classification trees that correctly classified the most sub parcels, were Tree39 (Table 21) and Tree 16 (Table 22). Both these classification trees were created with multi-way splitting criteria.

The impact of the label containing the length of crop rotation schemas on the prediction accuracy was also investigated. A model was created using the same property settings and input variables as Tree39, only the variable containing the length of crop rotation schemas was left out of the modeling process. In Paragraph 4.4 these results are displayed.

#### 4.3.2 Prediction accuracy according to label for crop rotation schemas

To investigate the accuracy of the prediction a bit more, the output result of Tree39 was compared with the length of the crop rotation schema and if the crop sequence of the rotation schema contained missing crop codes or not, see Table 23:

Table 23 Correctly and incorrectly classified sub parcels per length of a crop rotation schema split into crop sequence with missing crop types and without missing crop types

Length crop rotation	Missing crop codes	Correctly classified sub parcels		
		Incorrect	Correct	Percentage correct
1	YES	2.597	1.999	43,5
	NO	3.354	3.877	53,6
2	YES	612	609	49,9
	NO	123	147	54,4
3	YES	1.533	1.177	43,4
	NO	521	656	55,7
4	YES	2.060	1.760	46,1
	NO	623	749	54,6
5	YES	2.553	1.142	30,9
	NO	315	176	35,8
6	YES	1.771	1.268	41,7
Total		16.062	13.560	

Results show that a larger number of sub parcels is classified correctly for crop sequences that are complete (i.e. no missing crop type codes). For crop rotation schemas with the length of one till four years where the crop sequence is complete, more than 50% of the sub parcels are classified correctly. Accurate predictions for sequences with the length of five or six years are considerably worse when compared with prediction results of crop type sequences with a shorter rotation schema.

A further analysis is provided in Appendix M. An overview is given of the number of sub parcels that appear on a declared parcel and a table containing the number of sub parcels and declared parcels per area class. Furthermore, the accuracy of the sub parcels was researched per area class. Results show that in 57,2% of the cases a declared parcel is split up in more than one sub parcels. Also relative more sub parcels are misclassified when the area of a sub parcel is small. On average almost 64% of the sub parcels is predicted inaccurately when the area is smaller than 100 m<sup>2</sup> versus 51% when the area is larger than 5 ha.

### 4.3.3 Contingency tables

To explore the quality of the crop type prediction, contingency tables were created. The complete contingency tables were too large to fit in the text. Therefore, for clarification purposes two smaller contingency tables are displayed in Table 24 and Table 25. These tables are based on the top 20 declared crop types that appeared the most in the input data sets. Note that the total number of sub parcels for which the prediction accuracy was investigated, equaled 26.044 in Table 24 and 14.459 in Table 25. As the multi-way split classification trees correctly classified the most sub parcels, the prediction results from Tree39 and Tree16 were used. The percentages of correctly predicted crop types (user's accuracy) and the percentages of actual declared crop types codes that were predicted correctly (producer's accuracy) were added to the table. Also the crop type descriptions of the research performed in 2016 and the current research were added. Crop types that were predicted most accurately are displayed in bright green, and crop types with very poor prediction results are displayed in orange. Cells are highlighted pink when declared crop types were mixed up with another crop types more than a 100 times. Cells highlighted in dark orange represent the largest number of times a declared crop type was mixed up with another crop type. The light green cells on the diagonal element in the table display the number of sub parcels which were classified correctly.

Table 24 Contingency table - Multi-way split tree where labels for crop rotation schemas are present (tree 39) with the top 20 declared crop types

	Crop type description of research 2016	horticulture	arable crop type	arable crop type	potatoes	potatoes	potatoes	potatoes	Wheat	Wheat	Wheat	Wheat	Wheat	sugar beet	arable crop type	maize	arable crop type	grass	arable crop type	arable crop type	maize			
	Crop type description of research 2017	Tulip, bulbs/tubers	Rapeseed, winter	Union, seed/plant	Potatoes, consumption	Potatoe, seed NAK	Potatoe, seed TBM	Potatoe, starch	Wheat, winter-	Wheat, summer-	Barley, winter-	Barley, summer-	Rye (geen snijrogge)	Beet, sugar-	Alfalfa	maize, cut-	Union, seed	Grassland, temporary	Winter carrot, production	Chicory root, production	Maize, corncob mix			
<b>User's accuracy - predicted crop types of 2017</b>																								
	Crop code	1004	1922	1931	2014	2015	2016	2017	233	234	235	236	237	256	258	259	262	266	2785	2787	317	Total	% correct	
Producer's accuracy - actual declared crop types in 2017	1004	11		57	17	4				78			7	235		4	43	13			19	488	2%	
	1922		2		3					112		75	2		7		2						203	1%
	1931	12		55	10	6				13	2				38	46	31	90			1	304	18%	
	2014				854	163	8	21	241	1	2	7		78	14	315	52	168				11	1935	44%
	2015			13	126	667	114	4	141			4	3	45	1	33	38	132					1321	50%
	2016					1	11	64	1	4		147	6	2		9		12					257	4%
	2017				8	28	44	596	20	7	4	128		37		183	6	69					1130	53%
	233		84	11	143	38	5	2	4507	62	20	13		481	9	219	80	35			11	1	5721	79%
	234				11	5		3	132	36			5	2	22	4	44	14	23				301	12%
	235				60	5	1		170	1	105	2		24		89	10	2					469	22%
	236				15	36	2	82	83	44	1	325	24	146	2	71	15	6					852	38%
	237											17	200			15		10					242	83%
	256			3	4	98	60	7	64	898	57	13	168		1451	9	358	173	167	3	12	10	3555	41%
	258				11	1	2		129	10	2	1	4	31	85	13	1	24					314	27%
	259	15			108	28	12	37	226	5	14	42	3	144	3	3824	36	471	1			13	4982	77%
	262	28		19	65	49	225	2	389	6		30		241		68	299	163	2	12			1598	19%
	266	10		66	25	33	6	28	188	29		3	6	55	10	701	28	229	1	7	8	1433	16%	
	2785			4	45	9	4		74	15		1	1	175	2	31	99	12	6	5			483	1%
	2787	1		18	8	7			36	29		11		36		17	58	9			66		296	22%
	317				61										12	17						70	160	44%
	Total	77	89	247	1668	1140	441	903	7438	308	236	913	249	3260	139	6059	983	1635	13	133	113	26044		
	% correct	14%	2%	22%	51%	59%	2%	66%	61%	12%	44%	36%	80%	45%	61%	63%	30%	14%	46%	50%	62%			

The user's accuracy for rye was 80% (Table 24) while for starch potato and cut maize it was 66% and 63%, respectively. However, the producer accuracies for rye, starch potatoes and cut maize are 83%,



53% and 77% respectively. The producer's accuracy for winter wheat was 79% which was higher than the producer's accuracy for cut maize and starch potato.

The user's accuracy for rapeseed, seed potatoes (TBM) and summer wheat was 2%, 2% and 12% respectively. The producer's accuracy for these crop types was also very low: 1%, 4% and 12% respectively. However, the producer's accuracy for winter carrots (1%) and tulip bulbs (2%) was lower than the producer's accuracy for seed potatoes (TBM) and summer wheat.

When crop types were predicted inaccurately, a multitude of other crop types was predicted instead. Some of the prediction results stand out. The potato crop types 2014 and 2015 were often mixed up with other potato types and the producer's accuracy shows that the potato types are often mixed up with winter wheat, cut maize and grass. The user's accuracy for winter wheat shows that this crop type is mixed up with several other crop types very often; winter wheat was often predicted instead of the actual declared crop type. And even though the producer's accuracy for winter wheat is 79%, in case the prediction was inaccurate it was mixed up with sugar beets most of the time. The producer's accuracy for sugar beets shows that this crop type was often mixed up with winter wheat and cut maize. And finally the producer's accuracy for cut maize shows that very often temporary grassland was predicted instead of cut maize, and vice versa.

Table 25 on the next page shows the prediction results for the input data set where no label was present for crop rotation schemas. The user's accuracy for oats, perennial ryegrass and winter wheat were 58%, 53% and 48% respectively. The producer's accuracy for these crop types were 18%, 37% and 61% respectively. The producer's accuracy for summer wheat and tulip bulbs was relatively high, 55% and 37% respectively. The user's accuracy for winter barley, temporary grassland and chicory root was 3%, 13% and 13% respectively. The producer's accuracy for these crop types was also very low: 1%, 20% and 5% respectively. However, the producer's accuracy for grass seed is also very low, 8%.

Again when crop types were predicted inaccurate, a multitude of other crop types was predicted instead. Here also some of the prediction results stood out. The potato crop types were often mixed up and the producer's accuracy shows that potato crop type 2014 and 2015 were often mixed up with winter wheat, sugar beets, unions and temporary grasslands. Again the user's accuracy for winter wheat shows that this crop type was mixed up with several other crop types very often; winter wheat was often predicted instead of the actual declared crop type. The producer's accuracy shows that winter wheat again had the highest prediction accuracy percentage. However, when the prediction was inaccurate, the producer's accuracy shows that winter wheat was mixed up with potatoes, sugar beets, cut maize and union most of the time. The producer's accuracy also shows that instead of sugar beets, winter wheat was predicted most of the time, although potatoes, barley, cut maize, unions and temporary grassland were also predicted very often. The producer's accuracy also shows that cut maize was often mixed up with winter wheat, sugar beets and temporary grassland, unions were often mixed up with winter wheat and sugar beets, and temporary grassland was regularly replaced by cut maize and winter wheat.

Table 25 Contingency table - Multi-way split tree where labels for crop rotation schemas are not present (tree 16) with the top 20 declared crop types

	Crop type differentiation research 2016	horticulture	arable crop type	potatoes	potatoes	potatoes	Wheat	Wheat	Wheat	Wheat	Wheat	arable crop type	sugar-beet	arable crop type	maize	arable crop type	grass	arable crop type	arable crop type	grass	arable crop type			
	Crop type description of research 2017	Tulip, bulbs/tubers	Union, seed/plant	Potatoes, consumption	Potatoe, seed NAK	Potatoe, starch	Wheat, winter-	Wheat, summer-	Barley, winter-	Barley, summer-	Oats	Peas, green/yellow	Beet, sugar-	Alfalfa	maize, cut-	Union, seed	Grassland, temporary	Winter carrot, production	Chicory root, production	Perennial ryegrass	Grass seed			
User's accuracy - predicted crop types of 2017																								
	Crop code	1004	1931	2014	2015	2017	233	234	235	236	238	244	256	258	259	262	266	2785	2787	3506	383	Total	% correct	
Producer's accuracy - actual declared crop types in 2017	1004	98		17	7	4	45					6	23	5	9	15	6	29				264	37%	
	1931		20		11		14			3			26	13	4	14	18			10		133	15%	
	2014	14	4	564	54	38	347	25	24	54		8	108	1	47	183	96	43	3	6	5	1620	35%	
	2015	2	4	73	379	12	196	30		32	9		160	1	87	51	210	27			7	1280	30%	
	2017	4		16	27	191	40	18		251			48		82	15	19					711	27%	
	233	64	5	220	102	33	1860	86	9	51		2	188	7	108	158	61	31	21	12	16	3034	61%	
	234			17	2	9	49	223		20		4	38		20	12	9					3	406	55%
	235			3	4		48	5	2	12			31		31	11	1				3	151	1%	
	236				15	66	61	34	1	157	1		70		25	15	3						448	35%
	238				23	8	1	3		8	14	1	4				6	4		4			76	18%
	244		1	13	1		27			12		12	1				8						75	16%
	256	37		152	66	79	402	38	13	169		3	532	6	164	278	164	25	3				2131	25%
	258			26	43	6	63	3	2	3		8	45	34	32	21	11				1		298	11%
	259	3		22	26	57	190	32	11	51			115	1	464	83	169	53	2		3	1282	36%	
	262	45	30	36	21	6	249	7		14			127	14	78	365	59	17	20				1088	34%
	266	5	9	29	29	23	100	17		40		4	46	15	138	61	132	15			3	2	668	20%
	2785	6		54	46		63						49	2	6	57	11	51	1	1			347	15%
	2787	20		3	24		28	1					21		39	4			8				148	5%
	3506			32	8	3	19						20			1					49		132	37%
	383				3		83	2		5			25			3	27	4		1		14	167	8%
	Total	298	69	1277	891	535	3885	524	62	882	24	48	1677	99	1298	1420	981	291	63	92	43	14459		
	% correct	33%	29%	44%	43%	36%	48%	43%	3%	18%	58%	25%	32%	34%	36%	26%	13%	18%	13%	53%	33%			

However, the results of the complete contingency tables showed that some of the declared crop types in 2017 were never predicted. Four percent of the sub parcels contained declared crop types that were never predicted. For two-way split and multi-way split trees where rotation schemas were present in the input data, 101 out of 143 declared crop types were not predicted. These target values were found on 1.173 out of 29.622 sub parcels in the output result of the two-way split tree (Tree24), and 1.193 out of 29.622 sub parcels in the output result of the multi-way split tree (Tree39).

Among the input data where no labels were present for rotation schemas, for the two-way split tree 70 out of 123 declared crop type codes were not predicted (Tree 21), and for the multi-way split tree 64 out of 123 crop type codes were not predicted (Tree16). These target values were found on respectively 755 and 581 out of 17.079 sub parcels. This resulted in the following percentages, see Table 26:

Table 26 Sub parcels (%) containing crop types codes that were not predicted with classification trees

	Percentage of sub parcels containing crop type codes that were not predicted	
	Input data set where labels were present for the length of crop rotation schemas	Input data set where no labels were present for the length of crop rotation schemas
Two-way split tree	4,0%	4,4%
multi-way split tree	4,0%	3,4%

#### 4.3.4 Comparison of prediction accuracy of the former research performed in 2016 and the current research

The prediction result of this research and the research performed in 2016 were also compared. An important fact to mention here is that the research approaches were different which made it difficult to compare the accuracy results. In the former research crop types were aggregated. Only 11 different aggregated crop codes were used: grass, arable crop types, fallow land, forest, nature, horticulture, maize, wheat, potatoes, sugar beet and landscape elements. This research used the actual declared crop codes of the declaration information. The input data sets that represent the study area contained 143 crop codes used in the declaration of 2017.

To gain some perspective on the influence this different approach had on the prediction accuracy of crop types, the crop type description of the research performed in 2016 was added to the contingency tables, see Table 24 and Table 25. The 20 different crop types that were used to differentiate in this current research represent only aggregated 7 crop types in the research of 2016. If the aggregated crop type description of 2016 would have been used in this research, the prediction accuracy would have been higher, see Table 27 where the prediction accuracy of the contingency table based on Tree39 (Table 24) was adjusted for the aggregated crop type description of 2016.

Table 27 Adjusted prediction accuracy according to the aggregated crop type description of 2016 for contingency table based on Tree39

Crop type description of research 2016	horticulture	arable crop type	arable crop type	potatoes	potatoes	potatoes	potatoes	Wheat	Wheat	Wheat	Wheat	Wheat	sugar beet	arable crop type	maize	arable crop type	grass	arable crop type	arable crop type	maize		
	User's accuracy - predicted crop types of 2017																					
Crop code	1004	1922	1931	2014	2015	2016	2017	233	234	235	236	237	256	258	259	262	266	2785	2787	317	Total	% correct
1004	11		57	17	4			78			7		235		4	43	13		19		488	2%
1922		2		3				112		75	2		7		2						203	1%
1931	12		55	10	6			13	2				38		46	31	90		1		304	18%
2014				854	163	8	21	241	1	2	7		78	14	315	52	168			11	1935	44%
2015			13	126	667	114	4	141			4	3	45	1	33	38	132				1321	50%
2016					1	11	64	1	4		147	6	2		9		12				257	4%
2017				8	28	44	596	20	7	4	128		37		183	6	69				1130	53%
233		84	11	143	38	5	2	4507	62	20	13		481	9	219	80	35		11	1	5721	79%
234				11	5		3	132	36		5	2	22	4	44	14	23				301	12%
235				60	5	1		170	1	105	2		24		89	10	2				469	22%
236				15	36	2	82	83	44	1	325	24	146	2	71	15	6				852	38%
237											17	200			15		10				242	83%
256		3	4	98	60	7	64	898	57	13	168		1451	9	358	173	167	3	12	10	3555	41%
258				11	1	2		129	10	2	1	4	31	85	13	1	24				314	27%
259	15			108	28	12	37	226	5	14	42	3	144	3	3824	36	471	1		13	4982	77%
262	28		19	65	49	225	2	389	6		30		241		68	299	163	2	12		1598	19%
266	10		66	25	33	6	28	188	29		3	6	55	10	701	28	229	1	7	8	1433	16%
2785			4	45	9	4		74	15		1	1	175	2	31	99	12	6	5		483	1%
2787	1		18	8	7			36	29		11		36		17	58	9		66		296	22%
317				61									12		17					70	160	44%
Total	77	89	247	1668	1140	441	903	7438	308	236	913	249	3260	139	6059	983	1635	13	133	113	26044	
% correct	14%	2%	22%	51%	59%	2%	66%	61%	12%	44%	36%	80%	45%	61%	63%	30%	14%	46%	50%	62%		

The numbers of sub parcels mentioned in the dark green diagonal row of the table represent the sub parcels of which the crop types were classified correctly for this current research. If the aggregated crop types of the research performed in 2016 had been used to predict crop types, the number of sub parcels mentioned in the light green cells would have been classified correctly also. The pink cells represent the sub parcels where crop types were classified incorrectly for both of the studies. In Table 28 the number of sub parcels was summarized for the mentioned groups.

Table 28 Division of sub parcels according to the prediction accuracy of both studies (2016 and 2017)

	Total number of sub parcels	Percentage
predicted correctly for the research of 2016 and 2017	13.399	51,4%
predicted correctly for the research of 2016 only	1.439	5,5%
predicted incorrectly for the research of 2016 and 2017	11.206	43,0%

Based on the numbers in Table 28 the prediction accuracy for this research was 51,4% for the top 20 declared crop types. The prediction accuracy would have been higher if the aggregated crop type description of 2016 had been used, for more sub parcels would have classified correctly: 51,4% + 5,5% = 56,9%.

However, these numbers represent only the input data set where labels were present. Therefore the results of both contingency tables were combined to establish the prediction accuracy for the top 20 declared crop types in both the input data sets, see Table 29. Based on the numbers the prediction accuracy for this research was 45,8%. Again, the prediction accuracy would have been higher if the aggregated crop type description of 2016 had been used, for more sub parcels would have classified correctly: 45,8% + 5,7% = 51,5%.

Table 29 Division of sub parcels according to the prediction accuracy of both studies (2016 and 2017) for the combined contingency tables of the top 20 declared crop types

	Number of sub parcels where label is present for rotation schemas	Number of sub parcels where no label is present for rotation schemas	Percentage
Predicted correctly for the research of 2016 and 2017	13.399	5.169	45,8%
Predicted correctly for the research of 2016 only	1.439	885	5,7%
Predicted incorrectly for the research of 2016 and 2017	11.206	8.405	48,4%
Total	26.044	14.459	

Still, these accuracy numbers were based on the top 20 of declared crop type codes of both the input data sets and also based on sub parcels. The prediction accuracy of both studies was also compared for complete declared parcels. Again Tree39 and Tree16 were used to dissolve the prediction result of 2017. It became clear that overlap existed in the declaration information, see Appendix N. On 83 declared parcels overlap was found. For these parcels no prediction result was produced. For 9.971 declared parcels within the study area a result was produced. These results were compared with the prediction accuracy of the research performed in 2016, see Table 30. Figure 11 shows a graphical comparison of the prediction results of both studies for a small area southwest of Emmeloord.

Table 30 Correctly classified parcels in 2016 and the current research for complete declared parcels

Prediction result	Correct classified parcels in current research		Correctly classified parcels in research of 2016	
	Number of parcel	Percentage	Number of parcel	Percentage
True	4.256	42,7%	5.519	55.3%

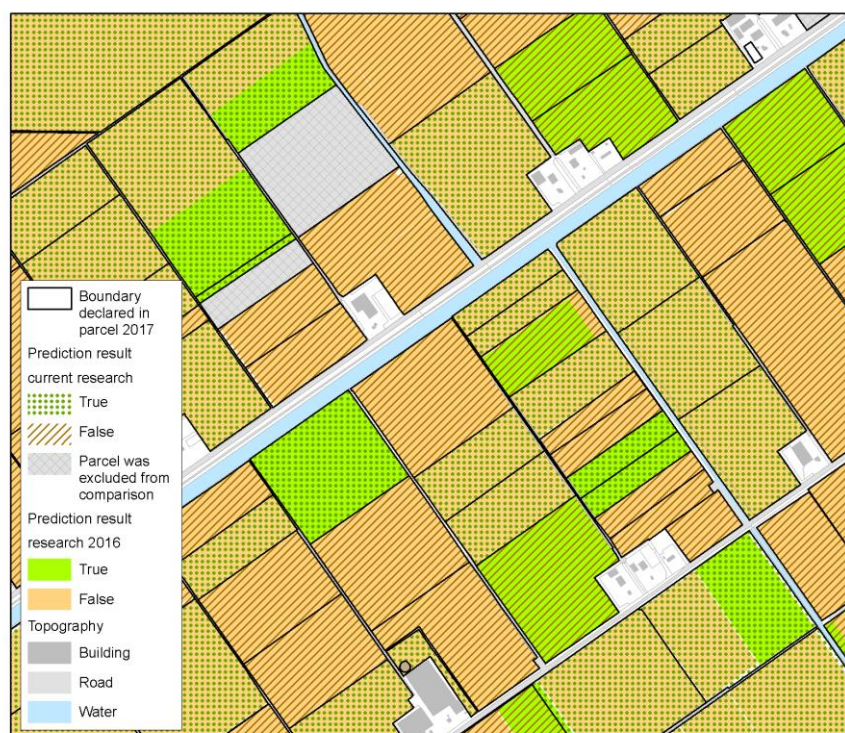


Figure 11 Comparison of prediction results of research performed by RVO.nl in 2016 with the current research based on declared parcel (location: agricultural area southwest of Emmeloord)

For 42,7% of the declared parcels the crop types of 2017 were classified correctly. The accuracy of the former research was higher, for 55,3% of the declared parcels the aggregated crop types were classified correctly.

#### 4.4 Variable importance

The model that predicted crop types the most accurate (Tree39, see Table 21) was used to investigate the importance of the variable containing the length of crop rotation systems. The impact of the variable was researched by creating a new model with the same property settings as Tree39. The only difference was that during the creation of Tree39 the input variable for crop rotation schemas was used as an input variable and during the creation of the new model (Tree2) this input variable was left out. The results of the comparison are displayed in Table 31:

Table 31 The prediction accuracy when the variable with the length of crop rotation schemas is used or not used in the modeling process

Tree name:	Tree39	Tree2	Difference
Label crop rotation schema present	Yes	No	
Misclassification rate (validation data set):	0,159716	0,187106	0,02739
Number of correctly classified sub parcels:	13.560	13.302	258
Percentage of correctly classified sub parcels:	45,78%	44,91%	0,87%

When the variable for crop rotation schemas was used to create a model for predicting crop types, a more accurate prediction result was produced than when the variable was left out of the modeling process. The model (Tree39) classified 258 more sub parcels correctly and the percentage of correctly classified sub parcels increased by 0.87%.

During the modeling process variable importance tables were created. These tables display the relative importance of the input variables used to create the classification trees. The variable with rank number 1 is the most important variable and splits the input data into homogeneous groups the best. The variable importance for Tree39 and Tree6 are displayed in Table 32 and Table 33:

Table 32 Relative variable importance of Tree39 (variable with length of crop rotation schemas was used)

Variable Importance				
Obs	NAME	LABEL	NRULES	IMPORTANCE
1	GWC_ADJ_2013		11	1.0000
2	GWC_ADJ_2010		15	0.5609
3	DSL_2010		7	0.4686
4	LAB_0_CRS		35	0.4391
5	GWC_ADJ_2012		28	0.4268
6	DSL_2014		6	0.4152
7	REP_DSL_2015	Replacement: DSL_2015	7	0.3907
8	DSL_2009		9	0.3485
9	DSL_2012		7	0.3188
10	DSL_2011		6	0.2760
11	DSL_2008		8	0.2596
12	GWC_ADJ_2011		18	0.2542
13	GWC_ADJ_2015		16	0.2510
14	DSL_2013		3	0.1852
15	EERSTE_BOD		7	0.1644
16	GWC_ADJ_2014		8	0.1597
17	GWC_ADJ_2008		10	0.1323
18	GWC_ADJ_2009		5	0.0982
19	BOD_EENVOU		8	0.0853
20	REP_BTS_2015	Replacement: BTS_2015	5	0.0789
21	REP_BTS_2013	Replacement: BTS_2013	5	0.0782
22	REP_BTS_2014	Replacement: BTS_2014	8	0.0777
23	GWT		5	0.0701
24	LG_neers1		7	0.0663
25	EERSTE_GWT		5	0.0653
26	LG_temp		3	0.0601
27	REP_BTH_2015	Replacement: BTH_2015	3	0.0576
28	REP_BIO_2015	Replacement: BIO_2015	2	0.0351
29	REP_BTH_2014	Replacement: BTH_2014	1	0.0250

Table 33 Relative variable importance of Tree6 (variable with length of crop rotation schemas was not used)

Variable Importance				
Obs	NAME	LABEL	NRULES	IMPORTANCE
1	GWC_ADJ_2013		13	1.0000
2	GWC_ADJ_2010		21	0.5565
3	DSL_2010		7	0.4696
4	DSL_2014		6	0.4161
5	REP_DSL_2015	Replacement: DSL_2015	8	0.3958
6	GWC_ADJ_2012		25	0.3840
7	DSL_2009		9	0.3525
8	DSL_2012		7	0.3204
9	GWC_ADJ_2015		29	0.3180
10	GWC_ADJ_2014		15	0.2923
11	DSL_2011		8	0.2820
12	DSL_2008		12	0.2745
13	EERSTE_BOD		12	0.2238
14	DSL_2013		3	0.1856
15	GWC_ADJ_2009		16	0.1765
16	GWC_ADJ_2008		12	0.1547
17	GWC_ADJ_2011		13	0.1502
18	LG_temp		9	0.0908
19	BOD_EENVOU		8	0.0851
20	REP_BTS_2015	Replacement: BTS_2015	6	0.0750
21	GWT		5	0.0730
22	REP_BTS_2013	Replacement: BTS_2013	4	0.0702
23	REP_BTH_2015	Replacement: BTH_2015	4	0.0646
24	REP_BTS_2014	Replacement: BTS_2014	4	0.0645
25	EERSTE_GWT		5	0.0635
26	LG_neers1		6	0.0606
27	REP_BIO_2015	Replacement: BIO_2015	2	0.0481
28	REP_BTH_2014	Replacement: BTH_2014	3	0.0371
29	REP_BTH_2013	Replacement: BTH_2013	1	0.0186

Table 32 displays the variable importance when the label with crop rotation schema information was used, Table 33 displays the variable importance when no label was used. Table 32 shows that the label for crop rotation schemas (column name is LAB\_O\_CRIS) was relatively important (row/rank 4). Of all the input variables present, it was used the most to split up the input data (column NRULES = 35 times). The variable was also deemed relatively important considering the extent it has increased the purity of the end result.

The variable importance tables in Table 32 and Table 33 also show the relative importance of other variables. Next to the variable for crop rotation schemas and the variables containing the crop types (variable names that start with GWC\_ADJ) the variables containing the user identifier information were the most important (all variable names start with DSL). The user identifier variables were used 53 times to split up the input data into more homogeneous groups (sum of column NRULES). Also the variable containing detailed information on soil type was relatively important. The variable containing information on organic farming was relatively unimportant. It was only used twice to split the input data and did not add much purity to the end result. This also applies to the variables representing the main farm types and the variables on long-term average precipitation, long-term average temperature and groundwater level.



## 5 Discussion

In this chapter the results are discussed.

### 5.1 Method for identifying the length of crop rotation schemas

The search patterns to identify crop rotation schemas were suboptimal. The crop sequences displayed in Table 12 suggest a one-year crop rotation schema or even a four-year crop rotation schema instead of a two-year rotation schema. The applied selections caused this error. The selection schema for one-year rotation schemas required a minimum sequence of four equal crop codes or when missing crop codes occurred in the sequence, five equal crop codes in a six-year period (2011 till 2016). The crop rotations mentioned in Table 12 have two deviating crop type codes, which occur in 2011 and 2013. Therefore these sequences are not marked as one-year or four-year crop rotation schemas and might be misrepresented.

The crop sequences of the input data where no labels were present for rotation schemas show that there are still crop rotation schemas present in the data set. All of these crop sequences have in common that more than one crop type is missing or more than one crop type deviates in the rotation schema. By adding the length of the rotation schema for these crop sequences to the input data set, more crop types might be predicted correctly as the variable containing the length of crop rotation schemas slightly increases the prediction accuracy.

Another option is to use the actual crop sequence in the modeling process instead of the length of crop rotation schemas, like the research mentioned in Paragraph 1.2 suggests. For both input data sets crop sequences were created. If crop sequences would be used as a variable, it would contain much more distinct categorical values than the variable containing the length of the rotation schemas. Modeling techniques like multinomial logistic regression and neural networks cannot handle this many distinct categorical values. Even classification trees need a lot of computing memory and the extent of these values might cause an error when creating a model. A solution would be to split the research area into several smaller areas, in order to diminish the extent of the variable values and the memory required to compute a model. Another option is to aggregate the crop type codes in the crop sequences (i.e. aggregating all potato crop type codes to one).

Finally, no research was found from which can be deduced how many lasting or stable crop rotation schemas are present in the Netherlands. Due to reasons like economy (it might be more profitable to cultivate a different crop type than before), pest and weed problems, change of parcel ownership, etc. crop rotation schemas could change easily and are very unstable. This reinforced the notion that when this research should be performed in a next year with additional new declaration information, the variable information on crop rotation schemas that was produced in this research is outdated and probably no longer accurate. With every new addition of declaration information, the variable for the length of the crop rotation schemas needs to be determined anew.

### 5.2 Selection of the prediction method

Part of the research was to select the prediction method that correctly classified the most (sub) parcels. In this research classification trees and neural network models were built in order to find the most accurate one. Results showed that classification trees performed much better than the neural network models, see tables 13 till 18. Part of the reason for this bad performance of the neural network model could be the suboptimal configuration of the models' properties. The autoneural tool of SAS EM was used to create neural network models. For this modeling technique fewer parameters need to be adjusted and it is more automated (more properties are predefined and fixed) than the neural tool.

Another reason why the autoneural tool did not produce good predictive results could be the omission of the variables for user identifiers and the detailed soil type information. The variable importance of the classification trees (Table 32 and Table 33 on Page 55) showed that variable information on user

identifiers are important in the prediction of crop types. However, these variables have many distinct values and therefore require a lot of computer memory. No neural network models could be built when using these variables, therefore they were omitted. Not using these variables may have contributed to the worse prediction results.

Classification trees produced much better prediction results (lower misclassification rates). Two types of trees were built: trees where the input data set was split in two with every splitting round (maximum branch = 2) and trees where the input data set was split into multiple groups with every splitting round (maximum branch > 2). The reason for making two types of classification trees was that trees where a two-way split was applied, would be easier to interpret. However, the target variable was multinomial which made the interpretation of the prediction result already difficult, even with two-way splitting criteria.

When classification trees were build based on two-way splitting criteria, the property maximum depth was set at its maximum value (50). SAS EM does not allow trees to grow deeper than 50 levels. Therefore several variable data with unique user identifiers and the detailed soil type information were left out of the modeling process. This resulted in more correctly classified sub parcels. This is contrary to the assumption that user information was important in detecting a potential crop rotation schema or a switch in crop type cultivation on a parcel. The results of the variable importance tables showed that the user identifier information was important in predicting crop types, see Table 32 and Table 33. A reason for these lower misclassification rates might be that this modeling approach is greedy. When input data are split in more subgroups, the decision rule is used that produces the most purity in the newly created subgroups. The purity gain might be optimal for this particular split, however, it might not be the optimal choice when the purity gain of the whole tree is considered. This greedy approach can create suboptimal classification trees (Murthy et al. 1995). By not using the variables for user identifiers, other variables had to be chosen for splitting up the input data which resulted in lower misclassification rates.

The fact that the maximum value for the property maximum depth was used for two-way split classification trees, indicated that the study area might be too large to create an optimal two-way split classification tree. Multi-way split classification trees had no problems concerning maximum settings for properties, however, no research was found that multi-way split classification trees perform better than two-way split trees and the results showed that the prediction accuracy was only slightly better. To avoid using the maximum value for the property maximum depth of a tree, a solution might be to stratify the current AOI. If a series of smaller study areas would be used to predict crop types, there would probably be a smaller range of target values to predict than in the current AOI. There would be less distinction in soil type, groundwater level, ownership or usage of parcels and other matters that influence the choice of crops to cultivate. This was in line with the concept of Tobler's first law of Geography where *"everything is related to everything else, but near things are more related than distant things"* (Tobler 1970). By keeping the research area small, the more similar the objects are in that area. This might increase the prediction accuracy.

Furthermore, classification trees may have a tendency to overfit the training data (Bramer 2007). Even though 10-fold cross validation was applied to prevent the tree from overfitting the training data set, the ranking of the models according to the misclassification rate was different from the ranking of the models according to the prediction accuracy, see Tables 19 till 22. Random forests are less prone to overfit the training data (Ali et al. 2012; Liberman 2017). By creating multiple classification trees and averaging the prediction results for all the created trees, a less biased prediction result is produced. It is assumed that this modeling technique produces better prediction results than a single classification tree. Because the random forest technique is not available in RVO.nl's version of SAS EM, software package R may be used to investigate this assumption.

### **5.3 Prediction accuracy of crop types**

The prediction accuracy for the data set with labels for rotation schemas was much higher than the prediction accuracy of the data set where no labels for rotation schemas were present, see Tables 19 till 22. The comparison of the prediction results for both data sets was not completely fair, as for the input data set with crop rotation schemas the extra variable with the length of crop rotation schemas was used. Therefore a classification tree was built for the data set where labels for rotation schemas were present without the variable containing the length of rotation schemas, see Table 31. Result showed that even without the variable the prediction results for the data set containing the labels for rotation schemas were still much better than for the data set where no labels were present. This indicated that the data with rotation labels are already more homogeneous and therefore easier to classify correctly than the data set without the rotation labels.

The prediction results for rotation schemas with a length of five and six years were worse than for the other lengths. The prediction accuracy for rotation schemas with the length of five years was on average a little more than 33% and for rotation schemas with the length of six years, the prediction accuracy was almost 42%, see Table 23. These less accurate results may have been caused by the fact that the crop sequence was too short for proper identification of these rotation lengths.

Also overlap was present in the input data set and was a potential problem. A parcel can be claimed by multiple users, also called a double claim. This meant that when a parcel was declared by more than one user, it was also included more than once in the feature file containing the declaration information. Therefore overlap exists in the declaration files, see Appendix N. By dissolving the model's outcome based on prediction results (correct or incorrect) most of the overlap disappeared. Also the area of most of the overlap that was still present in the dissolved prediction result was very small when compared to the declared parcels surface. For only 3 declared parcels the overlap area was larger than 5%. Therefore the overlap was deemed of little consequence for the prediction accuracy of declared parcels.

Another thing to consider in this analysis, was the relatively low accuracy of prediction results in small sub parcels, see Table 40 in Appendix M. The difference in prediction accuracy between sub parcels that were smaller than 100 m<sup>2</sup> and sub parcels that were larger than 5 ha is 12,4 percentage point (63,9% versus 51,5%). It could be argued that some of the very small sub parcels are actually errors and therefore should be characterized as slivers. The actual size of a declared parcel where small sub parcels belong to, should be considered also. Small sub parcels belonging to a large declared parcel are probably less relevant and more likely an error than small sub parcels belonging to a small declared parcel.

Also parcels without a declaration history posed a problem. Every year new parcels were declared and also formerly declared parcels disappeared, which makes the declaration area dynamic. Consequently a small number of the declared parcels could not be classified. When the location of the declared parcels in 2017 was compared with the location of the declared parcels in 2016, 2,7% could not be classified for these parcels were not declared in 2016 and therefore contained no declaration history. On the other hand, 2,0% of the parcels would be classified that were not declared in 2017 anymore, see Appendix O.

#### **5.3.1 Contingency tables**

In the contingency table for the data set with rotation labels, potatoes are more often mixed up with wheat, sugar beets and maize than other crop types. A reason might be that potatoes often appear in a crop rotation schema with wheat, sugar beets and maize. When these crop types form sequences that differ in length, then confusion is very likely. This is also the case for maize. Maize is often mixed up with grass. These crop rotation schemas were also found in the former research. That potatoes, sugar beets, wheat and maize appear often in a crop rotation schema, is due to the fact that these crop types are the most profitable and can be cultivated in the same environment (Engwerda 2015).

However, the complete contingency tables showed that not all crop type codes were predicted. Most of these classes were very small in size and therefore represented only 4% of the sub parcels in the input data, see Table 26. However, more than half of the individual declared crop type codes present in the study area, were never predicted! Machine learning algorithms are known for predicting mostly the majority classes (Hu et al. 2012; SAS 2013). In classification trees a minority class will only be predicted when it represents the dominant group of target values in the end-node of a branch. Still, most of the minority classes will be scattered across several branches most of the time. Literature suggests the use of class weights or priors for predicting more minority classes (Hu et al. 2012; SAS 2013). By adding class weights or priors, the difference in importance of majority classes versus minority classes for prediction is adjusted. This might lead to predicting more minority classes and increased prediction accuracy.

Another classification problem originated from the manner in which the model was created. The predictive model was created based on declared crop types of 2016. In order to predict crop types for 2017, the declaration information had to shift a year, see Paragraph 3.2.6 for the applied approach. The prediction of the model was based on the target values (declared crop types of 2016). However, crop types that were declared in 2017 had not always been declared in 2016 and therefore did not exist in the target values that were used to create the model. The reverse situation also existed. The complete contingency tables showed that for the data set containing labels for crop rotation schemas, crop types 1001 and 2033 were predicted although these crop type codes were never declared in 2017. For the data set containing no labels for crop rotation schemas the crop types 2033, 2700, 345, 854 and 991 were predicted although these crop type codes were never declared in 2017.

### **5.3.2 Comparison of current prediction results with research performed in 2016**

When the results of this current research were compared with the results of the research performed in 2016, results showed a less accurate prediction result for this current research, see Table 30. In total 42,7% of the declared parcels were correctly classified for this current research versus 55,3% correctly classified parcels for the former research of 2016. However, these results cannot be interpreted in a manner that this current approach performed worse than the former research. This can be explained by the fact that in the former research aggregated crop codes were used. Table 27 shows why the prediction accuracy for less (or aggregated) crop types produces a higher prediction accuracy. For instance, mixing up potato types for the current research would have resulted in an incorrect prediction, however, for the research of 2016 the crop type would have been predicted correctly. Therefore aggregation of crop types leads to a higher prediction accuracy.

In order to conclude if this research would predict crop types better or worse than the former research, all the crop type codes would have to be aggregated according to the division of crop types in the former research. Also one prediction result would have to be produced for a declared parcel. However, when the variable for the length of crop rotation schemas would then be computed, probably a lot of monocultures (length of the rotation schema = 1 year) will be found, for there are only 11 different crop descriptions used in 2016. Therefore it is debatable if this variable will improve the accuracy of the prediction under such circumstances. To conclude whether this research performed better in terms of prediction accuracy remains unclear. However, this research performs better than the former research of 2016 in terms of applicability. Individual crop types were predicted instead of aggregated crop types, and that information may be used in the declaration process. The individual predicted crop types might also provide support in other monitoring processes like the remote sensing monitoring of the project Pilot Monitoring Farmland. The prediction result of 2016 would have been less useful for these purposes.

## **5.4 Variable importance**

When the variable information was used in the modeling process, 258 more observations were predicted correctly (Table 31). This represents a slight accuracy increase of almost 0,9 percentage point. Although the variable did not increase the prediction accuracy that much, from a relative point of view it was very important. It was the fourth best variable that created the most purity in the

subgroups produced during modeling, and is used most of all variables to split the data into smaller more homogeneous groups of target values (NRULES = 35), see Table 32.

However, more variables were used to model crop type prediction. Results showed that next to the variables containing the declared crop types and the variable with labels for crop rotation schemas, all variables containing user identifiers were relatively important in predicting crop types, see Table 32 and Table 33. These variables had in common that they had relative many distinct values at field level and were used more often to split up the input data than the variables with few different values. This might be explained by the fact that there were many target values to be predicted, and therefore a lot of distinction was necessary to isolate groups with similar target values. The phenomenon that highly detailed information is used more often in the splitting procedure of a tree than coarse data, is well-known (Strobl et al. 2007). It is important that there is a correlation present between the detailed variable and the target variable. In case of a weak correlation between the detailed variable and the target variable, a created model might fit the training data well and produce a low misclassification rate, however, for new input data the misclassification rate might be much higher which would result in a lower prediction accuracy.

Still, there were also variables that were relatively unimportant. Results show that the variables containing information on organic farming, main farm types, long-term average precipitation and temperature and groundwater level did not contribute very much to the purity of the model's outcome and were not used often to split up the input data. The low rank of the variable for organic farming could be explained by the small group this variable represents and that there was only one variable value (value = BIO). Only 5,6% of the input data was characterized as an organic business. The low rank for long-term average temperature and precipitation might be explained by the (geometric) coarseness of the data and the few values these variables possess. The level of distinction for crop types is much more detailed in comparison. Splitting the input data based on the variables containing the temperature and precipitation information would not increase the purity of the outcome result very much. The low ranking of the main farm types might be explained by the fact that also variables were used with more distinct (sub) farm types. These variables were used more often in the splitting procedure, making the variables with the main farm type more or less superfluous. The variable containing groundwater levels also did not produce much purity by splitting the input data. Although the information was detailed from a geometric point of view (the information is extracted from the same feature files containing the detailed soil information), it did not have a lot of distinctive values. The information is also partly outdated for it was created between 1960 and 1990 and should be replaced by more up-to-date data.

The prediction accuracy for the study area of this research was not very high, 42,7% of the declared parcels were classified correctly. This average result is not good enough to use in a pre-filled declaration. However, some of the crop types were predicted rather well. Winter wheat and rye both had prediction accuracies over 75% based on Tree39 and winter wheat had the highest prediction accuracy of Tree16 (no rye was predicted for Tree 16). As a first attempt to investigate the applicability of prediction results, crops with a prediction accuracy of 75% or higher could be used in a test on pre-filling a declaration.

## 6 Conclusion and recommendations

In this final chapter the research questions are answered and the main conclusion is drawn. Furthermore, recommendations are given how this research could be improved upon when it is performed for a new year to come.

### 6.1 **RQ I: What methods can be used to identify the length of crop rotation schemas and predict crop types?**

Search patterns were used to identify the length of crop rotation. By creating selections in ArcGIS that compared crop type codes from different declaration years based on a pattern, rotation schemas were found ranging from one year to six years. The length of the crop rotation schema depended on the time it took to declare the same crop type again for the same (sub) parcel. In total 29.622 rotation schemas were found in the input data set representing the study area. In order to identify crop rotation schemas, the declared crop types were first adjusted for the change in crop type codification. Due to greening measures crop type codes changed during the research period. Therefore crop type codes that changed during the research period were replaced by corresponding crop types used in the declaration year of 2017 in order to create a consistent crop sequence from which crop rotation schemas could be deduced.

To predict the crop types for 2017, data mining techniques were applied. Supervised modeling techniques were used, for the crop types of 2016 were available as target variable. As all input variables and the target variables were nominal values, the following modeling techniques were appropriate: multinomial logistic regression, the classification tree, neural network models, Bayesian network models, random forest and support vector machines (SVM). However, considering the extent of the range of target values, logistic regression and SVM were not an option. From a data mining software point of view Bayesian network models and random forest were no option. Although the latest version of SAS Enterprise Miner provides the option to build random forest models and Bayesian network models, the version RVO.nl uses cannot create these models. Therefore only classification trees and neural network models were built in order to find the modeling technique that predicts crop types the most accurate.

### 6.2 **RQ II: Which variables available at RVO can be used for predicting crop types?**

The following variables were used in this research:

- Crop type (declaration data from 2008 till 2017)
- Farm type (2013, 2014, 2015, 2016)
- Climate data
  - o Long-term average temperature (growing season)
  - o Long-term average precipitation (growing season)
- Soil type
- Organic agriculture (2015, 2016)
- Owner data/Usage data (declaration data from 2008 till 2016)
- Groundwater level

### 6.3 **RQ III: To what extent are variables available in historical data sets at RVO capable of predicting crop types?**

In order to predict crop types, classification tree models and neural network models were built. For both the input data set were label were present for crop rotation schemas and the input data set where no labels were present for crop rotation schemas, a multi-way split classification tree correctly classified the most sub parcels. Tree39 (data set where labels for the length of crop rotation schemas were present) correctly classified 45,78% of the sub parcels. Tree16 (data set where no labels for the length of crop rotation schemas were present) correctly classified 31,29% of the sub parcels.

When the predicted results were aggregated to one prediction result per declared parcel, 42,7% of the declared parcels were classified correctly. These percentages concern the investigated study area only and represent arable land. This prediction accuracy is not high enough to use in a pre-filled declaration. Yet, some individual crop type codes of which the prediction accuracy was higher than 75% like winter wheat or rye might be good enough to pre-fill a declaration.

The low prediction accuracy indicated that classes were mixed up during the prediction. For instance, potato crop types were often mixed up with each other. Potatoes were also more often mixed up with crop types that were part of the same crop rotation schemas. Also part of the declared crop types were never predicted for this study area. Over half of the crop types that were declared in 2017, were never predicted. The reason is that machine learning algorithms predict mostly the majority classes and minority classes are scattered over the multiple branches.

Still, the prediction of crop types was possible. Some input variables were more important in predicting crop types than others. The variables containing the declared crop types and the user identifiers were relative the most important variables in predicting crop types. These variables were responsible for creating most part of the purity (or homogeneous subgroups of target values) during the splitting process of the modeling. Variables with many distinct values and a detailed geometry were more frequently used than variables with less variable values and coarser geometry. An exception was the variable with a label for the length of crop rotation schemas. The conclusion is that crop types can be predicted, although the accuracy needs to be improved further before it can be used to pre-fill the declaration information.

#### **6.4 RQ IV: To what extent does information on the length of crop rotation improve predicting crop types in the reference year?**

The variable containing the length of crop rotation schemas was also fairly important in predicting crop types. For this study area, the variable was responsible for 258 more correctly predicted fields compared to when the variable was not used during the modeling process. By using this variable 0,9% more sub parcels were classified correctly. From a relative point of view this variable is very important. It is the fourth best variable that created the most purity among the target values of the created model, and was used most of all variables to split the input data into smaller, more homogeneous groups of target values (NRULES = 35).

#### **6.5 Main Conclusion**

This research aimed to identify the length of pertinent crop rotation based on historical data of farmers declarations and assess whether this length improves the prediction of the crop type cultivated in the next year.

By adjusting for the change in crop type codification and using search patterns to find matching crop type codes, crop rotation schemas were discovered and the length in years was deduced from this information. Among the 185.108 sub parcels were 29.622 sub parcels present that contained a crop rotation schema. The length of the rotation schema was used as a new variable in the modeling process. By using the new variable, 0,9% more crop types were classified correctly. The new variable was relative important during the creation of the prediction model. In the splitting process the new variable was used most of all variables present. Compared to most other variables it was also more responsible for increasing the prediction accuracy of the model. Therefore it is concluded that adding the variable with the length of crop rotation schemas has improved the prediction of crop types cultivated in the next year.

#### **6.6 Recommendations**

As the results indicate, it is possible to predict cultivated crop types in the next year. Although 42,7% accuracy does not seem a lot, the percentage only represent the arable farmland of the Dutch Agriculture. However, there is room for improvement.

As was mentioned in the discussion, more crop rotation schemas exist in the input data than were used in this research. The applied selections should be adjusted so that the extra crop rotation schemas that were not detected during this research, are added as a label to the variable information. Also needs to be checked if the proper length was assigned to the crop rotation schemas. As the results of the crop rotation schemas with a length of five and six years were worse than for other lengths, it should also be considered to expand the research period and see if the prediction results improve for these rotation schemas.

Another recommendation is to use the actual crop sequence instead of the crop rotation schema. This information is very detailed and describes the declared crop types better than the variable containing the length of crop rotation schemas. However, to avoid problems during the modeling process due to the increased amount of distinct categorical values, the study area could be stratified. By splitting up the study area in smaller regions for modeling, the prediction accuracy might improve. Another option would be to aggregate the crop types used in the crop sequence. When aggregated crop codes would be used to create crop sequences, the number of distinct crop sequences would diminish and possible memory problems could be avoided.

To improve the variable input, an additional approach to adjust for the change in crop type codification has been thought of. In this research only those crop types were adjusted for which several old crop types were combined in one new crop type. For the crop types that were split up into multiple new crop types, it could be investigated if the old code could be replaced by one of the crop type codes it changed in. By detecting the crop types that are cultivated in the vicinity of the concerning crop type, it could be deduced which of the new crop types is cultivated the most. This crop type could then be used to replace the old crop type. The vicinity could be based on a buffer around the parcel containing the old crop type code and what new crop type should replace the old value could come from the most recent declaration information. Considering the analysis that has to be performed for individual declared parcels, this is a large analysis when it is performed for the whole of the Netherlands. There are no results that indicate if this additional approach improves the prediction of crop types. Therefore it is recommended to research this approach for a small area first.

Another recommendation concerns the selection of the sub parcels. During the combining of the declaration information, slivers came into existence due to the difference in geometries of overlaying parcels. Results show that the prediction accuracy in the small sub parcels is considerably worse than in the larger sub parcels. Therefore it is recommended that the selection of slivers is checked again if it can be adjusted. The area of the small sub parcel in combination with the area of the declared parcel the sub parcel belongs to, might create more distinction if a sub parcel is a sliver or not. The ratio of both areas could be added to the selection of slivers.

Furthermore, the prediction results might also improve if another modeling technique was used. Literature research mentioned good results were produced when Bayesian network models were used to predict crop types. Literature research also shows that random forests are much less prone to overfitting the training data and might produce better prediction results. As these modeling techniques are not yet available at RVO.nl. the software package R can be used. This is free software for statistical computing (R Core team 2018).

It is also recommended to stratify the agricultural area of the Netherlands. By splitting up the area into several smaller regions, it is assumed that crop types are more similar in that region. This leads to a smaller subset of crop types to predict and might therefore increase the prediction accuracy.

A lot of minority classes were present in the study area. Therefore it is recommended that the modeling technique uses class weights or priors in order to predict more minority classes. A low class weight or prior makes it harder for a model to predict a minority class, while with a higher class weight



or prior minority classes are more likely to be predicted. Both SAS EM and R have the option to set class weights and priors for classification trees.

Even though there are already many variables present that describe the choice of crop type to cultivate, an additional one might also indicate why certain crop types are cultivated more than others. Within the department of RVO.nl product yield prices are available. It can be conceived that from a production point of view an agricultural entrepreneur would want to increase his or her income as much as possible. Therefore products will be cultivated that sell for a good price. This variable might indicate why certain crop types are favored over other crop types and should be added as an input variable.

Other information which might also be of interest in crop type prediction, is to know if an agricultural business is under contract of a client (i.e. potato processing factory, etc.) or a freelance business. Businesses which are under contract are assumed to produce the required product of the client and therefore might have more stable crop rotation schemas. Freelance businesses on the other hand produce according to the market demand and are assumed to have less stable crop rotation schemas.

And finally, it is recommended to research the minimum prediction accuracy a crop type must have in order to be used to pre-fill a declaration. A starting point might be a minimum prediction accuracy of 75% although this percentage is not substantiated. Trial and error will gain more insight into the actual minimum accuracy that is required in order to pre-fill a declaration.

## References

- Abbey, R., T. He, T. Wang and SAS Institute (2017). "Methods of Multinomial Classification Using Support Vector Machines." *IEEE Transactions on Neural Networks* 13(2): pp. 415-425.
- Agresti, A. (1996). *An introduction to categorical data analysis*. Hoboken, John Wiley & Sons. ISBN 9780471226185.
- Ali, J., R. Khan, N. Ahmad and I. Maqsood (2012). "Random forests and decision trees." *International Journal of Computer Science Issues (IJCSI)* 9(5): pp. 272.
- Bacall, A. (2012). *The boss wants me to create a computer algorithm that can convert hindsight into foresight*, Cartoonstock.
- Bachinger, J. and P. Zander (2007). "ROTOR, a tool for generating and evaluating crop rotations for organic farming systems." *European Journal of Agronomy* 26(2): pp. 130-143.
- Bambrick, N. (2016). "Support Vector Machines: A Simple Explanation." Retrieved 30-07-2018, 2018, from <https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html>.
- Bleker, H. (2010). "28 625 Herziening van het Gemeenschappelijk Landbouwbeleid." Retrieved 12-12-2017, 2017, from <https://zoek.officielebekendmakingen.nl/kst-28625-108.html>.
- Boerenbusiness (2016). "Wat is dé Nederlandse akkerbouwplaats?" Retrieved 24-01-2018, 2018, from <http://www.boerenbusiness.nl/aardappelen/artikel/10870399/wat-is-de-nederlandse-akkerbouwplaats>.
- Bouty, C., F. Levavasseur, P. Martin and A. Barbottin (2015). Are crop sequence evolutions influenced by farm territory dynamics? In: 5. International Symposium for Farming Systems Design, Montpellier. AGRO2015. pp.
- Bramer, M. (2007). *Avoiding overfitting of decision trees. Principles of data mining*. London, Springer: pp. 119-134.
- Coursera (2018). "What Is A Random Forest and How Is It "Grown"?" Retrieved 30-06-2018, 2018, from <https://www.coursera.org/learn/machine-learning-data-analysis/lecture/FL9wx/what-is-a-random-forest-and-how-is-it-grown>.
- De Ville, B. and P. Neville (2013). *Decision trees for analytics using SAS Enterprise Miner*. Cary, SAS Institute. ISBN 9781612902524.
- Engwerda, J. (2015). "Veel akkerbouwgewassen brengen meer geld in kas." Retrieved 30-07-2018, 2018, from <https://www.boerderij.nl/Akkerbouw/Achtergrond/2015/10/Veel-akkerbouwgewassen-brengen-meer-geld-in-kas-2705405W/>.
- EUR-lex (2008). "COMMISSION REGULATION (EC) No 1242/2008 of 8 December 2008 establishing a Community typology for agricultural holdings." Retrieved 08-03-2018, 2018, from <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32008R1242>.
- EuropaNu (2017). "Landbouwbeleid (GLB)." Retrieved 12-12-2017, 2017, from [https://www.europa-nu.nl/id/vg9pir5eze8o/landbouwbeleid\\_glb](https://www.europa-nu.nl/id/vg9pir5eze8o/landbouwbeleid_glb).

- European Commission (2017). "Greening." Retrieved 08-03-2018, 2018, from [https://ec.europa.eu/agriculture/direct-support/greening\\_en](https://ec.europa.eu/agriculture/direct-support/greening_en).
- Feelders, A. (1999). Handling missing data in trees: surrogate splits or statistical imputation? In: European Conference on Principles of Data Mining and Knowledge Discovery, Prague. Berlin, Springer. pp. 329-334.
- Grise, J. (2013). Effect of Climate Change on Farmers' Choice of Crops: An Econometric Analysis. MSc thesis, University of Saskatchewan.
- Hall, P., J. Dean, I. K. Kabul and J. Silva (2014). An overview of machine learning with SAS® enterprise miner™. SAS Global Forum 2014. Washington, SAS Institute.
- Han, J., J. Pei and M. Kamber (2011). Data mining: concepts and techniques. Waltham, Elsevier. ISBN 9780123814791.
- Howden, S. M., J.-F. Soussana, F. N. Tubiello, N. Chhetri, M. Dunlop and H. Meinke (2007). "Adapting agriculture to climate change." Proceedings of the national academy of sciences 104(50): pp. 19691-19696.
- Hu, Y.-J., T.-H. Ku, R.-H. Jan, K. Wang, Y.-C. Tseng and S.-F. Yang (2012). "Decision tree-based learning to predict patient controlled analgesia consumption and readjustment." BMC medical informatics and decision making 12(1): pp. 131.
- IRS (2017). "Vruchtwisseling - teelthandleiding." Retrieved 12-12-2017, 2017, from <https://www.irs.nl/ziekten-en-plagen/teelthandleiding/5.3-vruchtwisseling>.
- Jacobsen, C., U. Zscherpel and P. Perner (1999). A comparison between neural networks and decision trees. Machine Learning and Data Mining in Pattern Recognition. P. Perner and M. Petrou. Berlin Heidelberg, Springer: pp. 144-158.
- Janssen, L. L. F. (1994). Methodology for updating terrain object data from remote sensing data. The application of Landsat TM data with respect to agricultural fields. Wageningen, Landbouwniversiteit Wageningen. ISBN 9789054851813.
- Kabat, P., R. Hutjes, B. Kruijt, G. Nabuurs, L. Higler, P. van der Meer, M. Schelhaas, E. van Ierland, A. Schapendonk and A. Verhagen (2000). Effecten van klimaatverandering op de vitaliteit van de functies in het landelijk gebied. Quicksan LNV-agenda klimaat: pp. 17-62.
- KNMI (2015a). "KNMI'14-klimaatsscenario's - Kaarten, grafieken en tabellen." Retrieved 18-03-2018, 2018, from <http://www.klimaatsscenarios.nl/getallen/overzicht.php?wel=neerslag&ws=kaart&wom=gemiddelde%20neerslag>.
- KNMI (2015b). "KNMI'14-klimaatsscenario's - Regionale verschillen." Retrieved 08-03-2018, 2018, from [http://www.klimaatsscenarios.nl/faq\\_klimaatsscenarios/regionale\\_verschillen.html](http://www.klimaatsscenarios.nl/faq_klimaatsscenarios/regionale_verschillen.html).
- Le Ber, F., M. Benoît, C. Schott, J.-F. Mari and C. Mignolet (2006). "Studying crop sequences with CarrotAge, a HMM-based data mining software." Ecological modelling 191(1): pp. 170-185.
- Leteinturier, B., J. Herman, F. De Longueville, L. Quintin and R. Oger (2006). "Adaptation of a crop sequence indicator based on a land parcel management system." Agriculture, Ecosystems & Environment 112(4): pp. 324-334.

- Levavasseur, F., P. Martin, C. Bouty, A. Barbottin, V. Bretagnolle, O. Thérond, O. Scheurer and N. Piskiewicz (2016). "RPG Explorer: A new tool to ease the analysis of agricultural landscape dynamics with the Land Parcel Identification System." *Computers and Electronics in Agriculture* 127: pp. 541-552.
- Liberman, N. (2017). "Decision Trees and Random Forests." Retrieved 30-07-2018, 2018, from <https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991>.
- Lobell, D. B., W. Schlenker and J. Costa-Roberts (2011). "Climate trends and global crop production since 1980." *Science* 333(6042): pp. 616-620.
- Mari, J.-F. and M. Benoît (2010). "Landscape regularity modelling for environmental challenges in agriculture." *Landscape ecology* 25(2): pp. 169-183.
- MathWorks (2018). "How many classes can a multi-class svm can classify?" Retrieved 30-06-2018, 2018, from <https://www.mathworks.com/matlabcentral/answers/388696-how-many-classes-can-a-multi-class-svm-can-classify>.
- Matignon, R. (2007). *Data mining using SAS enterprise miner*. Hoboken, John Wiley & Sons. ISBN 9780470149010.
- Murthy, S. K. and S. Salzberg (1995). *Decision Tree Induction: How Effective Is the Greedy Heuristic?* In: *The First International Conference on Knowledge Discovery and Data Mining*, Montreal. AAAI Press. pp. 222-227.
- Nakos, B. (2001). 'On the Assessment of Manual Line Simplification Based on Sliver Polygon Shape-Analysis'. 4th Workshop on Progress in Automated Map Generalisation Beijing, ICA.
- Osman, J., J. Inglada and J.-F. Dejoux (2015). "Assessment of a Markov logic model of crop rotations for early crop mapping." *Computers and Electronics in Agriculture* 113: pp. 234-243.
- Prairie farmer (2017). "Markets respond: From too much rain to not enough." Retrieved 30-06-2018, 2018, from <http://www.prairiefarmer.com/marketing/markets-respond-too-much-rain-not-enough>.
- Quan, M. and J. Liang (2017). "The influences of four types of soil on the growth, physiological and biochemical characteristics of *Lycoris aurea* (L'Her.) Herb." *Scientific Reports* 7(43284).
- R Core team (2018). "The R Project for Statistical Computing." Retrieved 01-08-2018, 2018, from <https://www.r-project.org/>.
- Refaeilzadeh, P., L. Tang and H. Liu (2009). *Cross-validation*. *Encyclopedia of database systems*. L. Liu and M. T. Özsu. Boston, Springer: pp. 532-538.
- ResearchGate (2014). "What is the most proper classifier for for big class numbers (30 or 55 classes)?" Retrieved 30-06-2018, 2018, from [https://www.researchgate.net/post/What\\_is\\_the\\_most\\_proper\\_classifier\\_for\\_for\\_big\\_class\\_numbers\\_30\\_or\\_55\\_classes](https://www.researchgate.net/post/What_is_the_most_proper_classifier_for_for_big_class_numbers_30_or_55_classes).
- ResearchGate (2016). "Is there an ideal ratio between a training set and validation set? Which trade-off would you suggest?" Retrieved 30-06-2018, 2018, from

[https://www.researchgate.net/post/Is\\_there\\_an\\_ideal\\_ratio\\_between\\_a\\_training\\_set\\_and\\_validation\\_set\\_Which\\_trade-off\\_would\\_you\\_suggest](https://www.researchgate.net/post/Is_there_an_ideal_ratio_between_a_training_set_and_validation_set_Which_trade-off_would_you_suggest).

Řezanková, H. and B. Everitt (2009). "Cluster analysis and categorical data." *Statistika 3*: pp. 216-232.

RVO.nl (2016). *Data inwinnen, hoezo?* Internal report.

RVO.nl (2017a). "Gecombineerde opgave." Retrieved 12-12-2017, 2017, from <https://mijn.rvo.nl/gecombineerde-opgave>.

RVO.nl (2017b). "Gemeenschappelijk Landbouwbeleid - GLB." Retrieved 12-12-2017, 2017, from <https://www.rvo.nl/onderwerpen/agrarisch-ondernemen/gemeenschappelijk-landbouwbeleid/gemeenschappelijk-landbouwbeleid>.

SARE (2012). "Crop Rotation and Farm Management." Retrieved 30-06-2018, 2018, from <https://www.sare.org/Learning-Center/Books/Crop-Rotation-on-Organic-Farms/Text-Version/How-Expert-Organic-Farmers-Manage-Crop-Rotations/Crop-Rotation-and-Farm-Management>.

SAS (2011), December 2011. "Getting Started with SAS® Enterprise Miner™ 7.1." Retrieved 20-12-2017, 2017, from <https://support.sas.com/documentation/cdl/en/emgsj/64144/PDF/default/emgsj.pdf>.

SAS (2013). "Using priors and decision weights in SAS® Enterprise Miner(tm)." Retrieved 01-08-2018, 2018, from <http://support.sas.com/kb/47/965.html>.

SAS (2017). "Working with Support Vector Machines." Retrieved 08-03-2018, 2018, from <http://documentation.sas.com/?docsetId=vaobjdmml&docsetTarget=n19iopwz2oo4izn1j9x7h6ydvopa.htm&docsetVersion=8.1&locale=en>.

SAS Communities Library (2017). "Getting the Most from your Random Forest." Retrieved 01-08-2018, from <https://communities.sas.com/t5/SAS-Communities-Library/Tip-Getting-the-Most-from-your-Random-Forest/ta-p/223949>.

Schönhart, M., E. Schmid and U. A. Schneider (2011). "CropRota—A crop rotation model to support integrated land use assessments." *European Journal of Agronomy* 34(4): pp. 263-277.

Seo, S. N. and R. Mendelsohn (2008). "An analysis of crop choice: Adapting to climate change in South American farms." *Ecological economics* 67(1): pp. 109-116.

Southern States (2018). "Crop production challenges in a high rainfall year." Retrieved 30-06-2018, 2018, from <https://www.southernstates.com/articles/crop-challenges-high-rainfall.aspx>.

Stack Overflow (2012). "Decision trees vs. Neural Networks." Retrieved 28-12-2017, 2017, from <https://softwareengineering.stackexchange.com/questions/157324/decision-trees-vs-neural-networks>.

StatQuest (2018). "StatQuest: Random Forests Part 1 - Building, Using and Evaluating." Retrieved 30-06-2018, 2018, from [https://www.youtube.com/watch?v=J4Wdy0Wc\\_xQ](https://www.youtube.com/watch?v=J4Wdy0Wc_xQ).

Steinmann, H.-H. and E. S. Dobers (2013). "Spatio-temporal analysis of crop rotations and crop sequence patterns in Northern Germany: potential implications on plant health and crop protection." *Journal of Plant Diseases and Protection* 120(2): pp. 85-94.

- Strobl, C., A.-L. Boulesteix, A. Zeileis and T. Hothorn (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution." *BMC bioinformatics* 8(1): pp. 25.
- Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region." *Economic geography* 46(sup1): pp. 234-240.
- Upton, G. J. (2016). *Categorical Data Analysis by Example*. Hoboken, John Wiley & Sons. ISBN 9781119307860.
- van Everdingen, W. and A. Wisman (2016). *NSO-typering 2016; Normen en uitgangspunten bij typering van agrarische bedrijven in Nederland*. Wageningen, Wageningen Economic Research.
- Wageningen University & Research (2018). "Methodiek Grondwaterdynamiek." Retrieved 30-07-2018, 2018, from <https://www.wur.nl/nl/Onderzoek-Resultaten/Onderzoeksinstituten/Environmental-Research/Faciliteiten-Producten/Software-en-modellen/Grondwaterdynamiek/Methodiek.htm>.
- Wang, J., R. Mendelsohn, A. Dinar and J. Huang (2010). "How Chinese farmers change crop choice to adapt to climate change." *Climate Change Economics* 1(03): pp. 167-185.
- Wang, J., K. Price and P. Rich (2001). "Spatial patterns of NDVI in response to precipitation and temperature in the central Great Plains." *International journal of remote sensing* 22(18): pp. 3827-3844.
- Wesseling, J. (1978). *De gevolgen van het verlagen van de grondwaterstanden voor landbouw en natuurgebieden*. Wageningen, Instituut voor Cultuurtechniek en Waterhuishouding.
- Wijnands, F. G. (2000). "Vruchtwisseling basis voor kwaliteits productie in biologisch bedrijf " PAV Bulletin Vollegrondsgroenteteelt.
- Xiao, Y., C. Mignolet, J.-F. Mari and M. Benoît (2014). "Modeling the spatial distribution of crop sequences at a large regional scale using land-cover survey data: A case from France." *Computers and Electronics in Agriculture* 102: pp. 51-63.
- Zhao, K. (2015). *Predictive Modeling Using Artificial Neural Networks in SAS® Enterprise Miner*. MWSUG 2015. Omaha, MidWest SAS Users Group.

## Appendix A – Selections to create label crop rotation schema

Crop rotation schema	Selection criteria (consistent)	Selection criteria (inconsistent)
1-year	<p>GWC_ADJ_2016 = GWC_ADJ_2015 AND  GWC_ADJ_2016 = GWC_ADJ_2014 AND  GWC_ADJ_2016 = GWC_ADJ_2013</p>	<p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2015= GWC_ADJ_2014 AND  GWC_ADJ_2015 = GWC_ADJ_2013 AND  GWC_ADJ_2015= GWC_ADJ_2012 AND  GWC_ADJ_2015 = GWC_ADJ_2011) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016= GWC_ADJ_2014 AND  GWC_ADJ_2016 = GWC_ADJ_2013 AND  GWC_ADJ_2016= GWC_ADJ_2012 AND  GWC_ADJ_2016 = GWC_ADJ_2011) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016= GWC_ADJ_2015 AND  GWC_ADJ_2016 = GWC_ADJ_2013 AND  GWC_ADJ_2016= GWC_ADJ_2012 AND  GWC_ADJ_2016 = GWC_ADJ_2011) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016= GWC_ADJ_2015 AND  GWC_ADJ_2016 = GWC_ADJ_2014 AND  GWC_ADJ_2016= GWC_ADJ_2012 AND  GWC_ADJ_2016 = GWC_ADJ_2011)</p>
2-year	<p>LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2014 AND  GWC_ADJ_2015 = GWC_ADJ_2013 AND  GWC_ADJ_2014 = GWC_ADJ_2012 AND  GWC_ADJ_2013 = GWC_ADJ_2011</p>	<p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2014 AND  GWC_ADJ_2014 = GWC_ADJ_2012 AND  GWC_ADJ_2013 = GWC_ADJ_2011) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2015 = GWC_ADJ_2013 AND  GWC_ADJ_2014 = GWC_ADJ_2012 AND  GWC_ADJ_2013 = GWC_ADJ_2011)</p>
3-year	<p>LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2013 AND  GWC_ADJ_2015 = GWC_ADJ_2012 AND  GWC_ADJ_2014 = GWC_ADJ_2011 AND  GWC_ADJ_2013 = GWC_ADJ_2010</p>	<p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2013 AND  GWC_ADJ_2015 = GWC_ADJ_2012 AND  GWC_ADJ_2014 = GWC_ADJ_2011 AND  GWC_ADJ_2012 = GWC_ADJ_2009) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2013 AND  GWC_ADJ_2015 = GWC_ADJ_2012 AND  GWC_ADJ_2013 = GWC_ADJ_2010 AND  GWC_ADJ_2012 = GWC_ADJ_2009) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2013 AND  GWC_ADJ_2014 = GWC_ADJ_2011 AND  GWC_ADJ_2013 = GWC_ADJ_2010 AND  GWC_ADJ_2012 = GWC_ADJ_2009) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2015 = GWC_ADJ_2012 AND  GWC_ADJ_2014 = GWC_ADJ_2011 AND  GWC_ADJ_2013 = GWC_ADJ_2010 AND  GWC_ADJ_2012 = GWC_ADJ_2009)</p>

4-year	<p>LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2012 AND  GWC_ADJ_2015 = GWC_ADJ_2011 AND  GWC_ADJ_2014 = GWC_ADJ_2010 AND  GWC_ADJ_2013 = GWC_ADJ_2009</p>	<p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2012 AND  GWC_ADJ_2015 = GWC_ADJ_2011 AND  GWC_ADJ_2014 = GWC_ADJ_2010 AND  GWC_ADJ_2013 = GWC_ADJ_2009) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2012 AND  GWC_ADJ_2015 = GWC_ADJ_2011 AND  GWC_ADJ_2014 = GWC_ADJ_2010 AND  GWC_ADJ_2012 = GWC_ADJ_2008) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2012 AND  GWC_ADJ_2015 = GWC_ADJ_2011 AND  GWC_ADJ_2013 = GWC_ADJ_2009 AND  GWC_ADJ_2012 = GWC_ADJ_2008) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2012 AND  GWC_ADJ_2014 = GWC_ADJ_2010 AND  GWC_ADJ_2013 = GWC_ADJ_2009 AND  GWC_ADJ_2012 = GWC_ADJ_2008) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2015 = GWC_ADJ_2011 AND  GWC_ADJ_2014 = GWC_ADJ_2010 AND  GWC_ADJ_2013 = GWC_ADJ_2009 AND  GWC_ADJ_2012 = GWC_ADJ_2008)</p>
5-year	<p>LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2011 AND  GWC_ADJ_2015 = GWC_ADJ_2010 AND  GWC_ADJ_2014 = GWC_ADJ_2009 AND  GWC_ADJ_2013 = GWC_ADJ_2008</p>	<p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2015 = GWC_ADJ_2010 AND  GWC_ADJ_2014 = GWC_ADJ_2009 AND  GWC_ADJ_2013 = GWC_ADJ_2008) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2011 AND  GWC_ADJ_2014 = GWC_ADJ_2009 AND  GWC_ADJ_2013 = GWC_ADJ_2008) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2011 AND  GWC_ADJ_2015 = GWC_ADJ_2010 AND  GWC_ADJ_2013 = GWC_ADJ_2008) OR</p> <p>(LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2011 AND  GWC_ADJ_2015 = GWC_ADJ_2010 AND  GWC_ADJ_2014 = GWC_ADJ_2009)</p>
6-year	<p>LAB_O_CRS IS NULL AND  GWC_ADJ_2016 = GWC_ADJ_2010 AND  GWC_ADJ_2015 = GWC_ADJ_2009 AND  GWC_ADJ_2014 = GWC_ADJ_2008</p>	



## Appendix B – Example of selections to replace missing data crop types

A model was created in ArcGIS to replace missing values of the variables containing the crop types. An example of a selection to replace missing values for variables containing the crop types of 2008 and 2009 is displayed here.

Example of selection criteria for imputing missing values in the declaration years 2008 and 2009				
Selection		Fieldname:		Fieldname:
GWC08_ADJ2 IS NULL AND LAB_O_CRS = 6	copy crop types of	[GWC_ADJ_2014]	copied into	GWC08_ADJ2
GWC08_ADJ2 IS NULL AND LAB_O_CRS = 5	copy crop types of	[GWC_ADJ_2013]	copied into	GWC08_ADJ2
GWC08_ADJ2 IS NULL AND LAB_O_CRS = 4	copy crop types of	[GWC_ADJ_2012]	copied into	GWC08_ADJ2
GWC08_ADJ2 IS NULL AND LAB_O_CRS = 3	copy crop types of	[GWC_ADJ_2011]	copied into	GWC08_ADJ2
GWC08_ADJ2 IS NULL AND LAB_O_CRS = 2	copy crop types of	[GWC_ADJ_2010]	copied into	GWC08_ADJ2
GWC08_ADJ2 IS NULL AND LAB_O_CRS = 2	copy crop types of	[GWC_ADJ_2012]	copied into	GWC08_ADJ2
GWC08_ADJ2 IS NULL AND LAB_O_CRS = 1	copy crop types of	[GWC_ADJ_2009]	copied into	GWC08_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 6	copy crop types of	[GWC_ADJ_2015]	copied into	GWC09_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 5	copy crop types of	[GWC_ADJ_2014]	copied into	GWC09_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 4	copy crop types of	[GWC_ADJ_2013]	copied into	GWC09_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 3	copy crop types of	[GWC_ADJ_2012]	copied into	GWC09_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 2	copy crop types of	[GWC_ADJ_2011]	copied into	GWC09_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 2	copy crop types of	[GWC_ADJ_2013]	copied into	GWC09_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 1	copy crop types of	[GWC_ADJ_2010]	copied into	GWC09_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 1	copy crop types of	[GWC_ADJ_2008]	copied into	GWC09_ADJ2
GWC09_ADJ2 IS NULL AND LAB_O_CRS = 2	copy crop types of	[GWC_ADJ_2015]	copied into	GWC09_ADJ2
The missing values in other declaration years have been imputed in the same manner, however are not mentioned here. Only an example is provided.				

Model name: 09\_PMF\_PREP\_IMPUTE\_MISSING\_DATA\_GWC

# Appendix C - Example of a SAS Enterprise Miner diagram

The screenshot displays the SAS Enterprise Miner interface. The main workspace shows a workflow diagram starting with a 'PMF AFST 2017 - LAB - MAX BR > 2 - Klaar' data source. This source feeds into 'StatExplore', 'Replace', and 'Data Partition' nodes. The 'Data Partition' node branches into 12 'Decision Tree' nodes. These nodes feed into 'Score' nodes (score (4), score (3), score (2)), which then feed into 'SAS Code (2)' nodes. A 'Control Point' node also feeds into a 'Model Comparison' node, which finally feeds into 'SAS Code (2)' nodes.

The bottom-left panel shows the 'Properties' window for a 'Decision Tree' node (Node ID: Tree39). The 'General' tab is active, showing the following properties:

Property	Value
Node ID	Tree39
Imported Data	...
Exported Data	...
Nodes	...
<b>Train</b>	
Variables	...
Interactive	...
Use Frozen Tree	No
Use Multiple Targets	No
Precision	4
<b>Splitting Rule</b>	
Interval Criterion	ProbF
Nominal Criterion	Sini
Ordinal Criterion	Entropy
Significance Level	0.2
Missing Values	Most correlated branch
Use Input Once	No
Maximum Branch	8
Maximum Depth	11
Minimum Categorical Size	5
Split Precision	4
<b>Nodes</b>	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
<b>Split Search</b>	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
<b>Subtree</b>	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification

The bottom-right panel shows the 'Nominal Criterion' section with the following text: 'Specifies the method of searching for and evaluating candidate splitting rules in the presence of a nominal target.'

## Appendix D – Class variable summary statistics

Class Variable Summary Statistics  
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	B10_2015	INPUT	3	28005		94.54	01	4.81
TRAIN	B0D_EENV0U	INPUT	16	0	Zeekeigronden	46.80	Humuspodzolen	11.84
TRAIN	BTH_2013	INPUT	9	244	Akkerbouwbedrijven	52.47	Graasdierbedrijven	27.43
TRAIN	BTH_2014	INPUT	9	1588	Akkerbouwbedrijven	51.96	Graasdierbedrijven	25.27
TRAIN	BTH_2015	INPUT	10	567	Akkerbouwbedrijven	51.83	Graasdierbedrijven	26.13
TRAIN	BTS_2013	INPUT	35	244	Overige akkerbouwbedrijven	28.31	Melkveebedrijven	20.14
TRAIN	BTS_2014	INPUT	37	1588	Overige akkerbouwbedrijven	27.35	Melkveebedrijven	18.03
TRAIN	BTS_2015	INPUT	39	567	Overige akkerbouwbedrijven	31.35	Melkveebedrijven	20.23
TRAIN	DSL_2008	INPUT	513	0	201803852	3.95	50013708	3.65
TRAIN	DSL_2009	INPUT	513	0	50013708	4.03	200359859	3.46
TRAIN	DSL_2010	INPUT	513	0	200104815	5.04	201803852	3.77
TRAIN	DSL_2011	INPUT	513	0	50013708	3.91	201803852	2.69
TRAIN	DSL_2012	INPUT	513	0	50013708	4.13	201459453	2.65
TRAIN	DSL_2013	INPUT	513	0	50013708	4.14	201803852	3.94
TRAIN	DSL_2014	INPUT	513	0	200104815	3.01	204625275	2.57
TRAIN	DSL_2015	INPUT	513	2	050013708	3.43	200359859	2.79
TRAIN	EERSTE_B0D	INPUT	185	0	Mn45A	10.06	Mn35A	8.94
TRAIN	EERSTE_GWT	INPUT	17	0	VI	37.78	IV	15.59
TRAIN	GWC_ADJ_2008	INPUT	65	1327	233	22.91	259	12.51
TRAIN	GWC_ADJ_2009	INPUT	61	707	233	22.23	259	12.97
TRAIN	GWC_ADJ_2010	INPUT	61	533	233	26.17	259	13.97
TRAIN	GWC_ADJ_2011	INPUT	62	505	233	23.57	259	16.00
TRAIN	GWC_ADJ_2012	INPUT	57	253	233	25.62	259	17.31
TRAIN	GWC_ADJ_2013	INPUT	56	146	233	24.25	259	18.30
TRAIN	GWC_ADJ_2014	INPUT	51	75	233	24.01	259	20.17
TRAIN	GWC_ADJ_2015	INPUT	86	51	233	24.88	259	18.63
TRAIN	GWT	INPUT	15	0	VI	33.90	-	21.84
TRAIN	LAE_0_CRS	INPUT	6	0	I	39.93	4	17.53
TRAIN	LG_neersl	INPUT	4	0	150 - 175	50.95	175 - 200	37.58
TRAIN	LG_temp	INPUT	5	0	9,0 - 9,3	40.91	8,4 - 8,7	17.69
TRAIN	GWC_ADJ_2016	TARGET	103	129	233	25.88	259	19.81

## Appendix E - Tables with misclassification rates for classification tree models

Table 34 Two-way split tree misclassification rates of data models where labels for crop rotation schemas are present

		Data set Classification tree: labels for crop rotation schemas are present, maximum branch = 2												
		Misclassification rate		Variable usage							Parameter settings			
Nr.	Tree name	Train	Validate	DSL 2008	DSL 2009	DSL 2010	DSL 2011	DSL 2012	DSL 2013	EERSTE BOD	Nominal criterion	Maximum branch	Maximum depth	Minimum categorical size
1	Tree15	0,1128277	0,1493995	No	No	No	No	No	Yes	No	Gini	2	50	5
2	Tree24	0,1098827	0,149624	No	No	No	Yes	Yes	Yes	No	Gini	2	50	3
3	Tree3	0,1146623	0,1511954	No	No	No	No	No	No	No	Gini	2	50	5
	Tree22	0,1138898	0,1520934	No	No	No	Yes	Yes	Yes	No	Gini	2	50	4
	Tree12	0,1179935	0,1520934	No	No	No	Yes	Yes	Yes	No	Gini	2	50	5
	Tree14	0,1153382	0,1531036	No	No	No	No	Yes	Yes	No	Gini	2	50	5
	Tree17	0,1157244	0,1537771	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	50	5
	Tree7	0,1164969	0,1542261	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	45	5
	Tree11	0,1188626	0,1552363	No	No	Yes	Yes	Yes	Yes	No	Gini	2	50	5
	Tree13	0,1176073	0,1559098	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	40	5
	Tree9	0,118766	0,156022	No	No	No	Yes	Yes	Yes	Yes	Gini	2	50	5
	Tree19	0,1210834	0,1572567	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	35	5
	Tree10	0,1194902	0,1574812	No	No	Yes	Yes	Yes	Yes	Yes	Gini	2	50	5
	Tree5	0,1167383	0,1591649	No	No	No	No	No	Yes	No	Entropy	2	50	5
	Tree25	0,1286632	0,1592771	No	No	No	Yes	Yes	Yes	No	Gini	2	50	6
	Tree2	0,1168348	0,1602873	No	No	No	No	Yes	Yes	No	Entropy	2	50	5
	Tree8	0,1260078	0,1630935	No	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	50	5
	Tree26	0,1379327	0,1682568	No	No	No	Yes	Yes	Yes	No	Gini	2	50	7
	Tree18	0,1318496	0,1694915	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	30	5
	Tree4	0,1307874	0,1713997	No	No	No	No	Yes	Yes	No	Gini	2	50	5
	Tree20	0,1430985	0,1767875	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	25	5
	Tree	0,1651137	0,1966551	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	20	5

Gini and entropy were compared by making identical models where only the nominal criterion property was different (Gini or entropy). Results can be found in Table 34. Tree 15 and tree 5 have identical property settings, except for the nominal criterion. Tree 14 and tree 2 also have identical property settings except for the nominal criterion. These results showed that Gini performed slightly better than entropy, however the difference of the misclassification rate of the models was small, only 1% for tree 15 and 5 and 0,7% for tree 14 and tree 2.

Table 35 Two-way split tree misclassification rates of data models where no labels for crop rotation schemas are present

		Data set Classification tree: labels for crop rotation schemas are not present, maximum branch = 2												
		Misclassification rate		Variable usage							Parameter settings			
Nr.	Tree name	Train	Validate	DSL 2008	DSL 2009	DSL 2010	DSL 2011	DSL 2012	DSL 2013	EERSTE BOD	Nominal criterion	Maximum branch	Maximum depth	Minimum categorical size
1	Tree25	0,1612189	0,24853	No	No	Yes	Yes	Yes	Yes	Yes	Gini	2	50	3
2	Tree20	0,1618148	0,254214	No	No	No	Yes	Yes	Yes	Yes	Gini	2	50	3
3	Tree15	0,1885427	0,2687181	No	No	Yes	Yes	Yes	Yes	Yes	Gini	2	50	5
	Tree21	0,182414	0,2687181	No	No	Yes	Yes	Yes	Yes	Yes	Gini	2	50	4
	Tree17	0,1916071	0,2700902	No	No	No	Yes	Yes	Yes	Yes	Gini	2	50	5
	Tree16	0,1907559	0,2716582	No	No	Yes	Yes	Yes	Yes	No	Gini	2	50	5
	Tree24	0,190841	0,2718542	No	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	50	5
	Tree18	0,1988424	0,2749902	No	No	No	No	Yes	Yes	Yes	Gini	2	50	5
	Tree19	0,1927137	0,2767542	No	No	No	Yes	Yes	Yes	No	Gini	2	50	5
	Tree8	0,1927988	0,2773422	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	50	5
	Tree6	0,1942458	0,2775382	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	45	5
	Tree4	0,2031835	0,2828303	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	40	5
	Tree5	0,2055669	0,2869463	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	35	5
	Tree22	0,2099932	0,2934143	No	No	Yes	Yes	Yes	Yes	Yes	Gini	2	50	6
	Tree23	0,2299115	0,3094865	No	No	Yes	Yes	Yes	Yes	Yes	Gini	2	50	7
	Tree3	0,2544263	0,3245786	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	30	5
	Tree2	0,2865169	0,3453548	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	25	5
	Tree	0,3340994	0,3874951	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Gini	2	20	5

Table 36 Multi-way split tree misclassification rates of data models where labels for crop rotation schemas are present

				Data set Classification tree: labels for crop rotation schemas are present, maximum branch > 2				
		Misclassification rate		Variable usage	Parameter settings			
Nr.	Tree name	Train	Validate	LAB_0_CRS	Nominal criterion	Maximum branch	Maximum depth	Minimum categorical size
1	Tree37	0,0681829	0,1420198	Yes	Gini	7	11	1
2	Tree38	0,0929635	0,1457394	Yes	Gini	7	11	2
3	Tree35	0,1119732	0,1532913	Yes	Gini	7	11	3
	Tree39	0,1254061	0,159716	Yes	Gini	8	11	5
	Tree28	0,1248242	0,1630974	Yes	Gini	7	11	5
	Tree25	0,1258911	0,1643372	Yes	Gini	6	11	5
	Tree36	0,1235634	0,1649008	Yes	Gini	7	11	4
	Tree20	0,1278793	0,1658025	Yes	Gini	7	10	5
	Tree24	0,1323408	0,1663661	Yes	Gini	5	11	5
	Tree23	0,1335532	0,1677187	Yes	Gini	7	9	5
	Tree10	0,1300131	0,1678314	Yes	Gini	6	10	5
	Tree	0,1403424	0,1730162	Yes	Gini	5	10	5
	Tree19	0,1356384	0,1743688	Yes	Gini	6	9	5
	Tree2	0,143446	0,1871055	No	Gini	8	11	5
	Tree16	0,1536298	0,1882326	Yes	Gini	5	9	5
	Tree27	0,1680811	0,2029982	Yes	Gini	4	11	5
	Tree26	0,1783619	0,2060415	Yes	Gini	3	11	5
	Tree17	0,1878182	0,2143823	Yes	Gini	3	10	5
	Tree18	0,1822414	0,2161858	Yes	Gini	4	10	5
	Tree21	0,2011542	0,2283589	Yes	Gini	3	9	5

Table 37 Multi-way split tree misclassification rates of data models where no labels for crop rotation schemas are present

				Data set Classification tree: labels for crop rotation schemas are not present, maximum branch > 2				
		Misclassification rate		Variable usage	Parameter settings			
Nr.	Tree name	Train	Validate	LAB_0_CRS	Nominal criterion	Maximum branch	Maximum depth	Minimum categorical size
1	Tree15	0,1181478	0,2191298	No	Gini	9	10	1
2	Tree16	0,147174	0,244022	No	Gini	9	10	2
3	Tree2	0,1737317	0,25343	No	Gini	9	10	3
	Tree19	0,1869254	0,2659741	No	Gini	7	10	3
	Tree3	0,1905005	0,2679341	No	Gini	9	10	4
	Tree10	0,2031835	0,2712662	No	Gini	7	10	5
	Tree14	0,2026728	0,2738142	No	Gini	9	10	5
	Tree8	0,2001192	0,2777342	No	Gini	8	10	5
	Tree18	0,1986721	0,2781262	No	Gini	7	10	4
	Tree20	0,2064181	0,2804782	No	Gini	11	10	5
	Tree21	0,2024174	0,2804782	No	Gini	10	10	5
	Tree11	0,2156112	0,2863583	No	Gini	6	10	5
	Tree17	0,2272727	0,2941984	No	Gini	7	10	6
	Tree9	0,2176541	0,2955704	No	Gini	5	12	5
	Tree12	0,2251447	0,3004704	No	Gini	5	11	5
	Tree13	0,2251447	0,3004704	No	Gini	5	11	5
	Tree5	0,2284644	0,3008624	No	Gini	9	10	6
	Tree	0,2440415	0,3081145	No	Gini	7	10	7
	Tree6	0,239445	0,3118385	No	Gini	5	10	5
	Tree4	0,2484678	0,3161505	No	Gini	9	10	7
	Tree7	0,2670242	0,3316347	No	Gini	5	9	5

## Appendix F - Classification tree properties description

### Decision Tree Node Train Properties

Variables	Use the Variables property to specify the properties of each variable that you want to use in the data source. Select the Ellipses Selector Button button to the right of the Variables property to open a variables table. You can specify whether to use a variable or generate a report. For input variables, the Use status determines whether a variable is included in the model. For target variables, the Use status determines which target is the primary target. You can have only one target variable with a Use status of Yes. When you train a model using the Interactive Decision Tree application, you can switch targets and use a target variable that has a Use status of No. The Explore functionality to view the distribution of a variable is available here.
Interactive	Use the Interactive property of the Decision Tree node to launch an interactive training session in the Interactive Decision Tree. Click the Ellipses Selector Button at the right of the Interactive Training property to launch the Enterprise Miner Interactive Decision Tree. For information about interactive training, see Interactive Decision Tree.
Use Frozen Tree	specifies whether a frozen tree definition should be used or if a new tree should be created during training. The Decision Tree node must have already been run before you can run the node with this property set to Yes.
Use Multiple Targets	specifies whether multiple targets should be available during training of the Decision Tree node.
Precision	specifies the number of decimal places displayed in the Variable Importance, Subtree Assessment, and Observation Based Importance tables and plots.

### Decision Tree Node Train Properties: Splitting Rule

Interval Criterion	<p>specify the method that you want to use to evaluate candidate splitting rules for interval variables and to search for the best one.</p> <p>Choose from the following splitting criteria:</p> <ul style="list-style-type: none"> <li><b>ProbF</b> — p-value of the F test that is associated with the node variance.</li> <li><b>Variance</b> — reduction in the square error from the node means.</li> </ul>
Nominal Criterion	<p>specify the method that you want to use to evaluate candidate splitting rules for nominal variables and to search for the best one.</p> <p>Choose from the following splitting criteria:</p> <ul style="list-style-type: none"> <li><b>ProbChisq</b> — p-value of Pearson Chi-square statistic for target versus the branch node.</li> <li><b>Entropy</b> — reduction in the entropy measure.</li> <li><b>Gini</b> — reduction in the Gini index.</li> </ul>
Ordinal Criterion	<p>specify the method that you want to use to evaluate candidate splitting rules for ordinal variables and to search for the best one.</p> <p>Choose from the following splitting criteria:</p> <ul style="list-style-type: none"> <li><b>Entropy</b> — reduction in the entropy measure, adjusted with ordinal</li> </ul>

	<p>distances.</p> <p><b>Gini</b> — reduction in the Gini index, adjusted with ordinal distances.</p>
Significance Level	<p>Specifies the maximum acceptable p-value for the worth of a candidate splitting rule when a <b>ProbChisq</b> or <b>ProbF</b> criterion method is selected. The measure of worth depends on the value of the Criterion property. For criterion methods that are based on p-values, the threshold is a maximum acceptable p-value. For other criteria, the threshold is the minimum acceptable increase in the measure of worth. Permissible values are real numbers greater than 0 and less than or equal to 1.</p>
Missing Values	<p>Use the Missing Values property of the <b>Decision Tree</b> node to specify how splitting rules handle observations that contain missing values for a variable. The default value is Use in search.</p> <p>Select from the following available missing value policies:</p> <p><b>Use in search</b> — uses missing values in the calculation of the worth of a splitting rule. This consequently produces a splitting rule that assigns the missing values to the branch that maximizes the worth of the split. This is a desirable option when the existence of a missing value is predictive of a target value.</p> <p><b>Most correlated branch</b> — assigns the observation to the branch with the smallest residual sum of squares among observations that contain missing values.</p> <p><b>Largest branch</b> — assigns the observations that contain missing values to the largest branch.</p>
Use Input Once	<p>The Use Input Once property of the Decision Tree node specifies whether a splitting variable can be used only once, or whether a splitting variable can be repeatedly used in splitting rules that apply to descendant nodes. The default value for the Use Input Once property is No.</p>
Maximum Branch	<p>Use the Maximum Branch property of the Decision Tree node to specify the maximum number of branches that you want a splitting rule to produce. Permissible values for the Maximum Branch property are integers between 2 and 100. The minimum value of 2 results in binary splits. The default value for the Maximum Branch property is 2.</p>
Maximum Depth	<p>Use the Maximum Depth property of the Decision Tree node to specify the maximum number of generations of nodes that you want to allow in your decision tree. The original node is the root node. Children of the root node are the first generation. Permissible values are integers between 1 and 50. The default number of generations for the Maximum Depth property is 6.</p>
Minimum Categorical Size	<p>Use the Minimum Categorical Size property of the Decision Tree node to specify the minimum number of training observations that a categorical value must have before the category can be used in a split search. Permissible values are integers greater than or equal to 1. The default value for the Minimum Categorical Size property is 5.</p>
Split Precision	<p>specifies the number of decimals displayed in the splitting values and average values displayed in nodes.</p>

## Decision Tree Node Train Properties: Node

Leaf Size	Use the Leaf Size property of the Decision Tree node to specify the minimum number of training observations that are allowed in a leaf node. Permissible values are integers greater than or equal to 1. The default setting is 5.
Number of Rules	Use the Number of Rules property of the Decision Tree node to specify the number of splitting rules that you want to save with each node. The tree uses only one rule. The remaining rules are saved for comparison. Permissible values are integers greater than or equal to 1. The default value for the Number of Rules property is 5.
Number of Surrogate Rules	Use the Number of Surrogate Rules property of the Decision Tree node to specify the maximum number of surrogate rules that the Decision Tree node seeks in each non-leaf node. The first surrogate rule is used when the main splitting rule relies on an input whose value is missing. Permissible values are nonnegative integers. The default value for the Number of Surrogate Rules property is 0.
Split Size	Use the Split Size property of the Decision Tree node to specify the smallest number of training observations that a node must have before it is eligible to be split. Permissible values are integers greater than or equal to 2. The Split Size property uses a default value of $(2 * \text{Leaf Size})$ , unless you specify an integer value that is greater than the calculated default value.

## Decision Tree Node Train Properties: Split Search

Use Decisions	Select Yes to use the decision information during the split search.
Use Priors	Select Yes to use the prior probabilities during the split search.
Exhaustive	Use the Exhaustive property of the Decision Tree node to specify the highest number of candidate splits that you want to find in an exhaustive search. The Exhaustive property applies to multi-way splits and to binary splits on nominal targets with more than two values. Permissible values are integers between 0 and 2,000,000,000. The default setting for the Exhaustive property is 5000.
Node Sample	Use the Node Sample property of the Decision Tree node to specify the maximum within-node sample size $n$ that you want to use to find splits. If the number of training observations in a node is larger than $n$ , then the split search for that node is based on a random sample of size $n$ . Permissible values are integers greater than or equal to 2. The default value for the Node Sample property is 20000.



## Decision Tree Node Train Properties: Subtree

Method	<p>Use the Method property to specify the method that you want to use to select a subtree from the fully grown tree for each possible number of leaves.</p> <p>The following subtree methods are available:</p> <p><b>Assessment</b> (default) — The smallest subtree with the best assessment value. The assessment value depends on the setting that you choose for the Assessment Measure property. Validation data set is used if available.</p> <p><b>Largest</b> — The largest (full) tree is selected.</p> <p><b>N</b> — The largest subtree with at most N leaves is selected. Use the Number of Leaves property to specify the value of N, the number of leaves.</p>
Number of Leaves	<p>When the Method property of the <b>Decision Tree</b> node is set to N, use the Number of Leaves property to specify the largest number of leaves that you want in a subtree of n leaves. Permitted values are integers greater than or equal to 1. The default value for the Number of Leaves property is 1.</p>
Assessment Measure	<p>Use the Assessment Measure property of the <b>Decision Tree</b> node to specify the method that you want to use to select the best tree, based on the validation data when the Method property is set to Assessment. If no validation data is available, training data is used.</p> <p>The available assessment measurements are as follows:</p> <p><b>Decision</b> (default setting) — The Decision method selects the tree that has the largest average profit and smallest average loss if a profit or loss matrix is defined. If no profit or loss matrix is defined, the value of the model assessment measure is reset in the training process, depending on the measurement level of the target. If the target is interval, the measure is set to Average Square Error. If the target is categorical, the measure is set to Misclassification.</p> <p><b>Average Square Error</b> — The Average Square Error method selects the tree that has the smallest average square error.</p> <p><b>Misclassification</b> — The Misclassification method selects the tree that has the smallest misclassification rate.</p> <p><b>Lift</b> — The Lift method evaluates the tree based on the prediction of the top n% of the ranked observations. Observations are ranked based on their posterior probabilities or predicted target values. For an interval target, it is the average predicted target value of the top n% observations. For a categorical target, it is the proportion of events in the top n% of the data. When you set the Measure property to Lift, you must use the Assessment Fraction property to specify the proportion for the top n% of cases.</p>
Assessment Fraction	<p>When the Assessment Measure property of the <b>Decision Tree</b> node is set to Lift, use the Assessment Fraction property to specify the proportion n (of the top n% of observations to use) during model assessment. Permissible values for the Percentage property are real numbers between 0 and 1. The default value for the Percentage property is 0.25. When a decision matrix has been defined and the Measure property is set to Lift, the percentage indicates the average profit or loss among the top n% of observations.</p>

## Decision Tree Node Train Properties: Cross Validation

Perform Cross Validation	Use the Perform Cross Validation property to specify whether to perform cross validation for each subtree in the sequence.
Number of Subsets	Use the Number of Subsets property to specify the number of cross validation subsets or folds.
Number of Repeats	Use the Number of Repeats property to specify the number of times to repeat cross validation. The estimates from repeated cross validation are the averages of the estimates from the individual cross validation runs.
Seed	Use the Seed property to specify the random number seed for generating the validation subsets.

## Decision Tree Node Train Properties: Observation-Based Importance

Observation Based Importance	Use the Observation Based Importance property to specify whether observation-based importance statistics should be generated. Variable importance is calculated using random forest tree methodology. See the section in this document on Variable Importance, as well as the Gradient Boosting Node documentation for more details about observation-based Importance.
Number Single Var Importance	Use the Number Single Var Import property to specify the number of variables for which one way importance statistics should be generated. The Number Single Var Import property is valid only when the Observation Based Importance property is enabled.

## Decision Tree Node Train Properties: P-Value Adjustment

Bonferroni Adjustment	When set to No, the Bonferroni Adjustment property of the Decision Tree node suppresses Bonferroni adjustments to the p-values. The default setting is Yes.
Time of Kass Adjustment	Use the Time of Kass Adjustment property of the Decision Tree node to indicate whether the Bonferroni adjustment should take place Before or After the split is chosen. The default setting is Before. The Time of Kass Adjustment property is ignored if the Bonferroni Adjustment property is set to No.
Inputs	When set to Yes, the Inputs property of the Decision Tree node adjusts the p-values for the number of inputs. The default setting for the Inputs property is No. When Inputs is set to Yes, you must use the Number of Inputs property to specify the number of inputs that you want to consider uncorrelated.
Number of Inputs	When the Inputs property of the Decision Tree node is set to Yes, use the Number of Inputs property to specify the number of inputs that you want to consider uncorrelated. The Number of Inputs property is ignored if the Inputs property is set to No. Permissible values are integers greater than or equal to 1. The default value for Number of Inputs is 1. If the specified value is greater than the number of input variables, then the Decision Tree node uses the number of input variable.
Split Adjustment	When set to Yes, the Split Adjustment property adjusts the p-values for the number of ancestor splits. The default setting for the Split Adjustment property is Yes.

## Decision Tree Node Train Properties: Output Variables

Leaf Variable	Set the Leaf Variable property of the Decision Tree node to No to indicate that you want to suppress the default creation of <code>_NODE_</code> variables in the output data. <code>_NODE_</code> variables store numeric identification numbers for each leaf that has observations assigned to it. The default setting for the Leaf Identifier node is Yes.
---------------	--

## Decision Tree Node Score Properties

Variable Selection	Use the Variable Selection property to specify whether variable selection should be performed based on importance values. If the Variable Selection property is set to Yes, all variables that have an importance value greater than or equal to 0.05 have the variable role set to Input. All other variables are set to Rejected. The default setting for the Variable Selection property is Yes.
Leaf Role	When the Leaf Identifier property of the Decision Tree node is set to Yes, use the Leaf Role property to specify the variable role that you want to apply to the created <code>_NODE_</code> variables. The selection choices are Segment, Input, and Rejected. The default setting is Segment.

## Appendix G – Splitting procedure for multi-way split classification trees

Correspondence with SAS Tech support about the manner in which the property 'maximum branch' is applied. The highlighted sentence stated that data in a tree is split up in subgroups according to the specified number for the 'maximum branch' or fewer:



Dear Lineke,

I hope this finds you well.

I attempted to reach you to discuss about your inquiry on the setting of maximum branch in the decision tree model, but without success. The Maximum Branch property sets an upper limit on the number of branches emanating from a node. Here is an excerpt from EM->Help -> Contents -> search for " Decision Tree Node" ->Decision Tree Node Train Properties: Splitting Rule:

Maximum Branch — specifies the maximum number of branches that you want a splitting rule to produce. Permissible values for the Maximum Branch property are integers between 2 and 100. The minimum value of 2 results in binary splits. The default value for the Maximum Branch property is 2.

Or if you click on "Maximum Branch" underneath the Splitting rule on the Property windows at the left side of EM GUI:

### Maximum Branch

Restricts the number of subsets that a splitting rule can produce to the specified number or fewer. For example, a value of 2 results in binary trees.

If you set the maximum branch to 5, it means you set all possible splits to have a maximum of 5 branches.

Please let me know if you have any follow-up questions. If not, I will proceed to archive the track.

Thank you.

Kindest regards,  
Shānshan Cóng  
Technical Consultant  
SAS Institute B.V.  
Tel: 035 699 6969  
<http://www.sas.com/nl>  
[support@snl.sas.com](mailto:support@snl.sas.com)  
SAS® | THE POWER TO KNOW®

## Appendix H - Autoneural network properties description

### AutoNeural Node General Properties

Node ID	<p>The Node ID property displays the ID that Enterprise Miner assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first AutoNeural node added to a diagram will have a Node ID of AutoNeural. The second AutoNeural node added to a diagram will have a Node ID of AutoNeural2, and so on.</p>
Imported Data	<p>the Imported Data property provides access to the Imported Data — AutoNeural window. The Imported Data — AutoNeural window contains a lists of the ports that provide data sources to the AutoNeural node. Select the Ellipses Selector Buttonbutton to the right of the Imported Data property to open a table of the imported data.</p> <p>If data exists for an imported data source, you can select the row in the imported data table and click:</p> <ul style="list-style-type: none"><li><b>Browse</b> to open a window where you can browse the data set.</li><li><b>Explore</b> to open the Explore window, where you can sample and plot the data.</li><li><b>Properties</b> to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contains summary information (metadata) about the table and the variables.</li></ul>
Exported Data	<p>the Exported Data property provides access to the Exported Data — AutoNeural window. The Exported Data — AutoNeural window contains a list of the output data ports that the AutoNeural node creates data for when it runs. Select the Ellipses Selector Buttonbutton to the right of the Exported Data property to open a table of the exported data.</p> <p>If data exists for an imported data source, you can select the row in the imported data table and click:</p> <ul style="list-style-type: none"><li><b>Browse</b> to open a window where you can browse the data set.</li><li><b>Explore</b> to open the Explore window, where you can sample and plot the data.</li><li><b>Properties</b> to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contains summary information (metadata) about the table and the variables.</li></ul>
Notes	<p>Select the Ellipses Selector Buttonbutton to the right of the Notes property to open a window that you can use to store notes of interest, such as data or configuration information.</p>

## AutoNeural Node Train Properties

Variables	<p>Use the Variables window to view variable information, and specify the Use and Report status of a variable. Select the Ellipses Selector Button to open a window containing the variables table. You can specify the Use and Report status of a variable, view the columns metadata, or open an Explore window to view a variable's sampling information, observation values, or a plot of variable distribution. In default, the Variables window displays columns for a variable's name, use, report, role, and level.</p> <p>You can use the following buttons to view additional metadata, or limit metadata:</p> <ul style="list-style-type: none"> <li><b>Apply</b> — Changes metadata based on the values supplied in the drop-down menus, check box, and selector field.</li> <li><b>Reset</b> — Changes metadata back to its state before use of the Apply button.</li> <li><b>Label</b> — Adds a column for a label for each variable.</li> <li><b>Mining</b> — Adds columns for the Order, Lower Limit, Upper Limit, Creator, Comment, and Format Type for each variable.</li> <li><b>Basic</b> — Adds columns for the Type, Format, Informat, and Length of each variable.</li> <li><b>Statistics</b> — Adds statistics metadata for each variable.</li> <li><b>Explore</b> — Opens an Explore window that allows you to view a variable's sampling information, observation values, or a plot of variable distribution.</li> </ul>
-----------	--

## AutoNeural Node Train Properties: Model Options

Architecture	<p>Use the Architecture property of the AutoNeural node to specify the neural network architecture that you want to use when you build the network.</p> <p>You can choose from the following networks:</p> <ul style="list-style-type: none"> <li><b>Single Layer</b> — (Default Setting) hidden layers are added in parallel. One or more activation functions can be used.</li> <li><b>Block Layers</b> — hidden layers are added as additional layers with a uniform number of neurons.</li> <li><b>Funnel Layers</b> — hidden nodes are added both to each existing hidden layer and to a new layer, so that the layers form a funnel pattern with a single node feeding into the output node.</li> <li><b>Cascade</b> — hidden nodes are added in a cascading fashion.</li> </ul>
Termination	<p>Use the Termination property of the AutoNeural node to specify the method that you want to use to end training. Training always stops when the maximum run time is exceeded. When training is terminated, the model that has the best selection criteria is retained.</p> <p>You can choose from the following termination methods:</p> <ul style="list-style-type: none"> <li><b>Overfitting</b> — (Default Setting) training stops when overfitting is detected.</li> <li><b>Training Error</b> — training stops when the reduction in training data set error is less than 0.001.</li> <li><b>Time Limit</b> — training stops when the time specified in the Total time property is exceeded.</li> </ul>

Train Action	<p>Use the Train Action property of the AutoNeural node to specify the method that you want to use when training.</p> <p>You can choose from the following actions:</p> <p><b>Train</b> — all selected functions are trained until stopping.</p> <p><b>Increment</b> — nodes are added one at a time. No activation function is reused. If a layer lowers the average error, it is retained. If it does not, then it is dropped from the model.</p> <p><b>Search</b> — (Default Setting) nodes are added according to the architecture. The best network is retained as the final model.</p>
Target Layer Error Function	<p>Each unit in a neural network produces a single computed value. Input and hidden units pass the computed values to other hidden or output units. The predicted value for output units is compared with the target value to compute the error function. Use the Target Layer Error Function property to specify the target layer error function for a user-defined network. Permissible values are</p> <p><b>Default</b> — For a categorical target variable, the default error function is multiple Bernoulli. For interval targets, the default error function is normal distribution.</p> <p><b>Normal</b> — may be used with any kind of target variable to predict the conditional mean. It is most often used with interval targets for which the noise distribution is approximately normal with constant variance. It can also be used with categorical targets for robust estimation of posterior probabilities.</p> <p><b>Cauchy</b> — The Cauchy distribution may be used with any kind of target variable. It is most often used when you want to predict the approximate conditional mode instead of the conditional mean.</p> <p><b>Logistic</b> — Logistic distribution may be used with any kind of target variable. It is most often used with interval targets where the noise distribution may contain outliers.</p> <p><b>Huber</b> — The Huber M-estimator is suitable for unbounded interval targets that have outliers or that have a moderate degree of inequality of the conditional variance and a symmetric distribution. Huber can also be used for categorical targets when you want to predict the mode rather than the posterior probability.</p> <p><b>Biweight</b> — The Biweight M-estimator may be used with any kind of target variable. It is most often used with interval targets where the noise distribution may contain severe outliers. Because of severe problems with local minima, you should obtain initial values by training with the Huber M-estimator before using the biweight M-estimator.</p> <p><b>Wave</b> — The Wave M-estimator may be used with any kind of target variable. It is most often used with interval targets where the noise distribution may contain severe outliers. Because of severe problems with local minima, you should obtain initial values by training with the Huber M-estimator before using the Wave M-estimator.</p> <p><b>Gamma</b> — The Gamma distribution may be used only with strictly positive interval target variables. It is most often used when the standard deviation of the noise is proportional to the mean of the target variable.</p> <p><b>Poisson</b> — The Poisson distribution is suitable for skewed, nonnegative interval targets, especially counts of rare events, where the conditional variance is proportional to the conditional mean.</p> <p><b>Bernoulli</b> — Suitable for a target that takes only the values zero and one. Same as a binomial distribution with one trial.</p> <p><b>Entropy</b> — The Entropy error function is cross-entropy or relative entropy for independent interval targets with values between zero and one inclusive.</p> <p><b>MBernoulli</b> — The Multiple Bernoulli error function is suitable for categorical (nominal or ordinal) targets.</p>

	<p><b>Multinomial</b> — The Multinomial error function is may be used only with two or more interval target variables with non-negative values, where each case represents a number of trials that are equal to the sum of the target values. The activation function must force the outputs to sum to one, like Softmax.</p> <p><b>Mentropy</b> — The Multiple Entropy error function is cross-entropy or relative entropy for targets that sum to 1. It is the same criterion as Kullback-Leibler divergence. Multiple entropy should normally be used only with two or more interval target variables with values that lie between zero and one inclusive, and that sum to one for each case. However, multiple entropy may also be used with nominal or ordinal targets, primarily for software testing. The activation function must force the outputs to sum to one, like Softmax.</p>
Maximum Iterations	Use the Maximum Iterations property of the AutoNeural node to specify the maximum number of iterations permitted during training. Permissible values are any integer between 1 and 50. The default setting is 8.
Number of Hidden Units	Use the Number of Hidden Units property of the AutoNeural node to specify the number of hidden units that you want to use. Permissible values are integers between 1 and 8. The default setting is 2.
Tolerance	Use the Tolerance property of the AutoNeural node to configure the extent of the preliminary search. Valid values are <p><b>Low</b> — no preliminary search is performed.</p> <p><b>Medium</b> — (Default Setting) preliminary statements are executed.</p> <p><b>High</b> — preliminary statements are executed, as well as applying the setting ABSCONV=0.001 to the training data.</p>
Total time	<p>The total amount of time allowed for training. The allowable choices are:</p> <p><b>Five Minutes</b>  <b>Ten Minutes</b>  <b>Thirty Minutes</b>  <b>One Hour</b>  <b>Two Hours</b>  <b>Four Hours</b>  <b>Seven Hours</b>  <b>One Day</b>  <b>Two Days</b>  <b>Four Days</b>  <b>Seven Days</b></p> <p>The default Total Time setting is 1 hour.</p>



## AutoNeural Node Train Properties: Increment and Search

Adjust Iterations	When set to No, the Adjust Iterations property of the AutoNeural node suppresses adjustments that are made to the Maximum Iterations property value setting when the Train Action is set to Search or Increment. If the Train Action is set to Search or Increment and Adjust Iterations is set to Yes, then the Maximum Iterations value is adjusted higher if the selected iteration equals the previously used Maximum Iterations. Similarly, the Maximum Iterations value can be adjusted lower if the selected iteration is significantly lower than the previously used Maximum Iterations. The default setting for the Adjust Iterations property is Yes.
Freeze Connections	Use the Freeze Connections property of the AutoNeural node to specify whether connections should be frozen. The default setting for the Freeze Connections property is No.
Total Number of Hidden Units	Use the Total Number of Hidden Units property of the AutoNeural node to specify the total hidden units ceiling when the Train Action is set to Search. When Search is performed, the total number of hidden units trained is calculated at each run. If the calculated number of total hidden units is greater than or equal to the value stored in Total Hidden, then training stops and a final model is selected. Permissible values are 5, 10, 20, 30, 40, 50, 75, or 100. The default setting is 30.
Final Training	Use the Final Training property of the AutoNeural node to indicate whether the final model should be trained again to allow the model to converge. If the Final Training property is set to Yes, the number of iterations that are used will correspond to the value set in the Final Iterations property.
Final Iterations	Use the Final Iterations property of the AutoNeural node to indicate the number of iterations to use when the Final Training property is set to Yes. The Final Iterations property is unavailable if the Final Training property is set to No. The Final Iterations property accepts integers greater than zero.

## AutoNeural Node Train Properties: Activation Functions

	<p>Use Activation Functions property group to set the use status for each of the available activation functions:</p> <p><b>Direct</b> — Use the Direct property of the AutoNeural node to indicate that you want to use the direct activation function during training. The default setting of the Direct property is Yes.</p> <p><b>Exponential</b> — Use the Exponential property of the AutoNeural node to indicate that you want to use the exponential activation function during training. The default setting of the Exponential property is No.</p> <p><b>Identity</b> — Use the Identity property of the AutoNeural node to indicate that you want to use the identity activation function during training. The default setting of the Identity property is No.</p> <p><b>Logistic</b> — Use the Logistic property of the AutoNeural node to indicate that you want to use the logistic activation function during training. The default setting of the Logistic property is No.</p> <p><b>Normal</b> — Use the Normal property of the AutoNeural node to indicate that you want to use the normal activation function during training. The default setting of the Normal property is Yes.</p> <p><b>Reciprocal</b> — Use the Reciprocal property of the AutoNeural node to indicate that you want to use the reciprocal activation function during training. The default setting of the Reciprocal property is No.</p> <p><b>Sine</b> — Use the Sine property of the AutoNeural node to indicate that you want to use the sine activation function during training. The default setting of the Sine property is Yes.</p> <p><b>Softmax</b> — Use the Softmax property of the AutoNeural node to indicate that you want to use the softmax activation function during training. The default setting of the Softmax property is No.</p> <p><b>Square</b> — Use the Square property of the AutoNeural node to indicate that you want to use the square activation function during training. The default setting of the Square property is No.</p> <p><b>Tanh</b> — Use the Tanh property of the AutoNeural node to indicate that you want to use the hyperbolic tangent activation function during training. The default setting of the Tanh property is Yes.</p>
--	--

## AutoNeural Node Score Properties

Hidden Units	Use the Hidden Units property of the AutoNeural node to specify whether or not you want to create hidden unit variables. The default setting of the Hidden Units property is No.
Residuals	Use the Residuals property of the AutoNeural node to specify whether or not you want to create residual variables. The default setting of the Residuals Property is Yes.
Standardization	Use the Standardization property of the AutoNeural node to specify whether or not you want to create standardization variables. The default setting of the Standardization property is No.

## Appendix I - Variable importance description

Extract from SAS Enterprise Miner help description:

### Variable Importance

Some tree node output variables are selected based on their relative importance, which is reported in the Variable Importance Table. This is an overview of the components of variable importance. The relative importance of an input variable  $u$  in subtree  $T$  is computed as follows:

$$I(u; T) \propto \sqrt{\sum_{\tau \in T} a(s_u, \tau) \Delta SSE(\tau)}$$

Here, the sum is over nodes  $\tau$  in  $T$ , and  $s_u$  denotes the primary or surrogate splitting rule using  $u$ .  $a(s_u, \tau)$  is the measure of agreement for the rule using  $u$  in node  $\tau$ :

$$a(s_u, \tau) = \begin{cases} 1 & \text{if } s_u \text{ is the primary splitting rule} \\ \text{agreement} & \text{if } s_u \text{ is a surrogate rule} \\ 0 & \text{otherwise} \end{cases}$$

$\Delta SSE(\tau)$  is the reduction in sum of square errors from the predicted values:

$$\Delta SSE(\tau) = SSE(\tau) - \sum_{b \in B(\tau)} SSE(\tau_b)$$

$$SSE(\tau) = \begin{cases} \sum_{i=1}^{N(\tau)} (Y_i - \hat{Y}(\tau))^2 & \text{for interval target } Y \\ \sum_{i=1}^{N(\tau)} \sum_{j=1}^J (\delta_{ij} - \hat{p}_j(\tau))^2 & \text{for target with } J \text{ categories} \end{cases}$$

These conditions apply:

- $B(\tau)$  = set of branches from  $\tau$
- $\tau_b$  = child node of  $\tau$  in branch  $b$
- $N(\tau)$  = number of observations in  $\tau$
- $\hat{Y}(\tau)$  = average  $Y$  in training data in  $\tau$
- $\delta_{ij}$  = 1 if  $Y_i = j$ , 0 otherwise
- $\hat{p}_j(\tau)$  = average  $\delta_{ij}$  in training data in  $\tau$

For a categorical target, the formula for  $SSE(\tau)$  reduces to

$$SSE(\tau) = \begin{cases} N(1 - \sum_{j=1}^J \hat{p}_j^2) & \text{for training data} \\ N(1 - \sum_{j=1}^J (2p_j - \hat{p}_j)\hat{p}_j) & \text{for validation data} \end{cases}$$

Here,  $p_j$  is the proportion of the validation data with target value  $j$ , and  $N$ ,  $p_j$ , and  $\hat{p}_j$  are evaluated in node  $\tau$ .

In the Decision Tree tool, the sum of square errors for a categorical target variable is always computed from the training data set. Therefore, the sum of square errors is equal to the Gini index.

Relative importance is established by computing how much error reduction in the predictive values is produced when a variable is applied as a primary or surrogate splitting rule. By doing this for all the input variables, relative importance of the variables can be computed

## Appendix J – Select sub parcels from combined variable data to use as input data for modeling (ArcGIS model)

Executed steps:

ArcGIS function	Description	Result (file)
Input file	BBR_PCR_identity_AOI	
Select	ORIGIN_OPP <1000	BBR_PCR_not_split_1000m2
Add field	OPP_PERC (double)	
Calculate field	OPP_PERC = ([SHAPE_Area] / [ORIGIN_OPP]) *100	
Summary statistics	per UNIEK (identifier) select maximum of OPP_PERC	SS_BBR_PCR_not_split_1000m2
Make feature layer		
Join field	Add OPP_PERC max to BBR_PCR_not_split_1000m2	
Table to table	Export table where OPP_PERC = MAX_OPP_PERC	TAB_BBR_PCR_not_split_1000m2_content_rows
Select layer by attribute	clear selection	
Dissolve	dissolve all feature parts based on the UNIEK identifier	BBR_PCR_not_split_1000m2_diss
Join field	Add all columns of TAB_BBR_PCR_not_split_1000m2_content_rows to feature file BBR_PCR_not_split_1000m2_diss based on the UNIEK identifier	<b>BBR_PCR_not_split_1000m2_diss</b>

ArcGIS function	Description	Result (file)
Input file	BBR_PCR_identity_AOI	
Select	ORIGIN_OPP >=1000	BBR_PCR_split
Add field	RATIO_AP (double)	
Calculate field	RATIO_AP = !SHAPE_Length! /math.sqrt( !SHAPE_Area!)	
Add field	Sliver (short)	
Make feature layer		
Select layer by attribute	SHAPE_Area <50 OR ( RATIO_AP > 7 AND SHAPE_Area >= 50 AND SHAPE_Area < 500) OR ( RATIO_AP > 19 AND SHAPE_Area >= 500 AND SHAPE_Area < 1000) OR ( RATIO_AP > 27 AND SHAPE_Area >= 1000)	
Calculate field	Sliver = 1	

Column Sliver: sliver are indicated with the value 1, sub parcels are indicated with the value NULL

Model name: 01\_PMF\_PREP\_SLIVER\_REMOVAL\_PARCEL\_SMALLER\_1000M2

## Appendix K - Arable area of municipalities present in study area (CBS 2016)

Rank (based on percentage arable Agriculture)	Rank (based on area arable Agriculture)	Municipality	Total Agricultural area (ha)	Arable Agriculture area (ha)	Percentage arable Agriculture	Percentage arable Agriculture based on area NOP (27044 ha)
18	1	Noordoostpolder	37088	27044	73%	100%
8	2	Dronten	23817	17967	75%	66%
31	5	Emmen	20585	13763	67%	51%
35	10	Oldambt	16340	10553	65%	39%
23	13	Schouwen-Duiveland	13575	9396	69%	35%
65	19	Aa en Hunze	15082	7230	48%	27%
101	77	Eijsden-Margraten	4551	1391	31%	5%
106	78	Gulpen-Witterm	4906	1365	28%	5%

Source: Boerenbusiness (2016)

The table shows the information of municipalities that are present in the study area of this research, except for the area containing the organic farms.

Some columns were added to the original table for more clarification:

Columns	Adjustment
Rank (based on percentage arable Agriculture)	From original table
Rank (based on area arable Agriculture)	Added to the original table
Municipality	From original table
Total Agricultural area (ha)	From original table
Area arable Agriculture (ha)	From original table
Percentage arable Agriculture	Added to the original table
Percentage arable Agriculture based on area NOP (27044 ha)	Added to the original table

## Appendix L – Crop rotation schemas in the unlabeled input data set

Crop sequence	Frequency of occurrences	Possible length crop rotation schema
1004 - 233 - 2014 - 233 - 256 - 233 - 2014 - 236 - 256	18	2
1004 - 233 - 2014 - 233 - 262 - 233 - 2015 - 262 - 176	10	2
1004 - 266 - 259 - 266 - 266 - 266 - 2014 - 266	14	2
1931 - 233 - 2014 - 233 - 256 - 233 - 2014 - 236 - 234	12	2
2014 - 233 - 256 - 233 - 2014 - 233 - 256 - 235 - 233	10	4
2015 - 1004 - 266 - 2015 - 263 - 672 - 2015 - 176 - 233	14	3
2015 - 233 - 256 - 2015 - 233 - 236 - 2015 - 259 - 233	16	3
2015 - 233 - 427 - 236 - 233 - 256 - 236 - 233 - 233	10	3
2015 - 263 - 259 - 2015 - 176 - 234 - 2015 - 263 - 672	10	3
2016 - 256 - 2016 - 233 - 2016 - 511 - 2016 - 256 - 233	10	2
2016 - 259 - 2016 - 233 - 2016 - 511 - 2016 - 256 - 233	10	2
2017 - 236 - 2016 - 256 - 316 - 236 - 2016 - 256 - 236	15	4
2017 - 259 - 2014 - 259 - 2016 - 259 - 2016 - 256 - 2016	17	2
233 - 1922 - 233 - 2015 - 233 - 1575 - 1575 - 2015 - 233	12	2
233 - 1922 - 233 - 233 - 233 - 1922 - 266 - 266 - 2034	12	2
233 - 2014 - 233 - 256 - 233 - 2014 - 262 - 672 - 233	10	2
233 - 2014 - 233 - 256 - 233 - 259 - - - 233 - 2014	10	2
233 - 233 - 233 - - - 233 - 233 - 259 - 259	12	1
233 - 233 - 233 - 256 - 233 - - - 233 - 259 - 259	14	1
233 - 234 - 233 - 2014 - 233 - 236 - 233 - 2014 - 236	12	2
233 - 256 - 233 - 2014 - 233 - 256 - 233 - -	10	2
233 - 259 - 233 - 2014 - 233 - 256 - 259 - 265 - 265	10	2
233 - 259 - 233 - 316 - 233 - 256 - 233 - - 233	12	2
233 - 262 - 233 - 2014 - 233 - 233 - 256 - 244 - 233	16	2
233 - 262 - 233 - 2014 - 233 - 256 - 233 - 2014 - 262	12	2
233 - 2785 - 233 - 262 - 233 - 2015 - 176 - 233 - 262	18	2
235 - 233 - 2014 - 233 - 256 - 233 - 2014 - 235 - 233	10	2
236 - 2016 - 256 - 2016 - 233 - 2016 - 233 - 2016 -	10	2
236 - 2016 - 256 - 2016 - 233 - 2016 - 233 - 2016 - 236	10	2
236 - 2017 - 236 - 2016 - 236 - 256 - 2016 - 236 - 2016	14	2
256 - 233 - 2014 - 233 - 236 - 233 - 2014 - 233 - 234	12	2
256 - 233 - 2014 - 233 - 236 - 233 - 2014 - 233 - 262	12	2
256 - 233 - 2014 - 233 - 236 - 233 - 2014 - 236 - 256	12	2
256 - 233 - 2014 - 233 - 262 - 233 - 2014 - 233 - 256	11	4
256 - 233 - 262 - 233 - 2014 - 233 - 316 - 233 - 259	16	2
256 - 259 - 2015 - 259 - 259 - 2014 - 259 - 259	16	2
256 - 259 - 259 - 259 - 256 - 259 - 259 - 256 -	10	4
256 - 259 - 259 - 259 - 256 - 259 - 259 - 256 - 259	16	1
259 - 2014 - 259 - 259 - 259 - 2014 - 672 - 266 - 266	10	4
259 - 2016 - 256 - 2016 - 234 - 2016 - 256 - 2016 - 259	11	2
259 - 2017 - 259 - 259 - 259 - 2016 - 259 - - 259	12	4
259 - 234 - 259 - 233 - 259 - 265 - 265 - 265 - 265	11	2
259 - 256 - 259 - 233 - 259 - 256 - 233 - 259 - 233	18	2
259 - 256 - 259 - 233 - 259 - 259 - 233 - 256 - 233	14	2
259 - 259 - 259 - 233 - 256 - 233 - 2014 - 233 - 256	10	1
259 - 259 - 259 - 266 - 266 - 266 - 266 - 266 - 266	24	1
259 - 262 - 2015 - 176 - 233 - 2015 - 176 - 233 - 2015	11	3
259 - 266 - 259 - 266 - 176 - 266 - 266 - 266 - 266	112	2
259 - 266 - 259 - 266 - 266 - 266 - 233 - 256 - 233	20	2
262 - 233 - 2014 - 233 - 256 - 233 - 2014 - 233 - 262	12	2
265 - 265 - 266 - 266 - 266 - 266 - 266 - 266 - 266	12	1
266 - 266 - 266 - 176 - 266 - 259 - 176 - 266 - 266	16	1
266 - 266 - 266 - 176 - 266 - 259 - 2014 - 266 - 266	12	1
266 - 266 - 266 - 2014 - 266 - 262 - 672 - 258 - 2014	10	1
266 - 266 - 266 - 258 - 266 - 265 - 265 - 863 -	40	1
266 - 266 - 266 - 258 - 266 - 265 - 863 - 863 -	24	1

## Appendix M - Analyses of sub parcels

Due to combining the declaration information based on a spatial overlay, the declared parcels were split up into sub parcels (this information mattered for predicting crop types) and slivers (errors). For both the input data sets used to create models, ranges of sub parcels present per unique declared parcels were investigated, see Table 38:

Table 38 Number of sub parcels per declared parcel for both input data sets

Number of sub parcels per unique declared parcel	Frequency of occurrence	Percentage
1	5.278	42,8%
2 - 5	4.962	40,3%
6 - 10	1.276	10,4%
More than 10	802	6,5%
Total	12.318	100,0%

In almost 43% of the cases only one sub parcel is present for a declared parcel. For another 40% two till 5 sub parcels are present per unique declared parcel. Another 10% contains 6 till 10 sub parcels per unique declared parcel and the remaining 6,5% has more than 10 sub parcels per unique declared parcel.

The division of the area classes for sub parcels and declared parcels differs also. Therefore the number of sub parcels and the declared parcels per area class were compared, see Table 39 for a division of this information into area classes:

Table 39 Number of sub parcels and declared parcels per area class

Area class	Sub parcels		Complete declared parcels	
	Total number	Percentage	Total number	Percentage
< 100 m <sup>2</sup>	2.028	4,3%	6	0,0%
> 100 m <sup>2</sup> and < 250 m <sup>2</sup>	2.789	6,0%	15	0,1%
> 250 m <sup>2</sup> and < 500 m <sup>2</sup>	4.249	9,1%	41	0,3%
> 500 m <sup>2</sup> and < 1000 m <sup>2</sup>	5.304	11,4%	96	0,8%
> 1000 m <sup>2</sup> and < 2500 m <sup>2</sup>	5.771	12,4%	279	2,3%
> 2500 m <sup>2</sup> and < 5000 m <sup>2</sup>	5.278	11,3%	431	3,5%
> 5000 m <sup>2</sup> and < 10000 m <sup>2</sup>	5.351	11,5%	864	7,0%
> 10000 m <sup>2</sup> and < 25000 m <sup>2</sup>	7.948	17,0%	2.621	21,3%
> 25000 m <sup>2</sup> and < 50000 m <sup>2</sup>	5.561	11,9%	3.542	28,8%
> 50000 m <sup>2</sup>	2.422	5,2%	4.423	35,9%
Total	46.701	100,0%	12.318	100,0%

Relative many small sub parcels are present compared to the division of declared parcels with a small area. Almost 20% of the sub parcels are smaller than 500 m<sup>2</sup> compared to 0,4% of the declared parcels. On average almost 3,8 sub parcel exists for every declared parcel.

The accuracy of the sub parcels is researched per area class, based on the most accurate prediction results for labeled and unlabeled data (presence of crop rotation schemas). The prediction result of Tree39 (see Table 21) and Tree16 (see Table 22) were used, see Table 40:



Table 40 Number of correctly and incorrectly classified sub parcels for tree39 and tree16 per area class

Area class	Classified sub parcels				Total
	Total incorrect classification	Percentage incorrect classification	Total correct classification	Percentage correct classification	
Area: < 100 m2	1.296	63,9%	732	36,1%	2.028
Area: > 100 m2 and < 250 m2	1.742	62,5%	1.047	37,5%	2.789
Area: > 250 m2 and < 500 m2	1.468	65,9%	759	34,1%	2.227
Area: > 500 m2 and < 1000 m2	2.871	61,5%	1.801	38,5%	4.672
Area: > 1000 m2 and < 2500 m2	4.831	62,0%	2.962	38,0%	7.793
Area: > 2500 m2 and < 5000 m2	3.122	59,2%	2.156	40,8%	5.278
Area: > 5000 m2 and < 10000 m2	3.265	59,8%	2.197	40,2%	5.462
Area: > 10000 m2 and < 25000 m2	4.875	56,8%	3.705	43,2%	8.580
Area: > 25000 m2 and < 50000 m2	3.080	56,5%	2.370	43,5%	5.450
Area: > 50000 m2	1.247	51,5%	1.175	48,5%	2.422
Total	27.797	59,5%	18.904	40,5%	46.701

Results show that relative more crop types are predicted inaccurately when the area of a sub parcel is small. On average almost 64% of the sub parcels is predicted inaccurately when the area is smaller than 100 m2 versus 51% when the area is larger than 5 ha.

## Appendix N – Double claims

Parcels can be claimed by more than one applicant, therefore overlap existed in the input data sets, see Figure 12. Part of the overlap in the declaration information was caused by poorly aligned parcels however overlap was also caused by double claims. Usually this double claim would disappear as soon as the issue was resolved by approving only the declaration of the applicant who was using the double claimed parcel in the declaration year. Unfortunately, during the administration process not all double claims were updated, and therefore overlap still existed in the declaration information. It is difficult to almost impossible to identify the rightful claim. Most of these overlaps occurred in the older declaration information. At that time many declarations were still delivered on paper, including the handling of the double claim.

The ArcGIS tooling intersect was used to extract the overlap from the input data set. To extract the area that was involved in the overlap, the ArcGIS tooling dissolve was used. In the input data set where labels were present for crop rotation schemas, 12.634 out of 29.622 sub parcels overlapped. These numbers represent 4.608 ha out of 32.817 ha in total which is 14,0%. In the input data set where no labels were present for crop rotation schemas, 7.010 out of 17.079 sub parcels overlapped. These numbers represent 2.265 ha out of 13.973 ha in total which is 16,2%.



Figure 12 Overlapping parcels (location: agricultural area southwest of Emmeloord)

To minimize the overlap when prediction results were created for complete declared parcels of 2017, the prediction results were dissolved based on correct or incorrect classified parcels for 2016 and 2017. Most of the overlap disappeared, only total 83 overlaps remained. The area involved in the overlap was 6 ha in total. In 80 cases the area of the overlap was less than 5% of the surface of the declared parcel. The declared parcels where overlap existed were left out of the prediction results.

## Appendix O – Declared parcels in 2016 and 2017

The boundaries of declared parcels are not fixed. Figure 13 shows an example of a parcel that was declared in 2016 however not again declared in 2017, and vice versa.



Figure 13 Example of difference in declared crop types for 2016 and 2017  
(location: agricultural area northeast of Tollebeek near Emmeloord)

When the declaration information of 2016 and 2017 was compared based on a select by location (based on intersecting polygons for the whole of the Netherlands), 22.283 parcels were declared in 2017 that were not declared in 2016. This is 2,7% of all parcels that were declared in 2017. In 2016 16.357 were declared that were not declared again in 2017. This is 2,0% of all parcels that were declared in 2016.