
Big data analysis for smart farming

Results of T02 project in theme food security

Kempenaar, C.^{1a}, Lokhorst, C.^{1b}, Bleumer, E.J.B.^{1b}, Veerkamp, R.F.^{1b},
Been, Th.^{1a}, Evert, F.K. van^{1a}, Boogaardt, M.J.^{1c}, Ge, L.^{1c}, Wolfert, J.^{1c},
Verdouw, C.N.^{1c}, Bekkum, M.A. van², Feldbrugge, L.², Verhoosel, J.P.C.²,
Waaij, B.D. van der², Persie, M. van³, and Noorbergen, H.³



¹ Wageningen University & Research (^a WPR, ^b WLR & ^c LEI)

² TNO

³ NLR

© 2016

**Wageningen University & Research, Wageningen Plant Research,
Business Unit Agrosystems Research**

Address : P.O. Box 616, 6700 AP Wageningen, The Netherlands
: Wageningen Campus, Droevendaalsesteeg 1, Wageningen, The Netherlands
Tel. : +31 317 48 04 98
Fax : +31 317 41 80 94
E-mail : corne.kempenaar@wur.nl
Internet : www.wageningenUR.nl/en/pri

Table of contents

	page
Preface	1
Summary	2
1. Introduction	7
1.1 Rationale	7
1.2 Precision farming and smart agri-food chains	7
1.3 Big data technology	8
1.4 Objectives, approach and report structure	9
2. Literature study	11
2.1 Introduction	11
2.2 Methodology	11
2.3 Conceptual framework	12
2.3.1 Farm processes	13
2.3.2 Farm management	13
2.3.3 Data chain	14
2.3.4 Network management organization	14
2.3.5 Network management technology	15
2.4 Results	15
2.4.1 Drivers for Big Data in Smart Farming	15
2.4.2 Business processes	16
2.4.3 Stakeholder network	19
2.4.4 Network management	20
2.4.5 Challenges	22
2.5 Conclusions and recommendations	22
3. Description of a dairy farm case	25
3.1 Introduction	25
3.2 Case: Estimation of feed efficiency of individual dairy cows	25
4. Semantics and linked data	28
4.1 Linked data and ontologies	28
4.1.1 Linked data, big data, open data	28
4.1.2 The essence of Linked Data	29
4.1.3 Ontologies: engineering Linked Data	31
4.2 Application to the use case	31
4.2.1 Ontology matching approach	31
4.2.2 Triplification, triple stores and SPARQL queries	33
4.3 Insights	35
5. Analysis with Machine Learning methods	36
5.1 Introduction	36
5.2 Data description	36
5.2.1 Dairy Campus data	36
5.2.2 KNMI data	37

5.3	Machine learning techniques	37
5.3.1	Artificial Neural Networks	37
5.3.2	Support Vector Machines	38
5.3.3	Comparison	39
5.4	Approach feed intake prediction	39
5.4.1	Data inspection	39
5.4.2	Data cleaning	42
5.4.3	Initial parameter determination	43
5.4.4	Data set division	44
5.4.5	Model setup	44
5.4.6	Fine-tuning the model	45
5.4.7	Input investigations	46
5.5	Results	46
5.6	Insights in machine learning techniques for analysis	48
6.	Remote sensing based grass growth analysis	49
6.1	Introduction	49
6.2	Source data	49
6.2.1	Dairy Campus grazing experiment	49
6.2.2	Satellite data	50
6.2.3	UAS recordings	51
6.2.4	Field measurements	51
6.2.5	Overview of all measurements	52
6.3	Remote sensing based grass growth monitoring	52
6.3.1	Principle	52
6.3.2	Vegetation index measurement	52
6.3.3	Vegetation index in relation to grass harvest weight	55
6.3.4	Vegetation index in relation to grass growth	57
6.3.5	UAS based height measurement	61
6.3.6	UAS based height in relation to grass growth	64
6.4	Big Data technologies	66
6.5	Conclusions and recommendations	66
7.	Dairy Campus ICT infrastructure for the dairy farm case	68
7.1	Introduction	68
7.2	Data infrastructure Dairy Campus	68
7.3	Data for the dairy farm case	69
7.3.1	Animal selection	69
7.3.2	Output	69
7.3.3	Input	69
7.3.4	Data availability	69
7.4	Internet access and expected infrastructure	72
8.	General discussion and insights	74
	Literature	78

Preface

New technologies become available to support agro and food production. New concepts like Smart Farming are driving innovations and the way we work together. Also research has to adapt and be aware what these new technologies, like Big Data and Internet of Things, and new concepts can bring them. There is an increasing need for cooperation according triple helix concepts and multidisciplinary research to tackle complex issues. The basis for this project is coming from the idea of the Ministry of Economic Affairs who would like to stimulate TO2 institutes to work more together. We thank the ministry to support the creation of an arena where the TO2 institutes DLO, TNO and NLR could work together and explore the field of BigData and concepts like Smart Farming.

The authors

Summary

In this report we describe results of a one-year TO2 institutes project on the development of big data technologies within the milk production chain. The goal of this project is to 'create' an integration platform for big data analysis for smart farming and to develop a show case. This includes both technical (hard/software) and organizational integration (developing business ecosystem) and combining and linking of data and models. DLO, NLR and TNO worked together in 2015 towards the realization of an IT data infrastructure that makes it possible to solve to connect data from different sources and models in an effective and safe way, ontology problems, specific analysis tools develop, opportunities and risks to identify and assess the acquired knowledge and experience and present it in a smart farming show case, from 'grass to glass'.

In the project we combine domain specific databases with generic ICT- tools and -infrastructure. A data transfer agreement was agreed upon the use of data in the project and to safeguard IP of third parties. NLR contributed data and expertise on (analysis of) satellite and drone data to the project. WLR and PRI contributed domain knowledge, databases (e.g. on genotype and phenotype of cows, feed uptake, milk production and crop management) and models, LEI contributed socio-economic expertise on big data, and TNO provided expertise on ICT-infrastructure, data protection, ontology (in cooperation with PRI and WLR) and analysis tools.

The project was organised in work packages: (1) literature study on big data and smart farming that is relevant to the project, (2) case definition plus outline of big data research questions and required data-infrastructure, (3) semantics, ontology and Linked Open Data (LOD), (4) analysis tools, and (5) synthesis plus answering of the research questions.

(1) Literature study on big data and smart farming that is relevant to the project

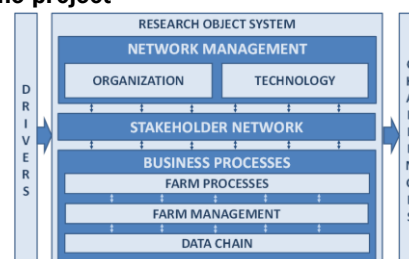
For the literature research a conceptual framework (see figure) was used to analyse both scientific and some grey literature. Based on the findings from the literature review several conclusions can be drawn on the state-of-the-art of Big Data applications in Smart Farming. First of all, Big Data in Smart Farming is still in an early development stage. This is based on the fact there are only limited scientific publications available on this topic and much information had to be derived from 'grey literature'.

- *What role does Big Data play in Smart Farming?*

Big Data is changing the scope and organisation of farming through a pull-push mechanism. Global issues such as food security and safety, sustainability and as a result efficiency improvement are tried to be addressed by Big Data applications. These issues make that the scope of Big Data applications extends far beyond farming alone, but covers the entire supply chain. The Internet of Things development, wirelessly connecting all kind of objects and devices in farming and the supply chain, is producing many new data that are real-time accessible. Analytics is a key success factor to create value out of these data. Many new and innovative start-up companies are eager to sell and deploy all kind of applications to farmers of which the most important ones are related to sensor deployment, benchmarking, predictive modelling and risk management.

- *What stakeholders are involved and how are they organized?*

There are first of all the traditional players in agriculture such as input suppliers and technology suppliers for which there is a clear move towards Big Data as their most important business model. Most of them are pushing their own platforms and solutions to farmers, which are often proprietary and rather closed environments although a tendency towards more openness is observed. This is stimulated by farmers - organized in cooperatives or coalitions - that are concerned about data privacy and security and also want to create value with their own data or at least want to benefit from Big Data solutions. Beside the traditional players we see that Big Data is also attracting many new entrants which are often start-ups supported by either large private investors or large ICT or non-agricultural tech companies. Also public institutions aim to open up public data that can be combined with private data. These developments raise issues around data ownership, value of data and privacy and security. The architecture and infrastructure of Big Data solutions are also significantly determining how stakeholder networks are organized. On the one hand there is a tendency towards closed, proprietary systems and on the other hand towards more open systems based on open source,



standards and interfaces. Further development of Big Data applications may therefore likely effect two supply chain scenarios: one with further integration of the supply chain in which farmers become franchisers; another in which farmers are empowered by Big Data and open collaboration and can easily switch between suppliers, share data with government and participate in short supply chains rather than integrated long supply chains.

- *What are the expected changes that are caused by Big Data developments?*

Big Data will cause major changes in scope and organization of Smart Farming. Business analytics at a scale and speed that was never seen before will be a real game changer, continuously reinventing new business models. It can be expected that farm management and operations will drastically change by access to real-time data, real-time forecasting and tracking of physical items and in combination with IoT developments in further automation and autonomous operation of the farm. It is already visible that Big Data will also cause major shifts in power relationships between the different players in the Big Data farming stakeholder network. The current development stage does however not reveal yet towards which main scenario Smart Farming will be developed.

- *What challenges need to be addressed in relation to the previous questions?*
 - *Data ownership* and related *privacy* and *security* issues – these issues have to be properly addressed, but when this is applied too strictly it can also slow down innovations;
 - *Data quality* - which has always been a key issue in farm management information systems, but is more challenging with big, real-time data;
 - *Intelligent processing* and *analytics* – for Big Data this is also more challenging because of the large amount of often unstructured, heterogeneous data which requires a smart interplay between skilled data scientists and domain experts;
 - *Sustainable integration* of Big Data sources – integration of many different data sources is challenging but because this is crucial for your business model this has to be done in a sustainable manner;
 - *Business models* that are attractive enough for solution providers but that also enable a fair share between the different stakeholders;
 - *Openness of platforms* that will accelerate solution development and innovation in general but also empower farmers in their position in supply chains.

The promise of Big Data in agriculture is alluring, but the challenges above have to be addressed for increased uptake of Big Data applications. Although there are certainly technical issues to be resolved we recommend to focus first on the governance issues that were identified and design suitable business models because these are currently the most inhibiting factors.

(2) Case definition plus outline of big data research questions and required data-infrastructure

The dairy case that was formulated was based on the question 'Is it possible to provide genotypic and phenotypic data to estimate the overall feed efficiency of dairy cows?'. The motivation is that feeding costs are a major part of the total costs for the production of milk. Feeding of milking cows consist general of grass, roughage, concentrates and additives. Most detailed information on cow level is known on concentrates. Grass and roughage intake are poorly available and only on group level. So present feed efficiency is based on concentrate efficiency. Feeding efficiency of all feeding components together is getting more and more important to realize an optimal milk production. With the end of the milk quota in 2015 dairy farming is now more soil-based. Therefore emphasis will be on maximizing milk production given the limited availability of land (and thus grass and roughage). The grass and roughage should be used without wasting. Animal behaviour that implies wasting of feeding should be avoided as much as possible. Therefore in this case we are interested in estimation of feed efficiency for all feed components. The Dairy Campus was used as experimental environment and historical data and infrastructure and observations in 2015 were used in this project. For the data infrastructure of Dairy campus some advices were formulated.

(3) Semantics, ontology and Linked Open Data (LOD)

Theoretically insight are presented on the following aspects of data sets that can be:

- Big: the extent to which data has volume, variety, velocity, veracity.
- Open: the extent to which data is reusable to others.
- Linked: the extent to which data is accessible (linkable) to other data.

Open data is data that "can be freely used, reused and redistributed by anyone. This means that data can be published and can be made publicly available under an open licence without necessarily linking it to other data sources. Linked Data on the other hand, can be linked to URIs from other data sources, using open standards such as RDF without necessarily being publicly available under an open licence. Both types of data may concern big data sets, but of course

this does not have to be the case. Organizing the data and enrich it by adding additional meaning, is the engineering or modelling part of Linked Data. Modelling makes the data more widely understandable and usable both within and across organizations. When creating linked data, one should employ proper engineering practices in order to create datasets of high quality that possibly make use of existing resources on the Web rather than creating them from scratch, and express the intended semantics correctly so that others (both machines and humans) can properly understand and reuse the datasets. Ontologies are a way to making a conceptual model of the data: they are “used to refer to a shared understanding of a domain of interest”. An ontology embodies a view of the domain of interest expressed in terms of concepts, their definitions in terms of properties and their inter-relationships. An ontology may take a variety of forms, but necessarily it will include a vocabulary of terms and some specification of their meaning (i.e. definition).

Data from Dairy Campus and Akkerweb were used to explore these different concepts. Summarizing, farmers can pose questions in terms of the concepts in our common ontology instead of the detailed and specific concepts of the Dairy Campus and Akkerweb data sources. The approach is in an experimental phase. We have reached a set-up by filling the triple stores for 3 farms with cow-data of 1 month which adds up to a total of 7 million triples.

(4) Analysis tools

Roughage intake

In a world where data is growing, the analysis of data moves from traditional methods to ‘big data’ methods, like data mining and machine learning. In this project we are focusing on modelling a complex process using machine learning techniques. The fields of statistics and machine learning are converging. Where data mining focuses on exploring data, finding correlations that are not apparent with traditional techniques, machine learning focuses on modelling the data, and learning it’s patterns. Machine learning algorithms use large amounts of data to learn algorithms that are able to predict the data. In general they require much more data than standard statistical techniques (since no pre-assumptions about the model are being made). An expected advantages of machine learning techniques is that the amount of domain knowledge required is minimal and that the resulting models (when set up and validated correctly) can model very complex (nonlinear) systems. Because it is possible to model non-linear systems, the models are good at generalizing new data points (inter and extrapolating). There are numerous machine learning algorithms available to choose from. We used a Dairy campus dataset, added KNMI data and we experimented with setting up an Artificial Neural Network.

We were able to construct an effective machine learning model to predict the roughage intake of cows. We are able to predict this with a precision of approximately 7.6%. This means that we are able to predict rather accurately the feed intake of cows that do not have labour expensive systems to monitor the roughage intake. These models form a good basis to develop proxies for daily roughage intake of individual dairy cows, based on routinely available data on the dairy farm. These proxies for daily roughage intake can be used in management modules to optimise the feed management of individual or groups of dairy cows or be aggregate to predict full lactation intake, when no or partial feed intake records are available.

Grass intake

In dairy farming grass forms a significant element in the food management and milk production process. The volume and quality of the grass production and consumption (preferably per individual cow) are relevant for the insight in and optimization of the milk production process, but also for the grass production process and in the end also for the investigation of the genotype and phenotype information in the framework of food efficiency and genomic selection. Technology for sensing, location specific treatment and data processing and analysis provide new opportunities for the optimization of the grass production process. In this project research has been done on establishing grass quantity by making use of remote sensing observations. Use has been made of observations from both satellites and drones. For this project a link could be made to a grazing experiment at the Dairy Campus in Leeuwarden. In the framework of this experiment a parcel was subdivided into multiple sub-parcels with varying grass stages. The parcels were observed by a drone on a weekly basis and field measurements of the grass length were collected on a daily basis. Also satellite observations were available at irregular intervals of one to four weeks (depending on the weather circumstances).

Big data techniques are relevant as the volume of remote sensing data tend to be large. Remote sensing data consist of a raster of observations covering the whole area. For each pixel reflections in multiple spectral bands may be recorded. For drone observations photos are collected with resolutions of some centimetres and with 60 to 80%

overlap. As a consequence in this project the data volume of one drone flight over the Dairy Campus is in the order of 1.5Gb. Big Data techniques can help in the management and processing of these data volumes. Secondly Big Data techniques can help in the combined analysis of multiple types of data: satellite and drone raster observations, field point measurements for height and dry weight, weather conditions, information on the grazing characteristics and cow situation. Relations between these features were investigated.

Multispectral photography, either from satellite or from UAS, can be a valuable instrument for the evaluation of the normalized vegetation index (NDVI) as an indicator for the grass biomass. From this study it can be concluded however that for doing this the radiometric and atmospheric calibration of the sensor data is of big importance. Especially because the NDVI variation for the different grass growth stages is only limited (0.6-0.8). Special points of attention for the calibration are the combination of observations from different sensors (relative calibration of different satellites and UAS) and the radiometric calibration and mosaicking of UAS data, especially when flying under cloudy circumstances. Strategies for calibration and usage of ground truth should be worked out and tested further.

Stereo photography from UAS can be a valuable instrument for the measurement of the actual grass heights. Important however is that the photos will be acquired with sufficient spatial resolution (cm level) and that attention is paid to accurate georeferencing (either by using accurate ground control points and/or by using RTK GPS positioning of the UAS). With the current 12cm photos relative variations in grass height could be observed, but the resolution was too coarse to extract accurate and reliable grass height information. As grass height variations in the order of 1cm are relevant, photos with a pixel size of 1cm need to be acquired which is well possible with current technology. An interesting and promising aspect is that with a single UAS stereo photo flight both information can be obtained on grass NDVI and grass height. The combination of these can be input to obtain more reliable values on the grass biomass.

Big data will offer capabilities for the processing of large amounts of data and for the analysis of relations between the measured and other available parameters. The current experiment showed that the amount of remote sensing data easily is in the order of Gigabytes per observation day. When the methodology is expanded to more parcels, farms and higher spatial detail this is scaled up further. Experiments with using Spark/Hadoop technology for raster data management and processing showed that scaling of the processing times is well possible.

(5) synthesis plus answering of the research questions

The creation of an integration platform for big data analysis for smart farming was far too ambitious for a one year activity with emphasis only on one use case. However, by doing this project we brought different kind of expertise and background together to explore parts of the solution. As described in the chapters it is much more complex than expected. Smart farming puts emphasis on the ICT- and decision making part of precision farming. It looks like a buzzword used to describe the concept of intelligent use of data-rich ICT-services and ICT-applications. It is presented as an extra on top of the concepts of precision agriculture and precision livestock farming. Smart farming has the potential to contribute to more sustainable agriculture. And big data use, if established, will support smart decisions and management. Organisations are evaluating to invest in big data technology and use? Big data technology represents a disruptive innovation that market orientated organisations will use to drive competitive advantage and governmental bodies to set and reach policy targets. The value of big data lies in the information and insight that organisations can draw from it, rather than in the data itself. Linking physical and socio-economic data, for example, may generate entirely new insights and market opportunities. So, the impact of big data for smart farming outreaches the impact of a single farmer and his processes, although at farm level maybe most of the big data will be created by implementing new sensing techniques that are high data sensitive.

The perspective arises that use of big data technology has the potential to dramatically change organisations. It will alter data availability, knowledge creation, decision making, production optimization and competitiveness. A fair question is: "Will big data bring real benefits to organisations and society, or will it end up a hype with only one or two companies benefiting from it?". We can answer this question within ca. 20 years from now. Presently, the authors of this report have positive expectations of big data applications in the field of smart farming.

1. Introduction

1.1 Rationale

FAO (2009) stated that 'a more sustainable agriculture' is one of the big challenges of global human population. Agriculture has to produce more food, feed, fuel, flowers, etc., with less use of natural resources and with less adverse side effects on the environment and society and the expectations are that precision agriculture and the acronym smart farming will be necessary to achieve this. More with less, is also a credo of Wageningen UR. Big data applications are likely to contribute to more efficient agro-food chains, and so likely to contribute to a more sustainable agriculture. This report describes a study on the application of big data technology in the context of smart farming in agriculture.



Big data is a buzzword used to describe a massive volume of both structured and unstructured data that is so large that it is difficult to process using traditional database and software techniques. Many public and private organisations are currently evaluating the potential of big data. And a small part of them is already investing in development and implementation of big data technology, or started to create added value with it. The perspective is that use of big data technology has the potential to dramatically change organisations. It will alter data availability, knowledge creation, decision making, production optimization, competitiveness, etc. A fair question is: "Will big data bring real benefits to organisations and society, or will it end up a hype with only one or two companies benefiting from it?" (Needle, 2015). We can answer this question within ca. 20 years from now. Presently, the authors of this report are curious and interested in the potential for big data applications in agriculture.

To make use of the big data mainstream, a lot of technological development and customization still has to take place. Show cases are also needed to demonstrate the benefits and to convince stakeholders. In this report we describe results of a project in which three TO2-institutes ([WUR](#), [TNO](#) and [NLR](#)) in The Netherlands in 2015 worked together on development of big data technologies within a (show)case of smart farming: the milk production chain, with the expectation that this will contribute to more sustainable agriculture.

1.2 Precision farming and smart agri-food chains

Smart farming is a relatively new concept arising within the more established management concepts of precision farming and precision livestock farming. Precision farming is based on the management concept of observing, measuring and responding to inter and intra-field variability in crops, including aspects related to animal rearing (Kempenaar & Kocks, 2013; Lokhorst & Ipema, 2010). The benefits for farmers and for the environment and society are related to increased yields, increased profitability, better working conditions, increased animal health and welfare and reduction of pesticides and chemicals (sustainable production). Smart farming puts emphasis on the ICT- and decision making part of precision farming (Wolfert & Kempenaar, 2012). It is a buzzword used to describe the concept of intelligent use of data-rich ICT-services and ICT-applications. Smart farming has the potential to contribute to more sustainable agriculture. And big data use, if established, will support smart decisions and management.

Smart Farming extends the Precision Agriculture concept: the existing tasks for management and decision making based on data are **enhanced by context, situation and location awareness**. A corresponding task can be related to farm operations, farm logistics, food logistics, stakeholder network, etc. Real-time assisting features are necessary to carry out agile actions, especially in cases of suddenly changed operational conditions or other circumstances (e.g. weather or disease alert). Furthermore, the assisting features typically include intelligent assistance in implementation, maintenance and use of the technology. Figure 1.1 visualizes Smart Farming as a cyber-physical cycle of smart sensing

and monitoring, smart analyses & planning and smart control of farm operations that utilizes Big Data in some way. This concept is explained in more detail in chapter 2.

Over the last decades, we see the following trend in modern agriculture. Productivity development in agriculture has incrementally moved from scaling of assets to optimization of assets. Agricultural technologies as well as overall farm enterprises have grown in size and value over the years and a higher degree of input/output management supported by ICT technology has become essential for optimization of farm profitability and minimization of environmental impact. Optimization parameters such as fuel, labour, fertilizer, pesticides, soil and water preservation relative to yield and quality of the crop are just some of the parameters any arable farmer needs to balance on an operational, tactical and strategic level. And dairy farmers are optimizing milk production by considering the livestock properties (genetics), inputs (feed, water, medicines, housing, etc.) and output (milk, manure, calves,). As input costs are increasing and in some cases input levels are regulated, the cost of making the right or wrong decisions is increasing correspondingly. Furthermore, the dynamic nature of agriculture which is strongly influenced by external factors such as climate volatility and fluctuating crop prices makes these decisions even more difficult. Figure 1.1 visualizes Smart Farming as a cycle of smart sensing and monitoring, smart analysis & planning and smart control of farm operations that utilizes a cloud-based event management system.

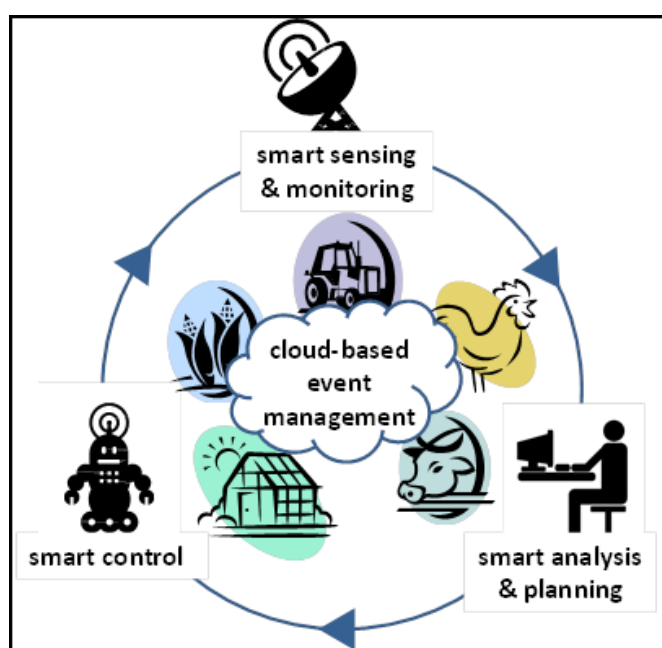


Figure 1.1. The cyber-physical management cycle of Smart Farming enhanced by cloud-based event and data management (Wolfert et al., 2014)

The last decades the relative isolated position of farmers has changed. They have become part of a dynamic chain system in which every part of the chain needs optimization, but also cooperation within the chain needs optimization. Thus farmers have become part of the quality chain control system, of marketing and logistics from basic genetic origin to consumer uptake (Trienekens et al., 2010).

1.3 Big data technology

Big data technology applies to ICT-tools and -infrastructure that allow access to handling, analysis and knowledge creation with big data. In scientific publications, the words electronic science or e-Science are also used when considering this domain. Strictly defined, e-Science is computationally intensive science that involves distributed networks or grid computing (Top et al., 2015, in prep.). E-Science, and also big data technology, allow world-wide collaboration in flexible research teams using advanced, user-friendly and web-based tools, services and repositories. E-

science/Big data technology will change how (big) data are generated, how new information and knowledge is created, and how data, information and knowledge is shared and applied.

In the next chapters more will be explained about what is needed to create value with the use of big data. Terms like data volume, velocity, variety, veracity, variability and value of data (six V's of big data) will be addressed. And data should be FAIR: Findable, Accessible, Interoperable en Reliable. Figure 1.2 depicts the process data make from devices that (autonomously) acquire data to knowledge and value creation to support decisions of users.

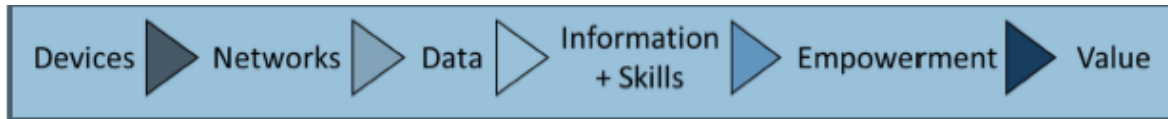


Figure 1.2. Schematic flow of data from devices to added value. Source: TNO.

Why are organisations evaluating to invest in big data technology and use? Big data technology represents a disruptive innovation that market orientated organisations will use to drive competitive advantage and governmental bodies to set and reach policy targets. The value of big data lies in the information and insight that organisation can draw from it, rather than in the data itself. Linking physical and socio-economic data, for example, may generate entirely new insights and market opportunities.

Also TO2-institutes participating in this project invest in big data technology. They formulated their big data strategies in strategy documents or position papers (DLO, 2015; TNO, 2015; NLR). A conceptual view on the big data domain by Wageningen UR is depicted in figure 1.3. Figure 1.3 shows examples of the various types of big data related technologies and methodologies and fields of expertise that are relevant for the subsequent higher levels of the Agri&Food knowledge chain. The big data field focuses on the horizontal as well as the vertical exchange of information and interaction. Therefore, strongly technology related infrastructures and fields of expertise are required as well as knowledge and methods in the field of policy analysis and knowledge brokerage are needed for the contextualization of the available data and gathered knowledge.

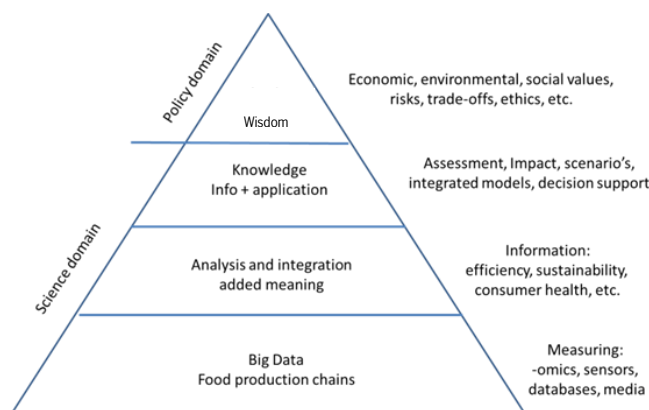


Figure 1.3. Conceptual view on big data domain. Source: DLO

However, there are also risks, for example those related to regulatory hazards and issues such as ownership and privacy. To secure value from big data technologies, organisations need a holistic and strategic plan for identifying opportunities and overcoming hurdles and managing risks.

1.4 Objectives, approach and report structure

In this report we describe results of a one-year TO2 institutes project on the development of big data technologies within the milk production chain. The goal of this project is to 'create' an integration platform for big data analysis for smart farming and to develop a show case. This includes both technical (hard /software) and organizational integration

(developing business ecosystem) and combining and linking of data and models. DLO, NLR and TNO worked together in 2015 towards the realization of an IT data infrastructure that makes it possible to solve to connect data from different sources and models in an effective and safe way, ontology problems, specific analysis tools develop, opportunities and risks to identify and assess the acquired knowledge and experience and present it in a smart farming show case, from 'grass to glass'.

In the project we combine domain specific databases with generic ICT- tools and -infrastructure. A data transfer agreement was agreed upon the use of data in the project and to safeguard IP of third parties. NLR contributed data and expertise on (analysis of) satellite data to the project. WLR and PRI contributed domain knowledge, databases (e.g. on genotype and phenotype of cows, feed uptake, milk production and crop management) and models, LEI contributed socio-economic expertise on big data, and TNO provided expertise on ICT-infrastructure, data protection, ontology (in cooperation with PRI and WLR) and analysis tools.

The project was organised in work packages: (1) literature study on big data and smart farming that is relevant to the project, (2) case definition plus outline of big data research questions and required data-infrastructure, (3) semantics, ontology and Linked Open Data (LOD), (4) analysis tools, and (5) synthesis plus answering of the big data questions. The conceptual framework is depicted in figure 1.4. Each work packages was led by a coordinator of one of the three institutes most affiliated with the core of the work package. In the chapters 2 to 5 we describe the relevant state of the art of the literature on big data use (chapter 2), the smart dairy farming case, research questions and requirements (chapter 3), semantic challenges and solutions (Chapter 4) and analysis tools (chapter 5 for cow centric data and chapter 6 for remote sensing data for grassland). In chapter 7, we describe what is achieved in one year time by bringing together the different databases and ICT-infrastructure and tools, in order to answer big data research questions in the milk production chain. Chapter 8 describes the general lessons learned.

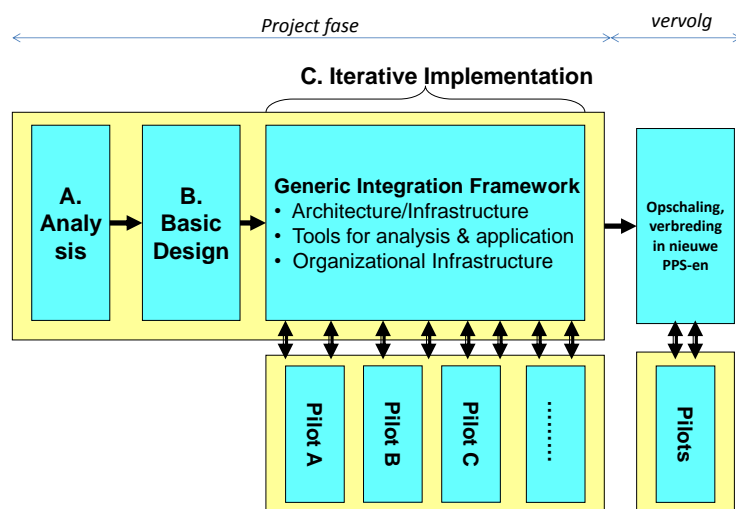


Figure 1.4 Project framework for integration.

2. Literature study

2.1 Introduction

As smart machines and sensors crop up on farms and farm data grow in quantity and scope, farming processes will become increasingly data-driven and data-enabled. Rapid developments in the Internet of Things and Cloud Computing are propelling the phenomenon of what is called Smart Farming (Sundmaeker et al., 2016). Smart Farming goes beyond the concept of Precision Agriculture by basing management tasks not only on location but also on data, enhanced by context- and situation awareness, triggered by real-time events (Wolfert et al., 2014). Real-time assisting reconfiguration features are required to carry out agile actions, especially in cases of suddenly changed operational conditions or other circumstances (e.g. weather or disease alert). These features typically include intelligent assistance in implementation, maintenance and use of the technology. Figure 1.1 summarizes the concept of Smart Farming along the management cycle as a cyber-physical system. In this picture it is already suggested that robots can play an important role in control, but it can be expected that the role of humans in analysis and planning is increasingly assisted by machines so that the cyber-physical cycle becomes almost autonomous. Humans will always be involved in the whole process but increasingly at a much higher intelligence level, leaving most operational activities to machines.

Big Data technologies are playing an essential, reciprocal role in this development: machines are equipped with all kind of sensors that measure data in their environment that is used for the machines' behaviour. This varies from relatively simple feedback mechanisms (e.g. a thermostat regulating temperature) to deep learning algorithms (e.g. to implement the right crop protection strategy). This is leveraged by combining with other, external Big Data sources such as weather or market data or benchmarks with other farms. Due to rapid developments in this area, a unifying definition of Big Data is difficult to give, but generally it is a term for data sets that are so large or complex that traditional data processing applications are inadequate (Wikipedia, 2016). Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale (Hashem et al., 2015). Big Data represents the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value (De Mauro et al., 2016). The Data FAIRport initiative emphasizes the more operational dimension of Big Data by providing the FAIR principle meaning that data should be Findable, Accessible, Interoperable and Re-usable (Data FAIRport, 2014). This also implies the importance of metadata *i.e.* 'data about'.

Both Big Data and Smart Farming are relatively new concepts, so it is expected that knowledge about their applications and their implications for research and development is not widely spread. Some authors refer to the advent of Big Data and related technology as another technology hype that may fail to materialize, others consider Big Data applications may have passed the 'peak of inflated expectations' in Gartner's Hype Cycle (Fenn and LeHong, 2011; Needle, 2015). This literature study aims to provide insight into the state-of-the-art of Big Data applications in relation to Smart Farming and to identify the most important research and development challenges to be addressed in the future. In studying the literature, attention is paid to both technical and socio-economic aspects. In the analysis there is a primary focus on the socio-economic impact Big Data will have on farm management and the whole network around it. The research questions to be addressed are:

1. What role does Big Data play in Smart Farming?
2. What stakeholders are involved and how are they organized?
3. What are the expected changes that are caused by Big Data developments?
4. What challenges need to be addressed in relation to the previous questions?

To answer these questions and to structure the review process, a conceptual framework for analysis has been developed, which is expected to be useful also for future analyses of developments in Big Data and Smart Farming.

2.2 Methodology

To address the research questions as outlined in the introduction, we surveyed literature between January 2010 and March 2015. This was done in three steps. In the first step we searched two major bibliographical databases, Web of

Science and Scopus, using all combinations of two groups of keywords of which the first group addresses Big Data (*i.e.* Big Data, data-driven innovation, data-driven value creation, internet of things, IoT) and the second group refers to farming (*i.e.* agriculture, farming, food, agri-food, precision agriculture). From these two databases 613 peer-reviewed articles were retrieved. These were scanned for relevance by identifying passages that were addressing the research questions. As a result, 20 were considered most relevant and 94 relevant. The remaining articles were considered not really relevant as they only tangentially touch upon Big Data or agriculture and therefore excluded from further reading and analysis. We found the number of relevant peer-reviewed literature not very high which can be explained because Big Data and Smart Farming are relatively new concepts. Especially the applications are rapidly evolving and expected not to be taken into account in peer-reviewed articles which are usually lagging behind. Therefore we decided to also include grey literature into our review. For that purpose we have used Google Scholar and the search engine LexisNexis for reports, magazines, blogs, and other web-items in English. This has resulted in 3 reports, 225 magazine articles, 319 blogs and 19 items on twitter. Each of the 319 blogs was evaluated on relevance based on its title and sentences containing the search terms. Also possible duplications were removed. The result was a short list containing 29 blogs that were evaluated by further reading. As a result, 9 blogs have been considered as presenting relevant information for our framework. Each of the 225 magazine articles was similarly evaluated on their relevance based on its title and sentences containing the search terms. After removing duplicates, the result is a short list of 25 articles. These 25 articles were then read through for further evaluation. Consequently 9 articles have been considered as containing relevant information for further analysis.

In the second step, we read the selected literature in detail to extract the information relevant to our research questions. Additional literature that had not been identified in the first step was retrieved in this step as well if they were referred to by the 'most relevant' literature. This 'snow-ball' approach has resulted in 11 additional articles and web-items from which relevant information was extracted as well. In the third step, the extracted information was analysed and synthesized following the conceptual framework as described in Section 2.3.

2.3 Conceptual framework

For this literature study a conceptual framework was developed to provide a systematic classification of issues and concepts for the analysis of Big Data applications in Smart Farming from a socio-economic perspective. This framework draws upon literature on network management and data-driven strategies.

The often-cited conceptual framework of Lambert & Cooper (2000) on network management comprises three closely interrelated elements: the network structure, the business processes, and the management components. The network structure consists of the member firms and the links between these firms. Business processes are the activities that produce a specific output of value to the customer. The management components are the managerial variables by which the business processes are integrated and managed across the network. The network management component is further divided into a technology and organisation component.

For our purpose the framework was tailored to networks for Big Data applications in Smart Farming as presented in Figure .1.

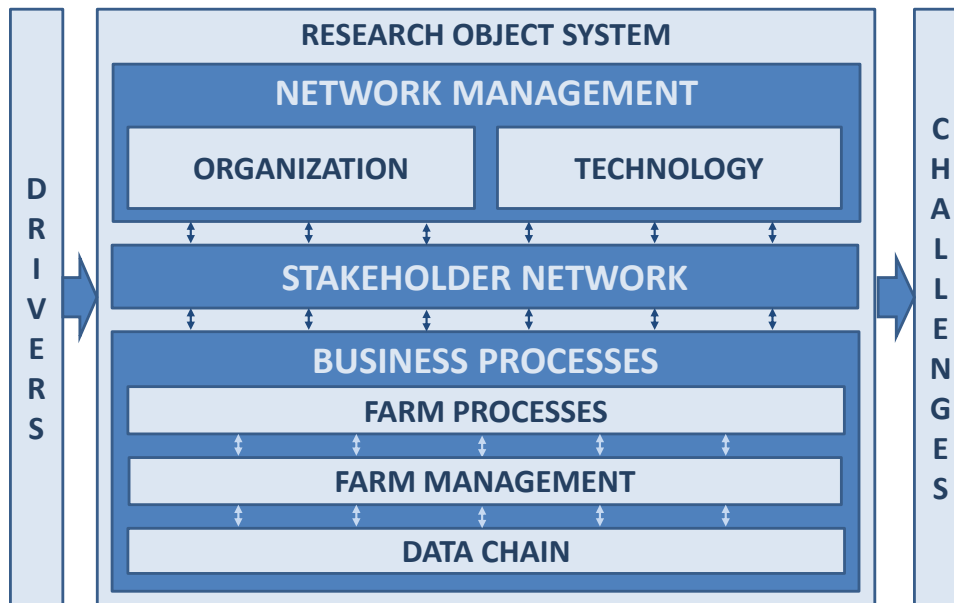


Figure 2.1 Conceptual framework for the literature analysis (adapted from Lambert and Cooper (2000))

In this object system for research, the business processes (lower layer) focus on the generation and use of Big Data in the management of farming processes. For this reason, we subdivided this part into the data chain, the farm management and the farm processes. The data chain interacts with farm processes and farm management processes through various decision making process in which information plays an important role. The stakeholder network (middle layer) comprises all stakeholders that are involved in these processes, not only users of Big Data but also companies that are specialised in data management. The technology component of network management (upper layer) focuses on the information infrastructure that supports the data chain. The organisational component focuses on the governance and business model of the data chain. Finally, several factors can be identified as key drivers for the development of Big Data in Smart Farming and as a result challenges can be derived from this development.

The next subsections provide a more detailed description of each subcomponents of the business processes layer and network management layer of the framework.

2.3.1 Farm processes

A business process is a set of logically related tasks performed to achieve a defined business outcome (Davenport and Short, 1990). An important foundation of business process approaches was laid by Porter (1985), who introduced the term 'value chain'. A firm's value chain is a system of interlinked processes, each adding value to the product or service. Based on this principle, business processes can be subdivided into primary and supporting business processes (Davenport, 1993; Porter, 1985). *Primary Business Processes* are those involved in the creation of the product, its marketing and delivery to the buyer (Porter, 1985). *Supporting Business Processes* facilitate the development, deployment and maintenance of resources required in primary processes.

2.3.2 Farm management

Management or control processes ensure that the business process objectives are achieved, even if disturbances occur. The basic idea of control is the introduction of a controller that measures system behaviour and corrects if measurements are not compliant with system objectives. Basically, this implies that they must have a feedback loop in which a norm, sensor, discriminator, decision maker, and effector are present (Beer, 1981; in 't Veld, 2002). As a consequence, the basic management functions are (Verdouw et al., 2015) (see also Figure 1.1):

- *Sensing and monitoring*: measurement of the actual performance of the object system. This can be done manually by a human observer or automated by using sensing technologies such as sensors or satellites. In addition, external data can be acquired to complement direct observations.

- *Analysis and decision making*: compares measurements with the norms that specify the desired performance (system objectives concerning e.g. quantity, quality and lead time aspects), signals deviations and decides on the appropriate intervention to remove the signalled disturbances.
- *Intervention*: plans and implements the chosen intervention to correct the object system's performance.

2.3.3 Data chain

The data chain refers to the sequence of activities from data capture to data marketing. Figure 2.2 illustrates the main steps in this chain.

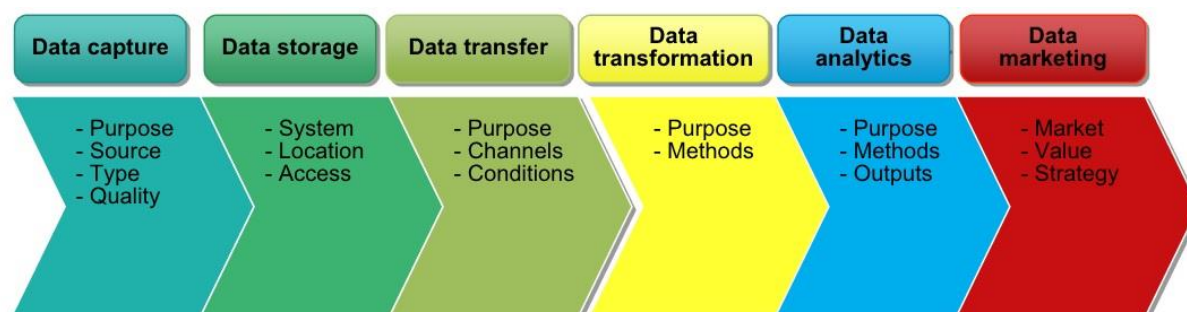


Figure 2.2 The data chain of Big Data applications, based on Chen et al. (2014b)

Being an integral part of business processes, the data chain consists necessarily of a technical layer that captures raw data and converts it into information and a business layer that makes decision and derives value from providing data services and business intelligence. The two layers can be interwoven in each stage and together they form the basis of what has come to be known as the 'data value chain' (Dumbill, 2014) (Table 2.1).

Table 2.1 Key stages of the data chain on technical and business layer

Layer of data chain	Stages of a data chain			
	Raw material	Processing	Transport	Marketing
Technical	Data generation and capture	Data janitorial work, Data transformation Data analytics	Data transfer	Data transfer Data analytics
Business	Data discovery Data warehousing	Interpreting data, Connecting data to decision (Obtaining business information and insight)	Information share and data integration	Data-driven services

2.3.4 Network management organization

The network management organisation deals with the behaviour of the stakeholders and how it can be influenced to accomplish the business process objectives. For the uptake and further development of Big Data applications, two interdependent aspects are considered relevant: governance and business model. Governance involves the formal and informal arrangements that govern cooperation within the stakeholder network. Lambert & Cooper (2000) identify the following components: management methods, power and leadership structure, risk and reward structure and culture and attitude. This includes the governance structure, which is a major issue in Network Theory. Three basic forms of network governance can be distinguished (Lazzarini et al., 2001):

- *Managerial Discretion*: discretionary actions by a coordinating agent, who centrally plans the flow of products and information;
- *Standardization*: standardized rules and shared mechanisms to orchestrate transactions;
- *Mutual Adjustment*: alignment of plans through mutual feedback processes and joint problem solving and decision making.

These forms correspond with the three forms of network governance presented by Provan and Kenis (2008): lead organisation-governed network, network administrative organisation, and shared participant-governed network. The choice of a particular network governance structure aims at mitigating all forms of contractual hazards found between the different contracting parties in such a way that transaction costs are minimized (Williamson, 1996). When studying hybrid forms of organization such as supply chain networks, two main dimensions should be identified: the allocation of decision rights, i.e., who has the authority to take strategic decisions within the supply chain network, and the inter-organizational mechanisms aiming at rewarding desirable behaviour and preventing undesirable behaviour (risk and rewarding mechanisms).

Despite agreement on the importance of business model to an organization's success, the concept is still fuzzy and vague, and there is little consensus regarding its compositional facets. Osterwalder (2004) defines business model as "... a conceptual tool that contains a set of elements and their relationships and allows *expressing a company's logic of earning money*". It is a description of the value a company offers to one or several segments of customers and the architecture of the firm and its network of partners for creating, marketing and delivering this value and relationship capital, in order to generate profitable and sustainable revenue streams." This definition reflects a so-called firm-centric view of business model. Another view on business model is the network-centric business model which builds upon value network theories (Al-Debei and Avison, 2010). The value network theories consider both financial and non-financial value of business transactions and exchanges. Both views are relevant to the network management of Big Data applications.

2.3.5 Network management technology

The network management technology includes all computers, networks, peripherals, systems software, application packages (application software), procedures, technical, information and communication standards (reference information models and coding and message standards) etc., that are used and necessary for adequate data management in the inter-organizational control of farming processes (van der Vorst et al., 2005). Components to be mentioned here encompass:

- Data resources stored in shared databases and a shared understanding of its content (shared data model of the database).
- Information systems and services that allow us to use and maintain these databases. An information system is used to process information necessary to perform useful activities using activities, facilities, methods and procedures.
- The whole set of formalised coding and message standards (both technically and content-wise) with associated procedures for use, connected to shared databases, which are necessary to allow seamless and error-free automated communication between business partners in a food supply chain network.
- The necessary technical infrastructure. None of the above can work if we don't have the connected set of computers (workstations of individual associates or people employed by or interested in the network and the database, communication and application servers and all associated peripherals) that will allow for its usage.

In conclusion, this framework now provides a coherent set of elements to describe and analyse the developments of Big Data in Smart Farming. The results are provided in chapter 2.4.

2.4 Results

2.4.1 Drivers for Big Data in Smart Farming

There has been a significant trend to consider the application of Big Data techniques and methods to agriculture as a major opportunity for application of the technology stack, for investment and for the realisation of additional value within the agri-food sector (Noyes, 2014; Sun et al., 2013; Yang, 2014). Big data applications in farming are not strictly about primary production, but play a major role in improving the efficiency of the entire supply chain and alleviating food security concerns (Chen et al., 2014a; Esmeyjer et al., 2015; Gilpin, 2015). Big data is the focus of in-depth, advanced, game-changing business analytics, at a scale and speed that the old approach of copying and cleansing all of it into a data warehouse is no longer appropriate (Devlin, 2012). Opportunities for Big Data applications in agriculture include benchmarking, sensor deployment and analytics, predictive modelling, and using better models to manage crop failure risk and to boost feed efficiency in livestock production (Faulkner and Cebul, 2014; Lesser, 2014). In conclusion, Big Data is to provide predictive insights to future outcomes of farming (predictive yield model, predictive feed intake

model, etc.), drive real-time operational decisions, and reinvent business processes for faster, innovative action and game-changing business models (Devlin, 2012). Decision-making in the future will be a complex mix of human and computer factors (Anonymous, 2014b). Big data is expected to cause changes to both the scope and the organisation of farming (Poppe et al., 2015). While there are doubts whether farmers' knowledge is about to be replaced by algorithms, Big Data applications are likely to change the way farms are operated and managed (Drucker, 2014). Key areas of change are real-time forecasting, tracking of physical items, and reinventing business processes (Devlin, 2012). Wider uptake of Big Data is likely to change both farm structures and the wider food chain in unexplored ways as what happened with the wider adoption of tractor and the introduction of pesticides in the 1950s. As with many technological innovations changes by Big Data applications in Smart Farming are driven by push-pull mechanisms. Pull, because there is a need for new technology to achieve certain goals. Push, because new technology enables people or organisations to achieve higher or new goals. This will be elaborated in the next subsections.

Pull factors

From a business perspective, farmers are seeking ways to improve profitability and efficiency by on the one hand looking for ways to reduce their costs and on the other hand obtaining better prices for their product. Therefore they need to take better, more optimal decisions and improve management control. While in the past advisory services were based on general knowledge that once was derived from research experiments, there is an increasing need for information and knowledge that is generated on-farm in its local-specific context. It is expected that Big Data technologies help to achieve these goals in a better way (Poppe et al., 2015; Sonka, 2015). A specific circumstance for farming is the influence of the weather and especially its volatility. Local-specific weather and climate data can help decision-making a lot (Lesser, 2014). A general driver can be the relief of paper work because of all kind of regulations in agri-food production (Poppe et al., 2015).

From a public perspective global food security is often mentioned as a main driver for further technological advancements (Gilpin, 2015; Lesser, 2014; Poppe et al., 2015). Besides, consumers are becoming more concerned about food safety and nutritional aspects of food related to health and well-being (Tong et al., 2015). In relation to that, Tong et al. (2015) mention the need for early warning systems instead of many ex-post analyses that are currently being done on historical data.

Push factors

A general future development is the Internet of Things (IoT) in which all kinds of devices – smart objects - are connected and interact with each other through local and global, often wireless network infrastructures (Porter and Heppelmann, 2014). Precision agriculture can be considered as an exponent of this development and is often mentioned as an important driver for Big Data (Lesser, 2014; Poppe et al., 2015). This is expected to lead to radical changes in farm management because of access to explicit information and decision-making capabilities that were previously not possible, either technically or economically (Sonka and IFAMR, 2014). As a consequence, there is a rise of many ag-tech companies that pushes this data-driven development further (Lesser, 2014).

Wireless data transfer technology also permits farmers to access their individual data from anywhere – whether they are at the farmhouse or meeting with buyers in Chicago – enabling them to make informed decisions about crop yield, harvesting, and how best to get their product to market (Faulkner and Cebul, 2014).

2.4.2 Business processes

Farm processes

Agricultural Big Data are known to be highly heterogeneous (Ishii, 2014; Li et al., 2014). Data collected from the field or the farm include information on planting, spraying, materials, yields, in-season imagery, soil types, weather, and other practices. Table 1.2 provides an overview of current Big Data applications in relation to different elements of Smart Farming in key farming sectors.

Table 1.2 Examples of Big Data applications/aspects in different Smart Farming processes (cf. Figure 1.1)

Cycle of Smart Farming	Arable	Livestock	Horticulture	Fishery
Smart sensing and monitoring	Robotics and sensors	Biometric sensing, GPS tracking	Robotics and sensors (temperature, humidity, CO ² , etc.), greenhouse computers	Automated Identification Systems (AIS)
Smart Analysis and Planning	Seeding, Planting, Soil typing, Crop health, yield modelling	Breeding, monitoring	Lighting, energy management	Surveillance, monitoring
Smart Control	Precision farming	Milk robots	Climate control, Precision control	Surveillance, monitoring
Big Data in the Cloud	Weather/climate data, Yield data, Soil types, Market information, agricultural census data	Livestock movements	Weather/climate, market information, social media,	Satellite data, Market data

There are in general three categories of data generation (Devlin, 2012; UNECE, 2013): (i) process-mediated (PM), (ii) machine-generated (MG) and (iii) human-sourced (HS). **PM data**, or the traditional business data, result from agricultural processes that record and monitor business events of interest, such as purchasing inputs, feeding, seeding, applying fertilizer, taking an order, etc. PM data are usually highly structured and include transactions, reference tables and relationships, as well as the metadata that define their context. Traditional business data are the vast majority of what IT managed and processed, in both operational and business information systems, usually structured and stored in relational database systems. **MG data** are derived from the vast increasing number of sensors and smart machines used to measure and record farming processes; this development is currently boosted by what is called the Internet of Things (IoT). MG data range from simple sensor records to complex computer logs and are typically well-structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the farming information stored and processed. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches. For Smart Farming, the potential of unmanned aerial vehicles (UAVs) has been well-recognized (Faulkner and Cebul, 2014; Holmes, 2014). Drones with infrared cameras, GPS technology, are transforming agriculture with their support for better decision making, risk management (Anonymous, 2014c). In livestock farming, smart dairy farms are replacing labour with robots in activities like feeding cows, cleaning the barn, and milking the cows (Anonymous, 2012). On arable farms, precision technology is increasingly used for managing information about each plant in the field (Vogt, 2013). With these new technologies data is not in traditional tables only, but can also appear in other formats like sounds or images (Sonka, 2015). In the meantime several advanced data analysis techniques have been developed that trigger the use of data in images or other formats (Lesser, 2014; Noyes, 2014). **HM data** is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. HM data are usually loosely structured and often ungoverned. In the context of Big Data and Smart Farming, human-sourced data have rarely been discussed except in relation to the marketing aspects (Verhoosel et al., 2016). Limited capacity with regard to the collection of relevant social media data and semantic integration of these data from a diversity of sources is considered to be a major challenge (Bennett, 2015). From the business perspective, the main data products along the Big Data value chain are (predictive) analytics that provide decision support to business processes at various levels. The use or analysis of sensor data or similar data must somehow fit into existing or reinvented business processes. Integration of data from a variety of sources, both traditional and new, with multiple tools, is the first prerequisite.

Farm management

As Big Data observers point out: big or small, Big Data is still data (Devlin, 2012). It must be managed and analysed to extract its full value. Developments in wireless networks, IoT, and cloud computing are essentially only means to obtain data and generate Big Data. The ultimate use of Big Data is to obtain the information or intelligence embodied or enabled by Big Data. Agricultural Big Data will have no real value without Big Data analytics (Sun et al., 2013). To obtain Big Data analytics, data from different sources need to be integrated into 'lagoons of data'. In this process, data quality issues are likely to arise due to errors and duplications in data. As shown in Figure 2.3, a series of operations on the raw data may be necessary to ensure the quality of data.

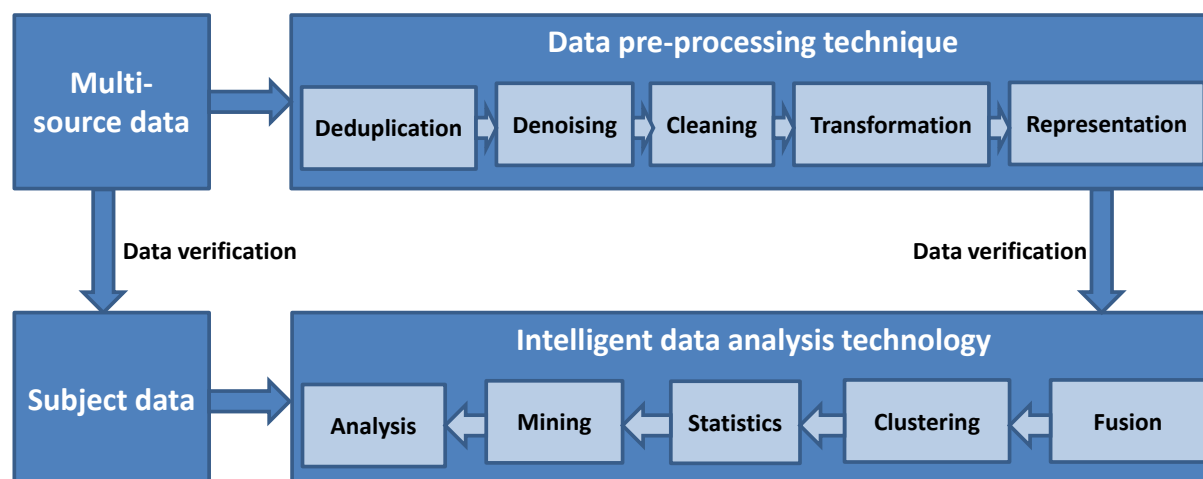


Figure 2.3 The flowchart of intelligent processing of agricultural Big Data (Source: Li, Chen, & Guo, (Li et al., 2014)

Since the advent of large-scale data collections or warehouses, the so-called data rich, information poor (DRIP) problems have been pervasive. The DRIP conundrum has been mitigated by the Big Data approach which has unleashed information in a manner that can support informed - yet, not necessarily defensible or valid - decisions or choices. Thus, by somewhat overcoming data quality issues with data quantity, data access restrictions with on-demand cloud computing, causative analysis with correlative data analytics, and model-driven with evidence-driven applications (Tien, 2013). Big data on its own can offer 'a-ha' insights, but it can only reliably deliver long-term business advantage when fully integrated with traditional data management and governance processes (Devlin, 2012). Big Data processing depends on traditional, process-mediated data and metadata to create the context and consistency needed for full, meaningful use. The results of Big Data processing must be fed back into traditional business processes to enable change and evolution of the business.

Table 2.2 State of the art of Big Data applications in Smart Farming and key issues

Stage of the data	State of the art	Key issues
Data capture	Sensors, Open data, data captured by UAVs, Biometric sensing, Genotype information, Reciprocal data	Availability, quality, formats
Data storage	Cloud-based platform, Hadoop Distributed File System (HDFS), hybrid storage systems, cloud-based data warehouse	Quick and safe access to data, costs
Data transfer	Wireless, cloud-based platform, Linked Open Data	Safety, agreements on responsibilities and liabilities, costs, wireless network
Data transformation	Machine learning algorithms, normalize, visualize, anonymize	Heterogeneity of data sources, automation of data cleansing and preparation
Data analytics	Yield models, Planting instructions, Benchmarking, Decision ontologies, Cognitive computing	Semantic heterogeneity, real-time analytics, scalability
Data marketing	Data visualisation	Ownership, privacy, new business models

Data chain

Table 2.2 summarizes the state-of-the-art features of Big Data applications in Smart Farming and the key issues corresponding to each stage of the Big Data chain that were found in literature.

2.4.3 Stakeholder network

In view of the technical changes brought forth by Big Data and Smart Farming, we seek to understand the stakeholder network around the farm. The literature suggests major shifts in roles of and power relations among different players in existing agri-food chains. We observed the changing roles of old and new software suppliers in relation to Big Data and farming and emerging landscape of data-driven initiatives with prominent role of big tech and data companies like Google and IBM. In Figure 2.4, the current landscape of data-driven initiatives is visualized.

The stakeholder networks exhibits a high degree of dynamics with new players taking over the roles played by other players and the incumbents assuming new roles in relation to agricultural Big Data. As opportunities for Big Data have surfaced in the agribusiness sector, big agriculture companies such as Monsanto and John Deere have spent hundreds of millions of dollars on technologies that use detailed data on soil type, seed variety, and weather to help farmers cut costs and increase yields (Faulkner and Cebul, 2014). Other players include various accelerators, incubators, venture capital firms, and corporate venture funds (Monsanto, DuPont, Syngenta, Bayer, DOW etc.) (Lane, 2015).

Monsanto has been pushing big-data analytics across all its business lines, from climate prediction to genetic engineering. It is trying to persuade more farmers to adopt its cloud services. Monsanto says farmers benefit most when they allow the company to analyse their data - along with that of other farmers - to help them find the best solutions for each patch of land (Guild, 2014).

While corporates are very much engaged with Big Data and agriculture, start-ups are at the heart of action, providing solutions across the value chain, from infrastructure and sensors all the way down to software that manages the many streams of data from across the farm. As the ag-tech space heats up, an increasing number of small tech start-ups are launching products giving their bigger counterparts a run for their money. In the USA, start-ups like FarmLogs (Guild, 2014), FarmLink (Hardy, 2014) and 640 Labs challenge agribusiness giants like Monsanto, Deere, DuPont Pioneer (Plume, 2014). One observes a swarm of data-service start-ups such as FarmBot (an integrated open-source precision agriculture system) and Climate Corporation. Their products are powered by many of the same data sources, particularly those that are freely available such as from weather services and Google Maps. They can also access data gathered by farm machines and transferred wirelessly to the cloud. Traditional agri-IT firms such as NEC and Dacom are active with a precision farming trial in Romania using environmental sensors and Big Data analytics software to maximize yields (NEC, 2014).

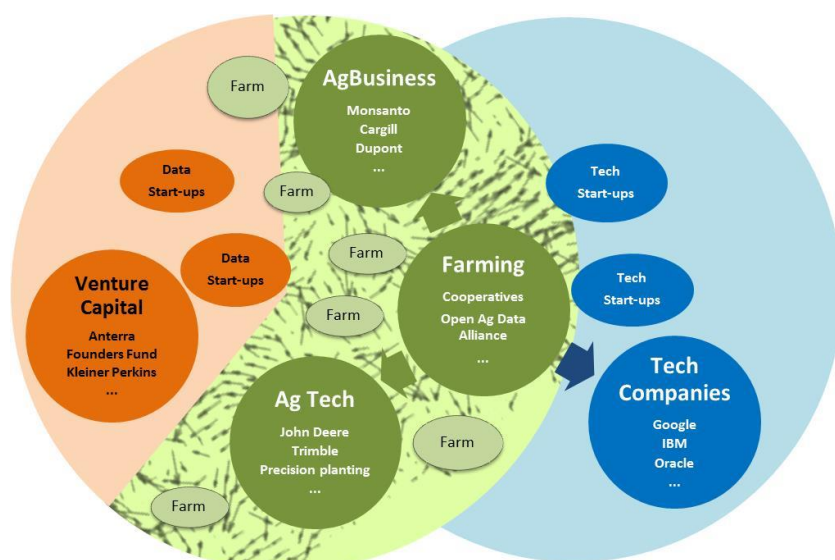


Figure 2.4 The landscape of the Big Data network with business players.

Venture capital firms are now keen on investing in agriculture technology companies such as Blue River Technology, a business focusing on the use of computer vision and robotics in agriculture (Royse, 2014). The new players to Smart Farming are tech companies that were traditionally not active in agriculture. For example, Japanese technology firms such as Fujitsu are helping farmers with their cloud based farming systems (Anonymous, 2014c). Fujitsu collects data (rainfall, humidity, soil temperatures) from a network of cameras and sensors across the country to help farmers in Japan better manage its crops and expenses (Carlson, 2012). Data processing specialists are likely to become partners of producers as Big Data delivers on its promise to fundamentally change the competitiveness of producers. Beside business players such as corporates and start-ups, there are many public institutions (e.g., universities, USDA, the American Farm Bureau Federation, GODAN) that are actively influencing Big Data applications in farming through their advocacy on open data and data-driven innovation or their emphasis on governance issues concerning data ownership and privacy issues. Well-known examples are the Big Data Coalition, Open Agriculture Data Alliance (OADA) and AgGateway. Public institutions like the USDA, for example, want to harness the power of agricultural data points created by connected farming equipment, drones, and even satellites to enable precision agriculture for policy objectives like food security and sustainability. Precision farming is considered to be the “holy grail” because it is the means by which the food supply and demand imbalance will be solved. To achieve that precision, farmers need a lot of data to inform their planting strategies. That is why USDA is investing in big, open data projects. It is expected that open data and Big Data will be combined together to provide farmers and consumers just the right kind of information to make the best decisions (Semantic Community, 2015).

2.4.4 Network management

Organization

Data ownership is an important issue in discussions on the governance of agricultural Big Data generated by smart machinery such as tractors from John Deere (Burrus, 2014). In particular, value and ownership of precision agricultural data have received much attention in business media (Haire, 2014). It has become a common practice to sign Big Data agreements on ownership and control data between farmers and agriculture technology providers (Anonymous, 2014a). Such agreements address questions such as: How can farmers make use of Big Data? Where does the data come from? How much data can we collect? Where is it stored? How do we make use of it? Who owns this data? Which companies are involved in data processing?

There is also a growing number of initiatives to address or ease privacy and security concerns. For example, the Big Data Coalition, a coalition of major farm organizations and agricultural technology providers in the USA, has set principles on data ownership, data collection, notice, third-party access and use, transparency and consistency, choice, portability, data availability, market speculation, liability and security safeguards (Haire, 2014). And AgGateway, a non-profit organization with more than 200 member companies in the USA, have drawn a white paper that presents ways to incorporate data privacy and standards (AgGateway, 2014). It provides users of farm data and their customers with issues to consider when establishing policies, procedures, and agreements on using that data instead of setting principles and privacy norms.

The ‘Ownership Principle’ of the Big Data Coalition states that “We believe farmers own information generated on their farming operations. However, it is the responsibility of the farmer to agree upon data use and sharing with the other stakeholders (...).” While having concerns about data ownership, farmers also see how much companies are investing in Big Data. In 2013, Monsanto paid nearly 1 billion US dollars to acquire The Climate Corporation, and more industry consolidation is expected. Farmers want to make sure they reap the profits from Big Data, too. Such change of thinking may lead to new business models that allow shared harvesting of value from data.

Big data applications in Smart Farming will potentially raise many power-related issues (Orts and Spigonardo, 2014). There might be companies emerging that gain much power because they get all the data. In the agri-food chain these could be input suppliers or commodity traders, leading to a further power shift in market positions (Lesser, 2014). This power shift can also lead to potential abuses of data e.g. by the GMO lobby or agricultural commodity markets or manipulation of companies (Noyes, 2014). Initially, these threats might not be obvious because for many applications small start-up companies with hardly any power are involved. However, it is a common business practice that these are acquired by bigger companies if they are successful and in this way the data still gets concentrated in the hands of one big player (Lesser, 2014). Gilpin (2015), for example, concluded that Big Data is both a huge opportunity as a potential threat for farmers.

Technology

To make Big Data applications for Smart Farming work, an appropriate technological infrastructure is essential. Although we could not find much information about used infrastructures in literature it can be expected that the applications from the AgTech and AgBusiness companies in Figure 2.4 are based on their existing infrastructure that is usually supplied by large software vendors. This has resulted in several proprietary platforms such as AGCO's AgCommand, John Deere's FarmSight or Monsanto's FieldScripts. Initially these platforms were quite closed and difficult to connect to by other third parties. However, they increasingly realize to be part of a system of systems (Porter and Heppelmann, 2014) resulting in more open platforms in which data is accessible through open Application Programming Interfaces (APIs). The tech- and data start-ups mainly rely on open standards (e.g. ISOBUS) through which they are able to combine different datasets. Moreover, Farmobile recently introduced a piece of hardware, the passive uplink communicator (PUC), which captures all machine data into a database that can be transmitted wirelessly (Young, 2016). In North America, several initiatives are undertaken to open up data transfer between several platforms and devices. The ISOBlue project facilitates data acquisition through the development of an open-source hardware platform and software libraries to forward ISOBUS messages to the cloud and develop applications for Android smartphones (Layton et al., 2014). The Open Ag Toolkit (OpenATK) endeavours to provide a specialized Farm Management Information System incorporating low-cost, widely available mobile computing technologies, internet-based cloud storage services, and user-centred design principles (Welte et al., 2013). One of the internet-based cloud storage services that is candidate in the OpenATK is Trello, which is also advocated by Ault et al. (Ault et al., 2013). They emphasize the capability to share data records easily between several workers within the farm or stakeholders outside the farm and the guarantee of long-term ownership of farmer's data.

In Europe, much work to realize an open infrastructure for data exchange and collaboration was done within the Future Internet programme. The focus of this programme was to realize a set of Generic Enablers (GEs) for e.g. cloud hosting, data and context management services, IoT services, security and Big Data Analysis which are common to all Future Internet applications for all kind of different sectors, called FIWARE (Wolfert et al., 2014). The SmartAgriFood proposed a conceptual architecture for Future Internet applications for the agri-food domain based on these FIWARE GEs (Kaloxylou et al., 2012; Kaloxylou et al., 2014). The Flspace project implemented this architecture into a real platform for business collaboration which is visualized in Figure 2.5 (Barmponakis et al., 2015; Wolfert et al., 2014).

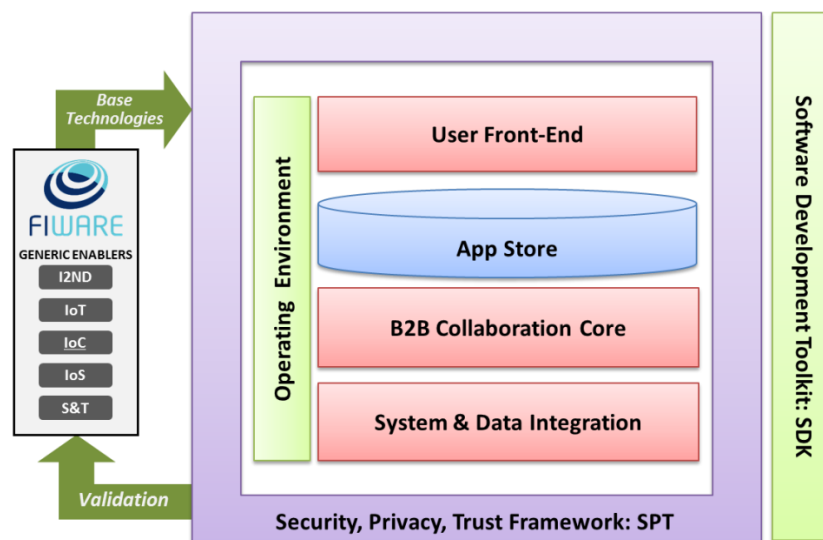


Figure 2.5 A high-level picture of the Flspace architecture based on FIWARE GEs.

Flspace uses FIWARE Generic Enablers (GEs) but has two particular extensions for business collaboration: the App Store and the Real-Time B2B collaboration core. These key components are connected with several other modules to enable system integration (e.g. with IoT), to ensure Security, Privacy and Trust in business collaboration and an Operating Environment and Software Development Kit to support an 'ecosystem' in which Apps for the Flspace store can be

developed. The Flspace platform will be approachable through various type of front-ends (e.g. web or smartphone), but also direct M2M communication is possible.

Because all mentioned open platforms result from recent projects, their challenges is still how they could be broadly adopted. For the Flspace platform, a first attempt was made in the FIWARE accelerator programme¹ in which several hundreds of start-ups were funded to develop apps and services and also received business support. Some of them were already successful in receiving further funding from private investors, but it is too early to determine the final success rate of this programme.

2.4.5 Challenges

The challenges for Big Data and Smart Farming found in literature can be broadly classified into technical and organizational ones of which the latter category is considered the most important (Orts and Spigonardo, 2014; Sonka, 2015). Moreover, most technical challenges will be solved if enough business opportunities for Big Data in Smart Farming can be created, so there needs to be a clear return on investment (Lesser, 2014). On the revenue side, there is a challenge to make solutions affordable for farmers, especially for those in developing countries (Kshetri, 2014). If there will be more users of Big Data applications it will lead in its turn to more valuable data, often referred to as the *reciprocal value* of Big Data (Van 't Spijker, 2014). This is a very important feature that needs to be carefully implemented in companies' strategies. On the costs side, the challenge is to automate data acquisition in such a way that there are virtually no costs (Sonka, 2015). Because on-farm data will generally remain in the hands of individual companies, investments are needed in a common pool infrastructure to transfer and integrate data and finally make applications out of it. Poppe et al. (2015) refer to this as Agricultural Business Collaboration and Data Exchange Facilities (ABCDEFs). An important question concerning these ABCDEFs is if these will be closed, proprietary systems such as currently Monsanto's FieldScripts or if these will be more open as proposed by e.g. the OpenATK or the Flspace platform. Finally, another business-related challenge of Big Data is how the potential of information across food systems can be utilized (Sonka, 2015).

One of the biggest challenges of Big Data governance is probably how to ensure privacy and security (Lesser, 2014; Orts and Spigonardo, 2014; Sonka and IFAMR, 2014; Van 't Spijker, 2014). Currently this is sometimes inhibiting developments when data are in silos, guarded by employees or companies because of this issue. They are afraid that data fall into the wrong hands (e.g. of competitors) (Gilpin, 2015). Hence privileged access to Big Data and building trust with farmers should be a starting point in developing applications (Van 't Spijker, 2014). Therefore new organizational linkages and modes of collaboration need to be formed in the agri-food chain (Sonka and IFAMR, 2014). In other words, it means the ability to quickly access the correct data sources to evaluate key performance/core processes and outcome indicators in building successful growth strategies (Yang, 2014).

All aforementioned challenges make that the current amounts of farm data is currently underutilized (Bennett, 2015). Another problem is that the availability and quality of the data is often poor and needs to be ensured before you can make use of it (Lesser, 2014; Orts and Spigonardo, 2014). A lack of integration is also reported as an important problem (Yang, 2014). Anonymization of data, so that it cannot be traced back to individual companies can also be a problem sometimes (Orts and Spigonardo, 2014). There are also attempts to include more open, governmental data (cf. the GODAN initiative), but a problem can be that the underlying systems were never designed for that or they contain many inconsistent, incompatible data (Orts and Spigonardo, 2014).

2.5 Conclusions and recommendations

Based on the findings from the literature review several conclusions can be drawn on the state-of-the-art of Big Data applications in Smart Farming. First of all, Big Data in Smart Farming is still in an early development stage. This is based on the fact there are only limited scientific publications available on this topic and much information had to be derived from 'grey literature'. Further conclusions, drawn as answers to the research questions we formulated in the introduction, are elaborated below.

¹ <https://www.fiware.org/fiware-accelerator-programme/>

What role does Big Data play in Smart Farming?

Big Data is changing the scope and organisation of farming through a pull-push mechanism. Global issues such as food security and safety, sustainability and as a result efficiency improvement are tried to be addressed by Big Data applications. These issues make that the scope of Big Data applications extends far beyond farming alone, but covers the entire supply chain. The Internet of Things development, wirelessly connecting all kind of objects and devices in farming and the supply chain, is producing many new data that are real-time accessible. This applies to all stages in the cyber-physical management cycle (Figure 1.1). Operations and transactions are most important sources of process-mediated data. Sensors and robots producing also non-traditional data such as images and videos provide many machine-generated data. Social media is an important source for human-sourced data. These big amounts of data provide access to explicit information and decision-making capabilities at a level that was not possible before. Analytics is a key success factor to create value out of these data. Many new and innovative start-up companies are eager to sell and deploy all kind of applications to farmers of which the most important ones are related to sensor deployment, benchmarking, predictive modelling and risk management.

What stakeholders are involved and how are they organized?

Referring to Figure 2.4, there are first of all the traditional players in agriculture such as input suppliers and technology suppliers for which there is a clear move towards Big Data as their most important business model. Most of them are pushing their own platforms and solutions to farmers, which are often proprietary and rather closed environments although a tendency towards more openness is observed. This is stimulated by farmers - organized in cooperatives or coalitions - that are concerned about data privacy and security and also want to create value with their own data or at least want to benefit from Big Data solutions. Beside the traditional players we see that Big Data is also attracting many new entrants which are often start-ups supported by either large private investors or large ICT or non-agricultural tech companies. Also public institutions aim to open up public data that can be combined with private data. These developments raise issues around data ownership, value of data and privacy and security. The architecture and infrastructure of Big Data solutions are also significantly determining how stakeholder networks are organized. On the one hand there is a tendency towards closed, proprietary systems and on the other hand towards more open systems based on open source, standards and interfaces. Further development of Big Data applications may therefore likely effect two supply chain scenarios: one with further integration of the supply chain in which farmers become franchisers; another in which farmers are empowered by Big Data and open collaboration and can easily switch between suppliers, share data with government and participate in short supply chains rather than integrated long supply chains.

What are the expected changes that are caused by Big Data developments?

From this review it can be concluded that Big Data will cause major changes in scope and organization of Smart Farming. Business analytics at a scale and speed that was never seen before will be a real game changer, continuously reinventing new business models. Referring to Figure 1.1, it can be expected that farm management and operations will drastically change by access to real-time data, real-time forecasting and tracking of physical items and in combination with IoT developments in further automation and autonomous operation of the farm. Taking also the previous research question into account, it is already visible that Big Data will also cause major shifts in power relationships between the different players in the Big Data farming stakeholder network. The current development stage does however not reveal yet towards which main scenario Smart Farming will be developed.

What challenges need to be addressed in relation to the previous questions?

A long list of key issues was already provided in Table 2.2, but the most important ones are:

- *Data ownership* and related *privacy* and *security* issues – these issues have to be properly addressed, but when this is applied too strictly it can also slow down innovations;
- *Data quality* - which has always been a key issue in farm management information systems, but is more challenging with big, real-time data;
- *Intelligent processing* and *analytics* – for Big Data this is also more challenging because of the large amount of often unstructured, heterogeneous data which requires a smart interplay between skilled data scientists and domain experts;
- *Sustainable integration* of Big Data sources – integration of many different data sources is challenging but because this is crucial for your business model this has to be done in a sustainable manner;
- *Business models* that are attractive enough for solution providers but that also enable a fair share between the different stakeholders;

- *Openness of platforms* that will accelerate solution development and innovation in general but also empower farmers in their position in supply chains.

The promise of Big Data in agriculture is alluring, but the challenges above have to be addressed for increased uptake of Big Data applications. Although there are certainly technical issues to be resolved we recommend to focus first on the governance issues that were identified and design suitable business models because these are currently the most inhibiting factors.

3. Description of a dairy farm case

3.1 Introduction

In plant and animal breeding big improvements have been made in the past decades by international cooperation in the field of genomic selection. Large amounts of data are used for this (big data). The development in the coming decennium is expected to be in the field of integrating phenotypic data. Phenotypic data are observable characteristics (or traits) of plants or animals. This will speed up the selection processes and in the end contribute to sustainable production systems. It is a challenge for DLO, NLR and TNO to disclose and analyse phenotypic data that are being collected by sensors for measuring performance and behaviour of animals and plants. New insights from the evolvments in big data field can help to realize this. A well-motivated consortium of companies and research institutes, also co-operating in the Breed4Food program, is willing to pick up this case. Research on the efficiency of phenotypic data collection and innovative analysis methods are needed to integrate these data efficient in breeding programs.

3.2 Case: Estimation of feed efficiency of individual dairy cows

Users

The primary user is a breeding expert that wants to incorporate phenotypic data in breeding research. This activity is scheduled several times a year.

An alternative user can be a dairy farmer who is interested in the intake of different feeding components to monitor the performance of the animals and in the possibilities in adjusting the feed intake if needed. This is an operational process with a daily/weekly/monthly frequency (see figure 3.1). Other operational processes are also livestock-oriented (health care, reproduction, milk production and replacement) or directed to land use of management. The dairy farm case focuses on feeding; the other operational processes are left aside.

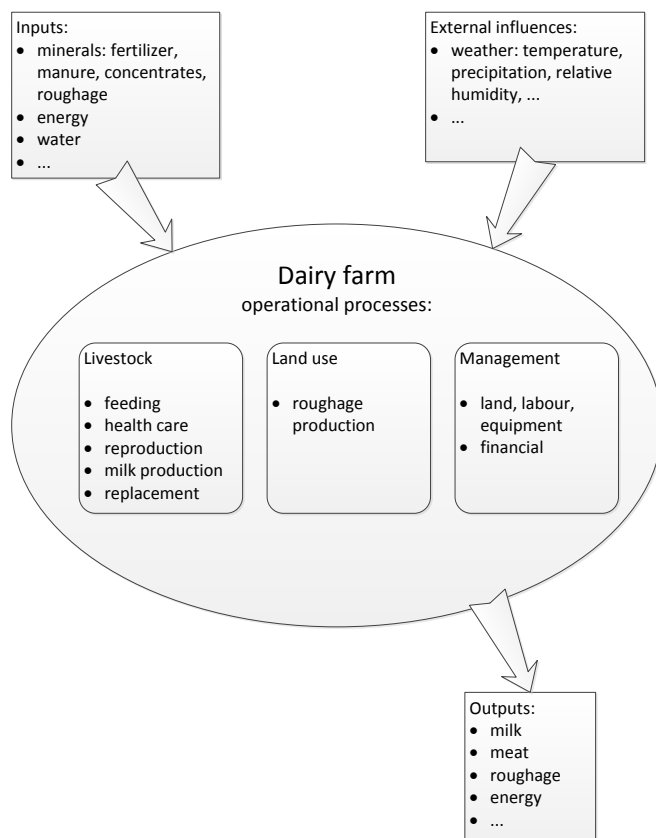


Figure 3.1 Scheme of operational processes on a dairy farm

Research question

Is it possible to provide genotypic and phenotypic data to estimate the overall feed efficiency of dairy cows?

Motivation

Feeding costs are a major part of the total costs for the production of milk. Feeding of milking cows consist generally of grass, roughage, concentrates and additives. Most detailed information on cow level is known on concentrates. Grass and roughage intake are poorly available and only on group level. So present feed efficiency is based on concentrate efficiency. Feeding efficiency of all feeding components together is getting more and more important to realize an optimal milk production. With the end of the milk quota in 2015 dairy farming is now more soil-based. Therefore emphasis will be on maximizing milk production given the limited availability of land (and thus grass and roughage). The grass and roughage should be used without wasting. Animal behaviour that implies wasting of feeding should be avoided as much as possible. Therefore in this case we are interested in estimation of feed efficiency for all feed components.

Requirements

Animal selection

- A selection of lactating dairy cows of the experimental farm Dairy Campus of Wageningen UR Livestock Research, as well as a data collection period. For the year 2015 cows that were also involved in grazing experiments were selected. Milk production and feed intake data in the preceding years should also be available for the selected cows. The parity is known, as well as other characteristics like the 305-day milk yield, days in milk and animal weight.
- Genotypic data should be available for the selected cows. This is not restricting as all cows of Dairy Campus are genotyped.
- It should be known to which group each selected cow belonged in each period and to which services access was permitted. These whereabouts are needed to know whether a cow was in a certain parcel in a given period or to relate the feed intake of a cow to the roughage given to a group.

Output:

- Production: the milk yield (kg) is recorded per cow and per milking.
- Milk composition: the fat and protein content should be recorded. For some animals also measurements from the Herd Navigator will be available: progesterone (relevant for fertility), LDH (relevant for mastitis detection), BHB and urea (monitoring of feeding).

Input:

- Concentrates supply (kg dry matter) is available per animal per day, as well as the composition expressed in VEM (intake of net energy), DVE (true protein digested in the small intestine), and OEB (degraded protein balance).
- Roughage intake (kg dry matter) and composition (VEM, DVE, OEB) should be available per cow per day in the ideal situation but is in practice available per group per day. Therefore roughage supply per group will be recorded on group level and later converted to animal level. Whereabouts are needed for this conversion. Silage analysis results and information on the mixtures are needed to determine the composition of mixed roughage rations. Roughage components are silage, fodder maize, hay and straw.
- Grass intake as eaten during grazing should ideally also be recorded in quantity (kg dm) and composition (VEM, DVE, OEB) but that is impossible in common practice. This is more complicated. It is expected that composition data are not available (might be available as results of analysis of fresh grass). Grass intake per cow might be calculated if it is known which animals were in which parcels during which period. Also here the whereabouts are important. It would ideal to have estimation of the quantity of grass available in each parcel per day. This can be measured in several ways: by mowing a subplot, measuring grass height, drone measurements or satellite measurements the available quantity can be estimated. By doing this regularly in combination with a start and end measurement and growth estimations, the quantities that are consumed by a herd of cows can be estimated. This is not perfect but nevertheless a big improvement to what is known now.
- Individualization of feed intake is needed if intake of concentrates, roughage or grass is given per group. Assignment to individual cows is necessary to notify differences between cows in feed intake or feeding behaviour. This individualization is possible by applying additional data on eating moments, eating duration,

rumination activity or maybe even time budget per cow per day. These data might be available from different sources or sensors.

In conclusion, all data that are needed to answer the research question might be available, some are ready-to-use, some only by estimation, individualization and other calculations. The implementation will be described in chapter 6 and 7.

4. Semantics and linked data

Dairy farmers are currently in an era of precision livestock farming in which information provision for decision support is becoming crucial to maintain a competitive advantage. Therefore, getting access to a variety of data sources on and off the farm that contain static and dynamic individual cow data is necessary in order to provide improved answers on daily questions around feeding, insemination, calving and milk production processes. This chapter addresses the use of ontologies and linked data technologies for combining these different sensor data sources to enable big data analysis for our use case in the dairy farming domain. We have made existing data sources accessible via linked data RDF mechanisms using OWL ontologies on Virtuoso and D2RQ triple stores. In addition, we have created a common ontology for the domain and mapped it to the existing ontologies of the different data sources. Furthermore, we verified this mapping using the ontology matching tools HerTUDA, AML, LogMap and YAM++. Finally, we have enabled the querying of the combined set of data sources using SPARQL on the common ontology. In the next section, we will first introduce the area of linked data and ontologies. In the subsequent section we describe the application of these technologies to our use case.

4.1 Linked data and ontologies

Smart farming relies on intelligent use of data-rich ICT-services and ICT-applications. One of the key enablers is the use of increasing amounts of data, resulting from sensory measurements on crops and soil, availability of thus far inaccessible information from satellites, detailed weather conditions, drone observations on individual plots of land, herd and animal movement, etc.

The entire concept of smart farming can then be catalysed by being able to relate and analyse all this data and then reasoning about it in order to generate new insights. The key question is how to make the transition to informed decision making, i.e. how do we take the right decision based on these vast amounts of data? How do we perform analysis and generate meaningful new insights? How do we support automation in this process?

This is where the concept of Linked Data plays a role: it captures the idea of enabling automated reasoning on related data sources and sharing them for reuse.

4.1.1 Linked data, big data, open data

When we talk about data sets, we can generally discern three related, but different aspects. Data can be:

- Big: the extent to which data has volume, variety, velocity, veracity.
- Open: the extent to which data is reusable to others.
- Linked: the extent to which data is accessible (linkable) to other data.

These three aspects characterize three dimensions of data, but are not necessarily the same. Data can address either one aspect or combine multiple. We can visualize the relation between the three aspects as follows:

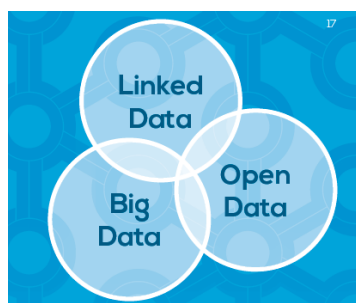


Figure 4.1 Relation between Big-, Linked- and Open data.

Open data is data that “can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and share alike “². This means that data can be published and can be made publicly available under an open licence without necessarily linking it to other data sources. Linked Data on the other hand, can be linked to URIs from other data sources, using open standards such as RDF without necessarily being publicly available under an open licence. Both types of data may concern big data sets, but of course this does not have to be the case.

The Big Data farm project is about big linked data, i.e. it concerns itself mainly with the aspects of big data sets, that need to be related or linked to make them meaningful and provide a basis for smart farming. We will thus turn our attention to the essence of Linked Data: how do we turn data sets into linked data and use it to our advantage?

4.1.2 The essence of Linked Data

Linked Data is part of the broader concept of the Semantic Web. We can describe Linked Data by means of the following definition³:

“Linked data is a set of design principles for sharing machine-readable data on the Web for use by public administrations, business and citizens.”

It is generally characterized by means of four design principles, that have been put forth by Tim Berners Lee in 2006⁴:

- Use Uniform Resource Identifiers (URIs) as names for things: the URIs may identify any kind of concept or object from the domain of discourse.
- Use HTTP URIs so that people can look up those names: use of the HTTP protocol makes data universally accessible.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL): the use of these formats (or similarly popular serialization formats such as Turtle, JSON, etc.) makes the information in the data interpretable.
- Include links to other URIs so that they can discover more things: this connects the data sets into a web of data, i.e. provides the sought after linkage.

These principles mean that Linked Data isn't just about putting data on the web and making it available, but about making links, so that a person or machine can explore the sources of data. With linked data, when you have some of it, you can find other, related data.

The notion of Linked Open Data covers and extends this idea. It represents Linked Data which is released under an open licence, which does not impede its reuse for free. The extent to which data can generally be considered to be both linked and open, was added in 2010 and is displayed in the well-known 5 star deployment scheme ⁵:

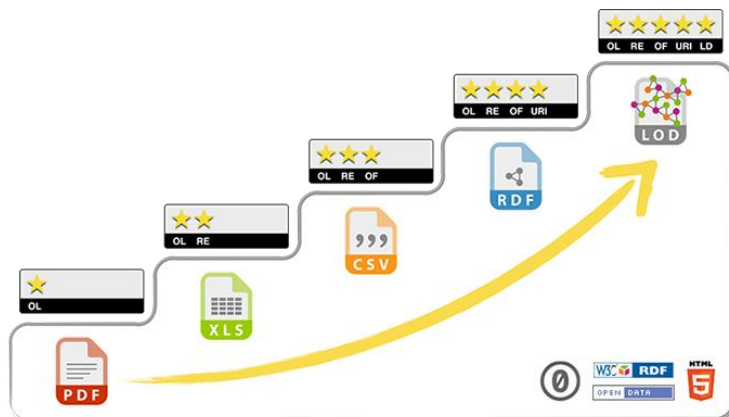


Figure 4.2. Five star deployment scheme of Linked Open Data.

The meaning of the 5 stars is as follows:

- ★ Available on the web (whatever format) but with an open licence, to be Open Data

² OpenDefinition.org

³ Open Data Support, <https://joinup.ec.europa.eu/community/ods/description>

⁴ <http://www.w3.org/DesignIssues/LinkedData.html>

⁵ <http://5stardata.info/en/>

- ★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)
- ★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)
- ★★★★ All the above plus, use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
- ★★★★★ All the above, plus: Link your data to other people's data to provide context

One of the essential features of Linked Data listed above, is that when someone accesses it, useful information is returned. This means that data is provided in an open standardized format, that can be interpreted and understood and be safely combined with other data.

The semantic web community, mainly organised in the W3C⁶, has therefore put significant standardization effort in supporting the notions of semantic web and linked data during the last ten years. Standards have been developed among others for:

- Publication: RDF (Resource Description Framework)⁷ is a metadata model for conceptual description of information, using a variety of syntax notations and serialization formats (such as e.g. JSON, N3, Turtle). It allows one to make statements about concepts or resources in the form of subject-predicate-object expressions (also known as triples in RDF terminology).
- Modelling: using OWL (Ontology Web Language)⁸, RDFS (RDF Schema)⁹ and e.g. SKOS¹⁰, it is possible to not only organize your data, but also enrich it by adding additional meaning. OWL is a knowledge representation language with fundamental logics behind it, available in a number of variants, allowing for different levels of expressiveness
- Querying: SPARQL is a language that can be used to query over multiple data sets, expressed and published in RDF. It allows defining patterns that are matched against RDF datasets, with an interpreter returning the data triples (RDF) that match the patterns. Consumers of Linked Data can thus extract information from data sets (even federated data sets) in a format that in turn allows reuse.



Figure 4.3. Standards in Linked Data and the Semantic Web.

Before we can utilize our data sets as Linked Data, we first need to make them available as Linked Data. This involves a number of steps including preparation of the data, modelling it according to semantic web standards, converting it into a format supported by the 5 star model and of course actually publishing it.

W3C have published best practices for publishing Linked Data¹¹ and have turned these into one of the foremost practical guides to Linked Data in the shape of the W3C Linked Data Cookbook¹². The Roadmap Linked Open Data¹³ is largely based on the former as well as Heath and Bizer's Linked Data book¹⁴.

⁶ <http://www.w3.org/standards/semanticweb/>

⁷ <http://www.w3.org/RDF/>

⁸ <http://www.w3.org/TR/owl2-overview/>

⁹ <http://www.w3.org/TR/rdf-schema/>

¹⁰ <http://www.w3.org/TR/skos-reference/>

¹¹ <http://www.w3.org/TR/ld-bp/>

¹² http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

¹³ <http://www.pilod.nl/wiki/BoekTNO/stappenplan>

¹⁴ <http://linkeddatabook.com/editions/1.0/>

4.1.3 Ontologies: engineering Linked Data

Organizing the data and enrich it by adding additional meaning, is the engineering or modelling part of Linked Data. Modelling makes the data more widely understandable and usable both within and across organizations. It is one of the essential steps presented in the guidelines and best practices for Linked Data mentioned above.

When creating linked data, one should employ proper engineering practices in order to create datasets of high quality that possibly make use of existing resources on the Web rather than creating them from scratch, and express the intended semantics correctly so that others (both machines and humans) can properly understand and reuse the datasets¹⁵.

Ontologies are a way to making a conceptual model of the data: they are “used to refer to a shared understanding of a domain of interest”¹⁶. An ontology embodies a view of the domain of interest expressed in terms of concepts, their definitions in terms of properties and their inter-relationships. An ontology may take a variety of forms, but necessarily it will include a vocabulary of terms and some specification of their meaning (i.e. definition).

When we apply ontology engineering to a data set and want to create an ontology, we first look for real world objects of interest, such as places, people, things, locations that are relevant to the scope of our view of the domain. These constitute the basis of our vocabulary. We can then employ domain expert knowledge, common sense and the context from which we view the domain to relate concepts and make links between them.

Finally, we can formalize our model and vocabulary by expressing it in a standardized formalization language, preferably in the Web Ontology Language OWL or RDFS.

4.2 Application to the use case

The use case of our project has been described in chapter 3. It focuses on the estimation of feed efficiency of individual lactating cows. It describes the information that is needed as input to do this estimation, such as cows, weight, whereabouts of the cow, milk yield, type of feed, feed parameters etc. This information is needed to do decision support for the dairy farmer on feed efficiency in relation to milk production. The big data analysis question in scope is:

“How much feed did an individual cow consume in a certain time period at a specific grassland parcel and how does this relate to the milk production in that period?”

By answering this question the grass input of an individual cow can be estimated in relation to the milk production. In the Dutch Smart Dairy Farm project, we have installed sensor equipment to monitor around 300 cows each at 7 dairy farms. These cows have been monitored during the last 2 years which has generated a huge amount of sensor data on grazing activity, feed intake, weight, temperature and milk production of individual cows stored in databases at each of the dairy farms. The amount of data recorded per cow is at least 1MB of sensor values per month, which adds up to 3.6GB of data per dairy farm per year. In addition, static cow data is available in a data warehouse at the national milk registration organization, including date of birth, ancestors and current farm. Finally, another existing data source contains satellite information on the amount of biomass in grasslands in the country that is important for measuring the feed intake of cows during grazing.

4.2.1 Ontology matching approach

In order to apply our linked data and ontology expertise, we selected one of the dairy farms (Dairy Campus) and studied the available data for the cows at this dairy farm. Based on this, we created with our ontology modelling tool TopBraid composer a small ontology with 12 concepts that covers among others the grasslands of the farm and the grazing

¹⁵ http://www.pilod.nl/wiki/Step_3:_Model_the_data

¹⁶ Ontologies: principles, methods and applications, Uschold & Gruninger, 1996

periods of the cows. This ontology contains the concept “perceel” which is Dutch for parcel. In addition, we selected the already existing data source with satellite information about biomass in grasslands (AkkerWeb, <http://www.akkerweb.nl>). This data source already had an ontology defined with 15 concepts that contains among others the concept “plot” which is similar to parcel but with different properties. The challenge was then to combine these two data sources in order to answer part of our big data analysis question. Combining these two data sources requires a generic ontology with concepts that are common to the domain and our specific use case. Therefore, we studied the description of the use case in more detail and derived from that the main common concepts. With TopBraid composer we created the common ontology with 28 concepts on feed efficiency. See Figure 4.4 for an excerpt of this common ontology.

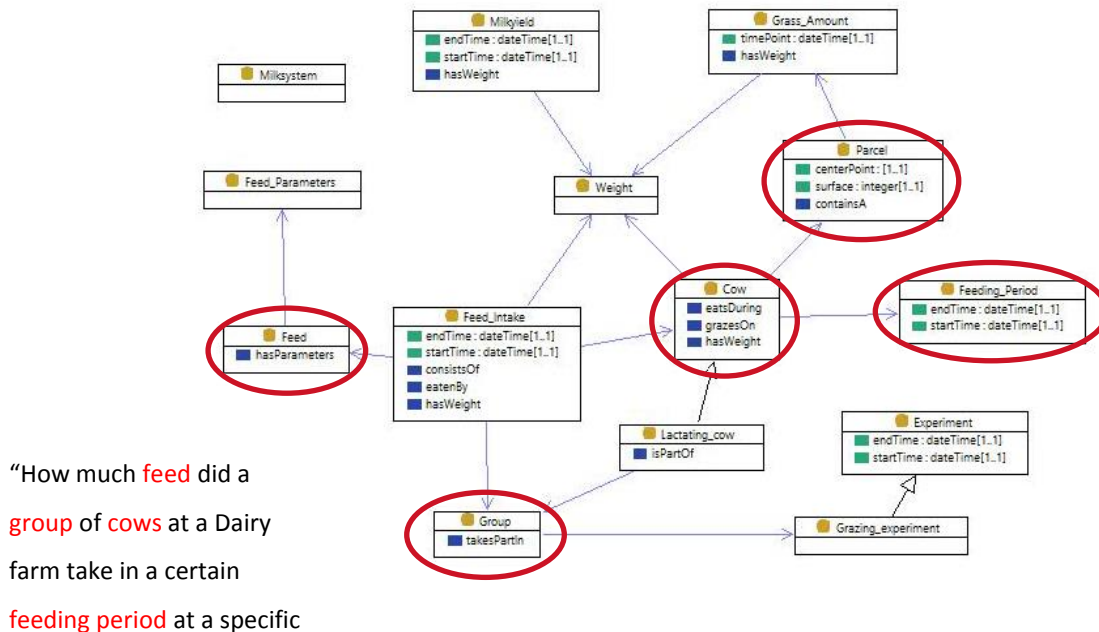


Figure 4.4 Common ontology excerpt for feed efficiency in dairy farming.

As one can see, the main concepts in our big data analysis question have been highlighted in red and are part of the common ontology. The next challenge was to find a match between the concepts and properties in the common ontology and both specific Dairy Campus and Akkerweb ontologies, especially regarding the concepts “parcel”, “perceel” and “plot”. As a first step to meet this challenge, we have created manual mappings between classes and properties in TopBraid using the existing `rdfs:subClassOf` and `owl:equivalentProperty` relational constructs. Based on relatively few and simple mappings we created initial alignments between properties and classes in the common ontology and the ontologies for Dairy Campus and Akkerweb (see Figure 4.5).

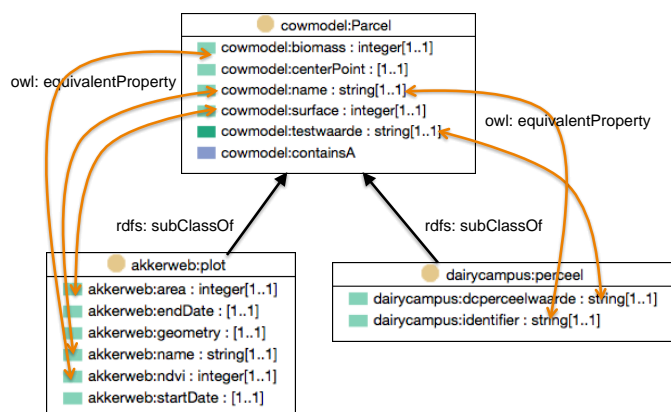


Figure 4.5. Mapping between classes and properties of the common model and the Dairy Campus and Akkerweb ontologies.

The next applied scientific question that we tried to answer was in what way our mapping exercise could also be executed by existing ontology matching tools. The use of a matching tool or system should provide us with opportunities to verify our current findings and better support our efforts in finding alignments between the other concepts in our ontologies. We used a literature survey of matching techniques and supporting matching systems in [1] to identify both a suitable matching technique and tools supporting that technique. By investigating existing techniques we deduced that language-based matching could be the appropriate type of matching since it focuses on syntactic element-level natural language processing of words.

In addition, there are numerous tools available that support this specific matching technology, mostly from academic efforts. Some however are no longer in active use, either being outdated or not maintained anymore [2].

We have selected several matching systems that support our requirement of language-based matching: HerTUDA [3,4], AgreementMaker Light (AML) [5], LogMap [6], and YAM++ [7]. We have started to investigate the possibilities of these tools to find alignments of concepts and properties in our ontologies. Initial efforts with the concepts shown in Figure 4.5 have not led to successful matches and alignments yet, however. The HerTUDA, LogMap and YAM++ tools were difficult to install and execute. The AML worked fine, but could not entirely find the relation between “parcel”, “perceel” and “plot”. Further analysis is required to find out whether this is due to inappropriate matching techniques or to the specific ontologies that we offered to the tool.

4.2.2 Triplification, triple stores and SPARQL queries

In order to show that the mapping of the common ontology to the specific ontologies works properly, we applied various tools to a separate subset of the data in the Dairy Campus and Akkerweb data sources. For Akkerweb, we derived a .csv file that was cleaned with LODRefine and transformed into a .rdf file using the Akkerweb ontology. The Dairy Campus data source was made accessible in a linked data RDF way via the D2RQ tool (www.d2rq.org) that enables access to relational databases using RDF triples and ontologies.

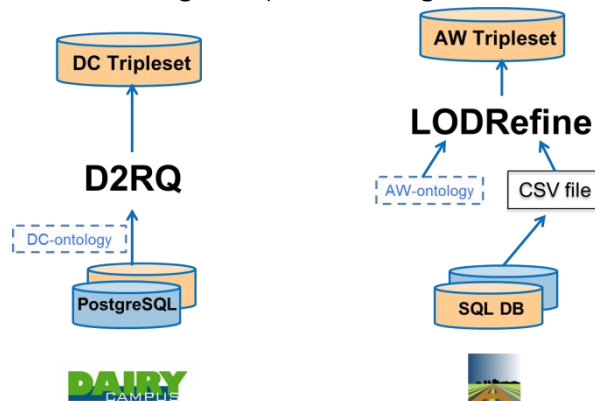


Figure 4.6. Triplification and use of the ontologies for the DairyCampus and Akkerweb data sources.

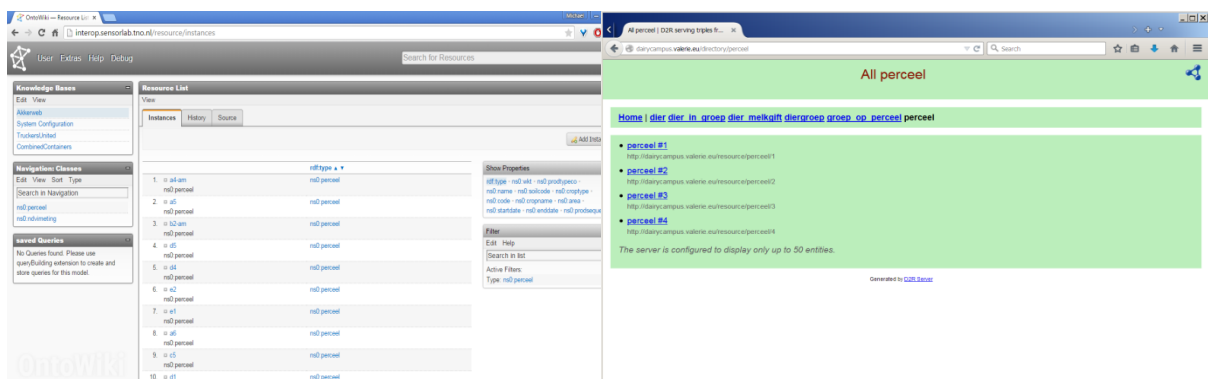
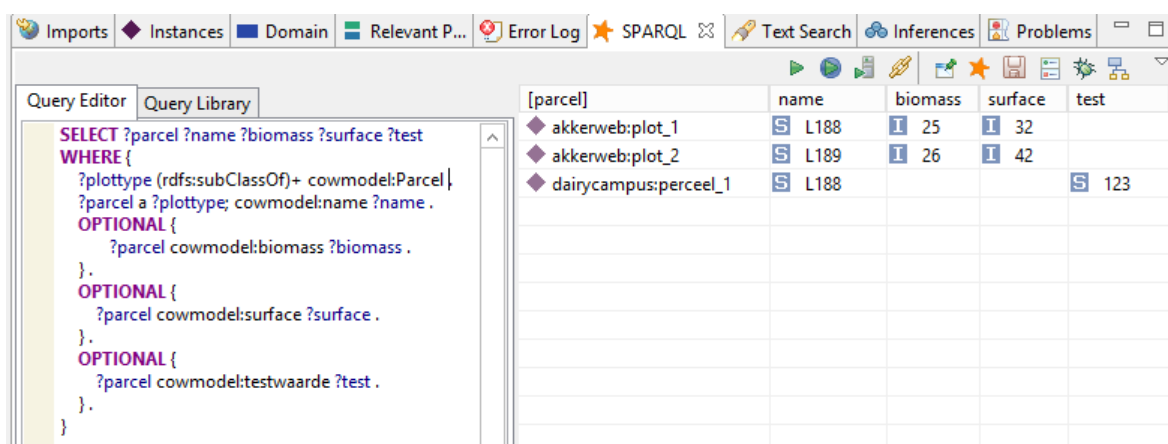


Figure 4.7. Screenshots of web-interfaces

This approach led to the publication of the Akkerweb triples on the Virtuoso triplestore via the OntoWIKI system of TNO and the DairyCampus D2RQ system. Both are accessible via a web-interface as depicted below.

A system that is based on the common ontology can take the big data question to create federated SPARQL queries on the Dairy Campus and Akkerweb triple stores using the mapped ontologies. As a result, farmers can pose questions in terms of the concepts in the common ontology instead of the detailed and specific concepts of the Dairy Campus and Akkerweb data sources. In order to show that the federated SPARQL queries resulted in the correct answers, we use the Topbraid composer SPARQL engine to define these queries and fire them towards the Dairy Campus and Akkerweb triplestores. We have built two select queries to show this.

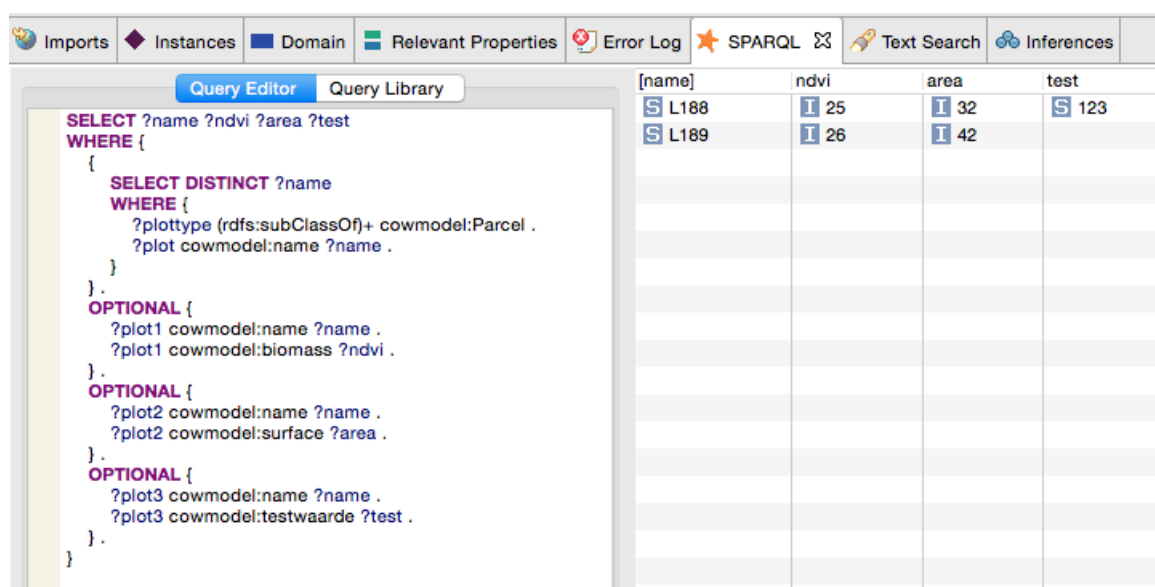
The first select query gives back all parcels thereby combining the “plots” and “percelen” that could be found via the mappings in the Dairy Campus and Akkerweb triplestores. For each parcel found, the properties name, biomass, surface and a test value to show the optional construct in a SPARQL query. The query and its results are shown in Figure 1.8a. As can be seen, the query retrieves both Akkerweb plots and Dairy Campus percelen. In addition, Akkerweb contains data about a plot with name “L188” and Dairy Campus contains data on a percel with an identifier “L188”. This means that both databases contain the same parcel and the properties can be combined.



[parcel]	name	biomass	surface	test
akkerweb:plot_1	S L188	I 25	I 32	
akkerweb:plot_2	S L189	I 26	I 42	
dairycampus:perceel_1	S L188			S 123

Figure 1.8a Select query on common ontology to retrieve all parcels.

The second select query gives also back all parcels but in addition combines the data per parcel that is present in both the DairyCampus and the Akkerweb triplestore. See Figure b below.



[name]	ndvi	area	test
S L188	I 25	I 32	S 123
S L189	I 26	I 42	

Figure 4.8b Select query to combine data from all parcels.

4.3 Insights

Summarizing, farmers can pose questions in terms of the concepts in our common ontology instead of the detailed and specific concepts of the Dairy Campus and Akkerweb data sources. Thereby, the farmer can use such a system for decision support purposes on various daily operations, such as which amount of feed to provide to which cow in which period, when to inseminate a specific cow and how to deal with the transition of a cow towards calving. The approach that is described in this chapter is currently in an experimental phase. We have reached a set-up by filling the triple stores for 3 farms with cow-data of 1 month which adds up to a total of 7 million triples. This needs to be upgraded to all farms with all data from 2014. Thereby, we can test the scalability of our system. In addition, we need to do more detailed analysis of the matching tools that we used and the reasons for not adequately solving the simple matching problem that we proposed.

5. Analysis with Machine Learning methods

5.1 Introduction

This chapter focuses on the research question “Prediction of the actual feed intake of cows using multiple sensor data (parity, weight, milk yield, milk quality, rumination and some environmental parameters), using big data machine learning techniques.” Paragraph 5.2 describes the data used. Paragraph 5.3 gives an overview of a number of relevant analysis techniques. Thereby discussing the research question “Which machine learning techniques and which combination of data sources are needed to make a good big data cow model for cow feed intake?” Paragraph 5.4 describes the approach taken for the feed intake prediction, it discusses the actual steps taken. The results coming out of these steps are presented in paragraph 5.5. Finally in paragraph 5.6 the conclusions of this study are presented.

5.2 Data description

The data used in this project come from two sources: Data about the behaviour of cows comes from the Dairy Campus in Lelystad. Data about the weather comes from a weather station of KNMI near Lelystad. An overview of these two data sets is presented in the following paragraphs. A more detailed description can be found in paragraph 5.4.1

5.2.1 Dairy Campus data

The Dairy Campus in Leeuwarden is the key data source used in the feed intake project. The sensor data is constructed by WUR Livestock Research out of a combination of different experiments they performed and it contains measurements from 21-5-2012 till 15-3-2015. This data set contains:

- 297 unique cows
- Covers in total 9 parities, spread over several cows
- 16468 measurement-sets, each measurement-set consists out of:
 - 12 parameters, 8 on a daily basis, 4 on a weekly basis.

From discussions with the domain experts followed that they were interested in modelling the roughage feed intake of cows on a weekly basis. When looking at the WUR dataset we saw that sampling the dataset at a weekly interval resulted in a dataset that was not large enough for big data analyses. It was therefore decided to extend the data set artificially by interpolating the weekly values to daily values (if they weren't already present). For the interpolation method we decided in collaboration with the experts from WUR Livestock Research that the weekly based measurements would be linear interpolated up to a daily basis. Resulting in 16468 measurement-sets, with 12 measurements on a daily basis. Table 5.1 shows the distribution of the cows over the different parities and how the number of measurement sets are available for each parity.

Table 5.1: Cow distribution over the different parities

Parity	#measurement-sets	#cows
1	5951	106
2	4300	70
3	4215	79
4	3696	61
5	2312	41
6	952	17
7	483	8
8	-	-
9	36	1

Looking through the data, 10331 measurement-sets with at least 10 days of data can be found, containing for each day values for all 12 parameters. This results in 194 partial lactations.

5.2.2 KNMI data

The weather data is used to see if it has any influence on the feed intake of the cow, in respect to the other sensor data. There is no scientific base for this, but based on a hunch of the researchers. The KNMI data set is a subset of an official weather data set coming from the KNMI organization. It contains only data for the Lelystad station, the station closest to the Dairy Campus location in Lelystad. It is limited to the same period as the Dairy Campus data set (21-5-2012 till 15-3-2015) and two parameters:

- Maximum daily outdoor temperature.
- Air humidity daily average.

The maximum daily outdoor temperature is used with the idea that the maximum temperature would be to most influential daily outdoor temperature reading. The daily average of the air humidity is chosen as an indicator of the type of day. Assuming that humidity plays a role in both the night and during the day. Both parameters are only validated by the fact that the experts agreed that this would be a good starting point to see if temperature and humidity have any influence at all.

The feed intake prediction research will take place based on both the Dairy Campus and the KNMI data set. The next paragraphs explain how the data is used and which (pre)processing steps have taken place to come to the end result.

5.3 Machine learning techniques

In a world where data is growing, the analysis of data moves from traditional methods to 'big data' methods, like data mining and machine learning. In this project we are focusing on modelling a complex process using machine learning techniques. Although the fields of statistics and machine learning are converging, there remain some important differences. The largest of them being that the core of traditional statistical techniques is making assumptions about the distribution of the data, and the underlying model where in machine learning no assumption about the model and the distribution are being made.

Where data mining focuses on exploring data, finding correlations that are not apparent with traditional techniques, machine learning focuses on modelling the data, learning it's patterns. Machine learning algorithms use large amounts of data to learn algorithms that are able to predict the data. In general they require much more data than standard statistical techniques (since no pre-assumptions about the model are being made).

The advantages of machine learning techniques are that the amount of domain knowledge required is minimal and that the resulting models (when set up and validated correctly) can model very complex (nonlinear) systems. Because it is possible to model non-linear systems, the models are good at generalizing new data points (inter and extrapolating).

Three general categories within machine learning techniques can be made:

1. Supervised learning, in which a ground-truth is provided
2. Unsupervised learning, in which no ground-truth is present
3. Reinforcement learning, used in dynamic environments where only sporadically feedback can be given.

Furthermore we can use most machine learning algorithms for several tasks, including curve fitting, classification and decision support. For the sake of this report we will only discuss some Supervised Curve Fitting methods.

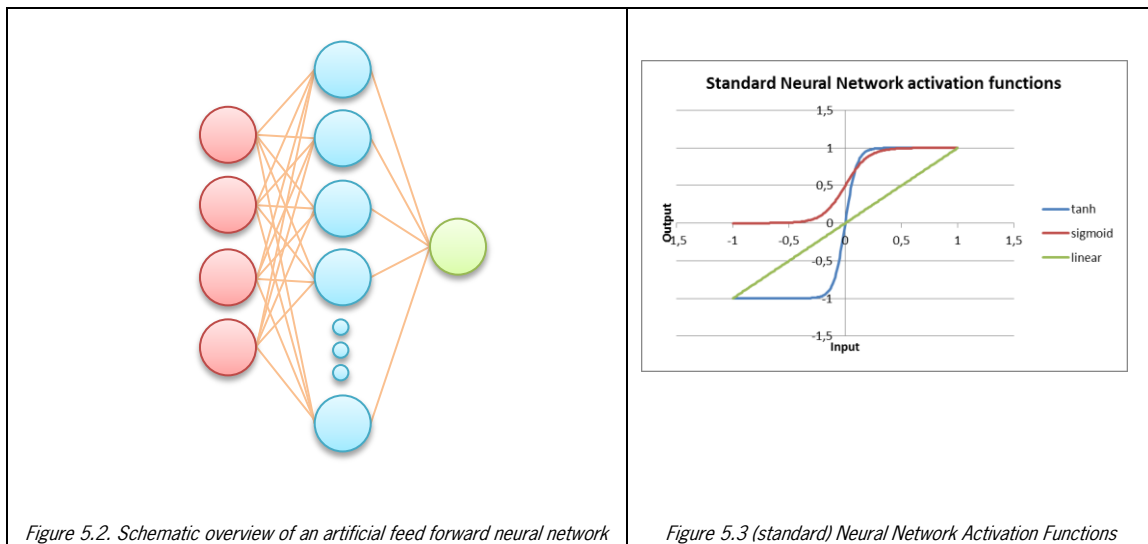
There are numerous machine learning algorithms available to choose from. We however focus on the two most popular methods that can be used for curve-fitting: "Artificial Neural Networks" and "Support Vector Machines".

5.3.1 Artificial Neural Networks

Inspired by the structure and function of biological neural networks, the basis of this method is a similar (simplified) interconnected group of artificial neurons which connect to each other. Each connection has a weight, and each neuron has a so called 'transfer function'. With the weights and transfer function we can *map* the inputs of the model, through a complex (highly configurable) collection of mathematical functions, to its corresponding output. Finding the right network topology (amount of neurons, and its connections), weights and transfer function is a difficult task that can be solved by using additional algorithms that *learn* these parameters.

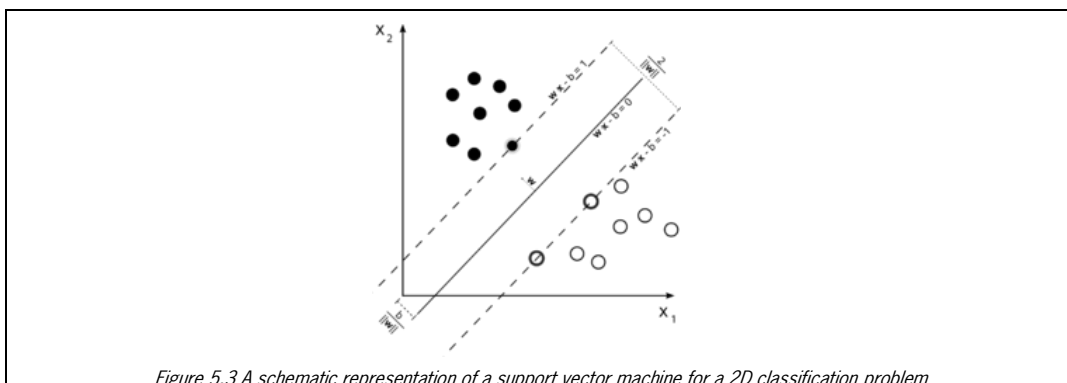
Learning is done most of the time in an iterative way, starting with an initial guess and converging to an optimum (local or global). For each iteration a test has to be performed to prevent the model from overfitting. Overfitting is a phenomenon that occurs when the model's performance is high on the training-data, but low on testing and/or unseen data. This balance is important to keep an eye on, since it is useless to have a model that can model the training data perfectly but cannot generalize and therefore cannot be used for anything else. This problem is generic and applies also for all statistical methods.

Figure 5.2 shows a simple neural network. The red nodes together form the input layer. Inside this layer we can normalize the data (e.g. so that it ranges from [-1 1]). The signals are then passed to the hidden layer (blue) where it is weighed, summed and mapped by an activation function (typical sigmoid or tanh, see Figure 5.3). The results from the hidden layers are then weighed, summed and mapped to the output layer (green), which can also function as a node that un-normalizes the data.



5.3.2 Support Vector Machines

Where Neural Networks take their inspiration from biology. Support Vector Machines originate from a purely mathematical field. A Support Vector Machine considers our measurement data to be a point in space, mapped in a particular way so that they form a straight line (or for classification problems, they are separated by a clear gap, the borders of these gaps are called the support vectors, see figure 5.3). Most systems are non-linear so this mapping will not always be possible straight away. For dealing with nonlinearity, Support Vector Machines apply a method called the 'kernel trick'. This trick is to map low dimensional data onto a higher dimensional plane (hyper plane) so the data will be linearly separable. Choosing a kernel function that provides the correct mapping, with its corresponding parameters, is a difficult problem. Multiple algorithms for finding the parameters can be used.



5.3.3 Comparison

Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are comparable to each other in many ways. It is even possible to express a single layer feed forward ANN in a linear SVM and multilayer ANNs can be expressed in terms of SVMs (see: http://ronan.collobert.com/pub/matos/2004_links_icml.pdf). However there are some side notes that have to be taken into account:

- The complexity of a SVM increases at least exponentially with the amount of data. This does not scale well with big data in mind. There are however methods of performing map-reduce techniques to solve the SVM mapping problem. The complexity of ANNs grows linearly with the amount of data, and can easily be adapted for online learning.
- The output of a SVM is straightforward (a set of support vectors) and can be interpreted. An ANN however is more like black-box.
- ANNs have to be monitored during learning stage so they will not over fit, this problem is however manageable when dividing the data in a training and a validation set.
- In ANN's it is possible to have multiple output nodes. This is not possible in standard SVMs. Modelling multiple outputs with a SVM means we need to set up multiple SVMs parallel to each other. In some cases this is preferable, when the output nodes are not inter-related to each other (this is for the ANN approach also more efficient, since the learning of one output does not introduce interference with the other output).
- Where an ANN is an universal function approximator (having numerous activations functions, the result will be a product of all these functions, thus increases with the number of hidden neuron), a SVMs performance is highly dependent on the choice of the kernel function.

Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are comparable to each other in many ways, like described in the previous section. The performance of both methods are also comparable for most applications. As a research tool both systems can be used, however when considering scalability the learning algorithms for SVMs are not favourable. It is possible to solve this by map-reduce strategies in which the workload is divided over multiple workers. ANNs do not have this problem and perform equal to SVMs. With Big Data in mind we therefore choose to conduct our experiments using ANNs.

5.4 Approach feed intake prediction

Several steps are taken to build an optimal model for predicting the feed intake of a cow using data driven techniques. Most of the pre-processing steps are generic and need to be done regardless of the chosen machine learning technique, while most of the optimization steps are algorithm specific.

The general approach is to develop an ANN model which predicts feed intake based on sensor data about the physical state of the cow, her milk production, the environment and her behaviour. The model will be trained using the special feed intake dataset of the WUR. After validation, it could be used on other cows to predict their feed intake, which is normally not known. The approach consists out of three steps:

- Pre-processing of the training data (5.4.1 data inspection, 5.4.2, data cleaning and 5.4.3 initial parameter determination).
- Neural network creation (5.4.4 data set division, 5.4.5 model set up)
- Analysis on test set cows (5.4.6 fine tuning the model, 5.4.7 input investigation)

5.4.1 Data inspection

Data inspection starts with getting a clear understanding of the data that will be used to develop our model.

Questions are:

1. What parameters are relevant for meeting our goals?
2. Does the data, and the underlying relationships, vary over time?
3. What is the level of detail of the data.

The main data set is provided by the WUR in excel format (see Table 5.2). Relevant parameters from the total set of parameters were selected using input from domain experts.

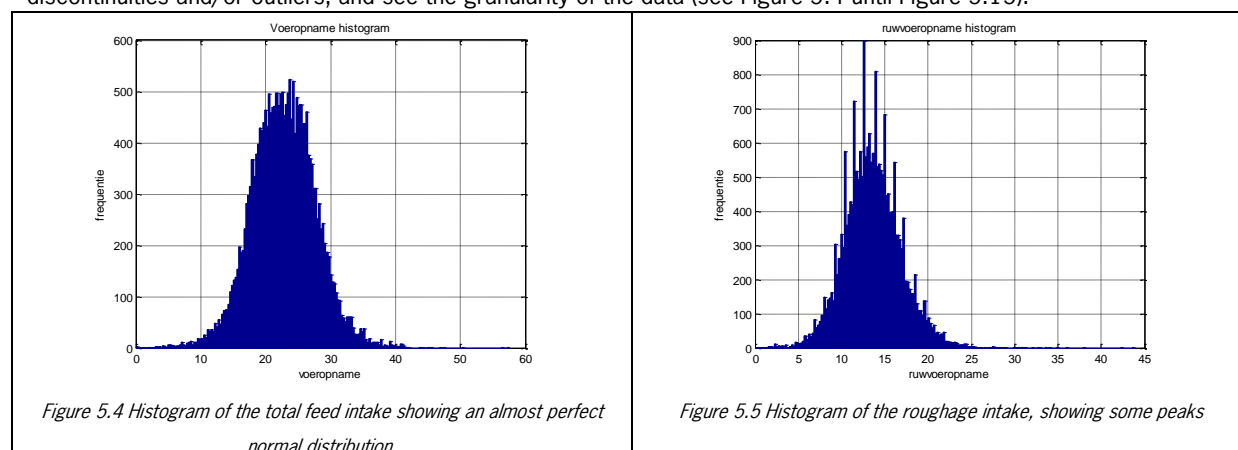
Table 5.2: Parameters that were chosen to be relevant

	ParameterVariable	Description
1	<i>DagNrInJaarDayInYear</i>	Day of the year
2	<i>LactdagLactDay</i>	Number of days since last calving
3	<i>DiernrAnimalNr</i>	Animal tracking number
4	<i>PariteitParity</i>	Parity number
5	<i>GewichtWeight</i>	Weight of the cow in kg
6	<i>MelkgiftMilk yield</i>	Cow's milk production for that day (kg)
7	<i>VoeropnameFeed intake</i>	Total feed intake for that day (kg dry)
8	<i>Ruwvoeropname roughage intake</i>	Roughage intake of the cow for that day (kg dry)
9	<i>Krachtvoeropname concentrates intake</i>	Concentrate intake of the cow for that day (kg dry)
10	<i>SumOfherkauwminuten rumination time</i>	Daily sum of 2-hourly ruminations (minutes)
11	<i>FPCM</i>	Fat-Protein Corrected Milk, i.e. standardised milk yield that day for fat 4% en protein 3,30% (weekly)
12	<i>MPR_gift test day milk yield</i>	Milk yield (during MPR=Melkgift) the weekly
13	<i>Vetperc fat percentage</i>	Fat percentage of the milk (weekly)
14	<i>Eiwperc protein percentage</i>	Protein percentage of the milk during weekly milk control
15	<i>Humidity</i>	Relative Humidity (weather outside)
16	<i>Temperature</i>	Outside Temperature (°C)

Data distribution

For most machine learning techniques (including Neural Networks), it is essential that the data is distributed evenly. This does not mean it has to follow a normal distribution; but it does require when a parameter is continuous (like in most curve fitting exercises), the data should contain the full range of that parameter with as few gaps as possible. Most machine learning techniques can overcome gaps in the data by interpolating/extrapolating to a certain extent. Furthermore categorical parameters should be evaluated. If for example a parameter d expresses which day of the week it is, it will range from 0 to 6. For any curve fitting algorithm this means that the gap between Sunday (6) and Monday (0) will be large, while the gap between Monday (0) and Tuesday (1) is smaller. To account for this phenomenon in some cases we can transform the data to a scale in which it is less pronounced (however accounting for this fully is not always possible).

For each parameter we've calculated a histogram, this way we can see how the data is distributed, spot any discontinuities and/or outliers, and see the granularity of the data (see Figure 5.4 until Figure 5.15).



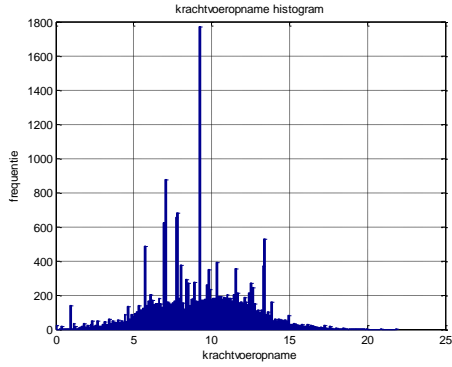


Figure 5.6 Histogram of the concentrate feed intake, showing several peaks at fixed dosages

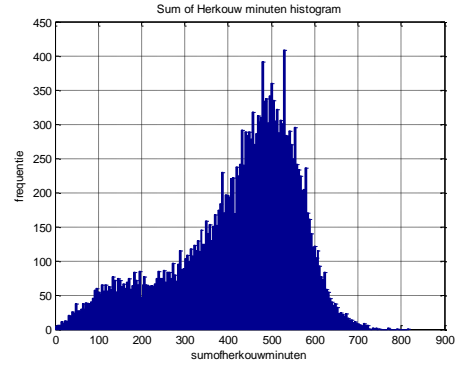


Figure 5.4 Histogram of the ruminant minutes.

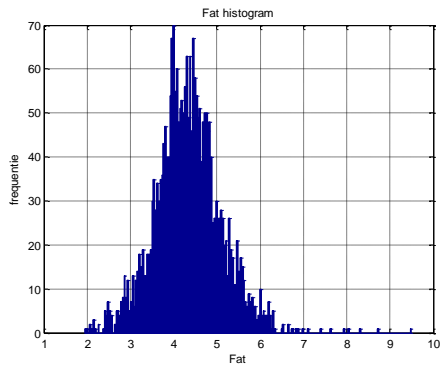


Figure 5.8 Histogram of the Fat percentage

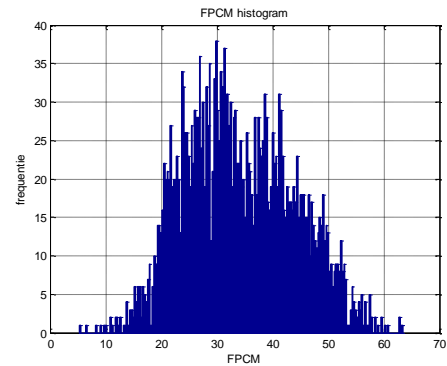


Figure 5.9 Histogram of the FPCM

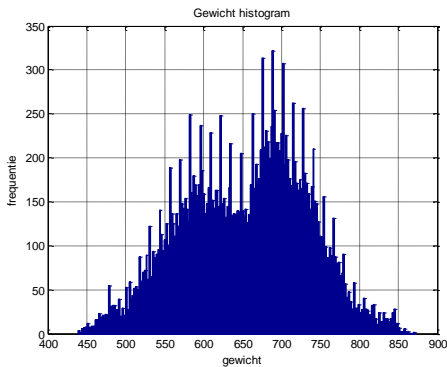


Figure 5.10 Histogram of the weight, showing two overlapping normal distributions

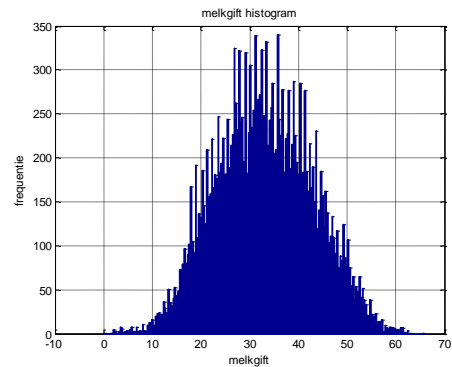


Figure 5.11 Histogram of the Milk yield

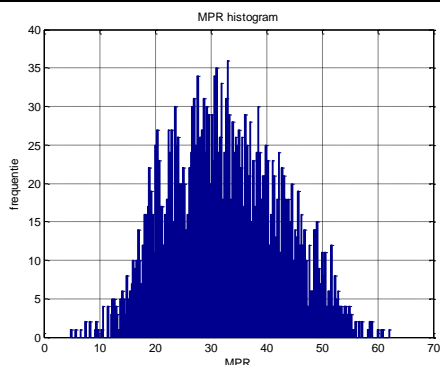


Figure 5.12 Histogram of the MPR

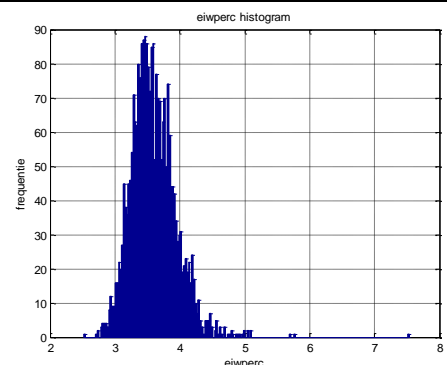
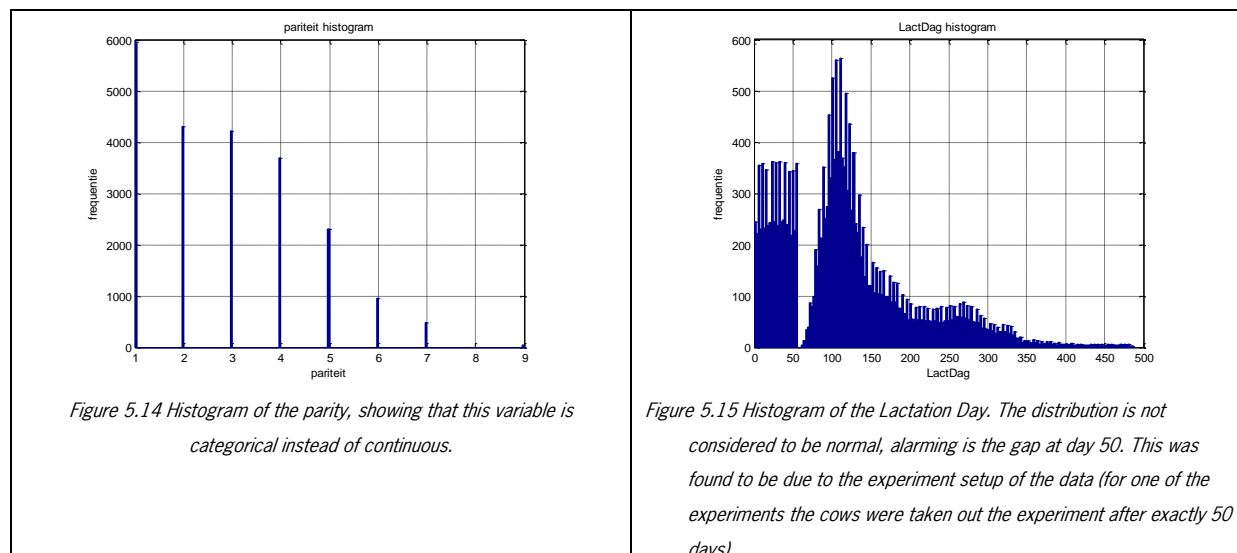


Figure 5.5 Histogram of the protein percentage



Findings:

- Most parameters have a normal distribution
- FPCM, MPR, fat percentage and protein percentage are only measured once a week
- Parity is a categorical parameter, not continuous
- Lactation days are not evenly represented in the dataset (we have a gap at 50 days due to experiments)

Alternative Findings coming from discussions with the domain experts:

- The data is gathered from experiments, this means that in certain cases the observed data does not represent normal cow behaviour.
- E.g. in several experiments the cows were fed using a TMR policy (Total Mix Ration), which means the roughage and concentrated feed is always proportionally to each other administered.

Table 5.3 Dairy Campus Projects Feed policy

Project	Non-TMR	Uncertain	TMR
1	Step_Con	-	-
2	A1,B1,C1,D1	A2,B2,C2,D2	A3,B3,C3,D3
3	A2,B2, C	-	A1,B1
4	11,21,31,41	-	12,22,32,42
5	-	-	1...16
6	1,2,3,4,5	-	-

The TMR projects represent 33,0% of the total dataset, the Uncertain projects represent 1,8% of the dataset. Which results that only 65,2% (=100% - 33,0% - 1,8%) of all data contains usable data from non-TMR projects.

5.4.2 Data cleaning

Based on our findings from the first data inspection we have chosen to clean up the data in the following steps:

1. Removal of incomplete data: Filter out the lines in the dataset where at least one of the fields is empty.
2. Filtering out TMR and unclear projects (see
3. Table 5.3).

Optionally we have interpolated the data for the entries that did not contain information (omitting filter step 1). This way we increased the resolution from one measurement vector once a week to a measurement vector once a day. We tested linear and sample hold interpolation methods. The domain experts told us that linear interpolation is valid (since the weekly parameters change gradually over time).

Finally we filtered out all lactations that did not contain at least 10 measurements.

5.4.3 Initial parameter determination

Any machine learning algorithm requires parameters/settings. These need to be set correctly in order for us to get the optimal performance. The most common method of finding the optimal parameters is by iteratively evaluate the parameters, adjust and update.

We start out with an initial guess of the parameter settings based on prior experience on machine learning experiments, looking at the complexity of the process we would like to model, and by the amount of data available.

- Ingredients:
 - We have approximately 16457 measurements vectors
 - The process complexity is moderate (we have at most 16 sensory inputs)
- Neural network settings:
 - Number of layers
 - Number of hidden layers
 - Activation function for each layer
 - Learning specific parameters (e.g. learning rate for backpropagation)

Initial guess:

- Given that the process complexity is considered to be moderate (given the number of inputs, and the number of measurements), we chose a single layered feed forward neural network.
- The activation function is chosen to be a hyperbolic tangent (tanh) for the hidden layers, and linear for the input and output layers. This setup is standard for a Neural Network used for curve fitting. An alternative activation function for the hidden layers could be a sigmoid function, which has the drawback of being in the range of [0 1] instead of [-1 1] for tanh (enabling a network with tanh activation functions to fit better with fewer hidden units).
- The number of hidden neurons is determined using an experiment in which we selected all 16 (=N) input parameters and measured the RMSerror (Root Mean Squared) of the resulting model tested on a separate dataset (see
- Figure 5.16). This resulted in choosing 35 as the number of hidden neurons (since increasing the number of neurons beyond 35 (=H) did not improve the model significantly). Choosing the number of hidden neurons too high will result in a high chance of overfitting the model on the data.
- A rule of thumb that can used to verify if the number of hidden units is chosen in a reasonable range is by looking at the number of Neural Network weights in relation to the number of measurements.
 - A single layer network with one output and N inputs and H hidden neurons has:
 - Number of weights = $N \cdot H + 2 \cdot H + 1$
 - In our case this results in 631weights ($16 \cdot 35 + 2 \cdot 35 + 1$) that need to be optimized. With 16457 measurements this results in a 1:26 ratio ($16457 / 631$)
 - Rule of thumb is that anything below a 1:20 ratio is considered to be sub-optimal for data driven machine learning techniques. With 1:26 we meet this rule of thumb.
- For the learning method we chose the Levenberg Marquardt optimizer instead of the 'standard' backpropagation algorithm. This is because we are performing an offline experiment, in which the relationships between input and output are fixed. With backpropagation the network weights can be updated in each iteration, providing a model that keeps learning and adapts to change. The advantage of Levenberg Marquardt is that it is a method that will optimize very quickly (it determines the partial derivatives of each of the network weights with respect to all measurement vectors, a Jacobian matrix), and does not require any settings. Downsides is that it requires the Jacobian Matrix to be present in memory, this cannot be done for large datasets; or distributed systems. When moving from small experiments like this towards big data experiments we can use learning algorithms like Back Propagation, or Rprop.

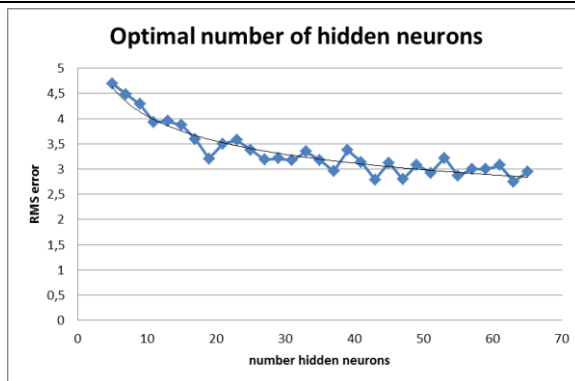


Figure 5.16 The RMS error with respect to the number of hidden neurons, trained on all 16 input parameters.

5.4.4 Data set division

To avoid overfitting the dataset, it is divided into a training set, a testing set and a validation set. The weights of the model are learned using the training set. Each epoch (a single presentation of the entire training set to the neural network), the network's performance is measured on the testing set. When we compare the performance of the testset with the training set and we see the performance of the testing set going *down*, while the performance of the training set is going *up*; we know the model is starting to overfit and that it is a good time to terminate the training phase. After the training phase the performance is measured on the validation set, just to be sure the performance we measure on the testing set equals the performance of the validation set.

Considering the 1:26 ratio described in section 5.4.3, we can conclude that we are on the edge of what is possible with ANNs. We therefore chose the testing set to be equal to the training set.

We've chosen the training set to be 80% of the data set, and the testing set to be 20%, randomly selected from the filtered data.

Table 5.4 All inputs for ANN model

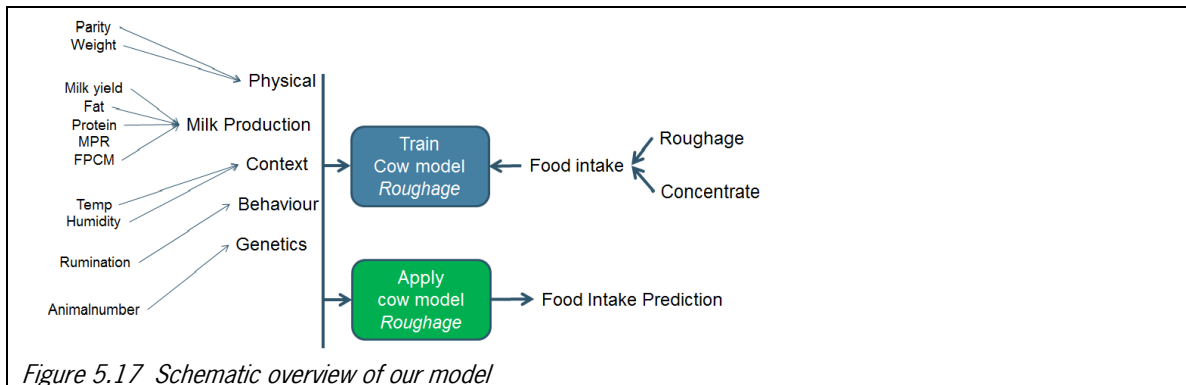
1. <i>DagNrInJaar</i>
2. <i>Lactdag</i>
3. <i>Diernr</i>
4. <i>Pariteit</i>
5. <i>Gewicht</i>
6. <i>Melkgift</i>
7. <i>Voeropname</i>
8. <i>Ruwvoeropname</i>
9. <i>Krachtvoeropname</i>
10. <i>SumOfherkauwminuten</i>
11. <i>FPCM</i>
12. <i>MPR_gift</i>
13. <i>Vetperc</i>
14. <i>Eiwperc</i>
15. <i>Humidity</i>
16. <i>Temperature</i>

5.4.5 Model setup

We started out with the parameters from section 5.4.3. From the inputs stated in section 5.4.1 we learn towards the roughage, and we do not use the total feed intake as an input (since this would not be fair, the roughage = total feed intake – concentrated feed intake). We also omit the MPR milk yield and FPCM, since this is similar to the daily milk yield and Vetperc and Eiwperc. Finally we also omit which day of the year it is. This leaves us with 11 inputs. Also see Table .

We then set up our model in Matlab. We read in all data from the .csv files, apply our filtering and pre-process steps so it can be processed by the neural network model.

The schematic setup is shown in Figure 5.. We make a distinction between Training the cow, and applying the cow model. The schematic shows that after training we can use this model on new measurements/cows that do not have data about their roughage, to predict their total feed intake.



5.4.6 Fine-tuning the model

We tried to leverage our results from our initial setup, so we could develop more accurate models. Two different steps are taken to optimize the neural network's settings, and select the most relevant neural network inputs. Note that both steps are dependent on each other. The optimal set of inputs can be influenced by the neural network settings and vice versa. This means we need to update both in an iterative way (the search space for optimizing both in the same time would be too large to be completed within polynomial time).

1. Finding the optimal neural network settings:

This step is equal to the step described in section 5.4.3 where we loop over all possible configurations, calculate the RMS error and choose one that fits best. This step is done at the start of the modelling, and at the end of all experiments (with the final input parameters, from step 2), to validate that the neural network settings do not need optimization. If there is a significant difference between the performance at the beginning of the experiments vs the end of the experiments, we need to run step 2 again with the new found settings.

2. Finding the optimal set of inputs:

This step is more elaborate than step 1. Finding the optimal set of inputs is a search-problem. It is possible to use sophisticated methods to reduce the search space, however given that the number of inputs is not that large we chose to compute all possible combinations *brute force* :

- The neural network settings are fixed (either by the results from our initial guess, or from step 1).
- Next we compute the RMSerror for all possible combinations of input parameters, the target/goal is the roughage. This leaves us with 11 input parameters, thus $2^{11} - 1 = 2047$ possibilities (note: we originally started out with all inputs, so $2^{16} - 1 = 65535$ possibilities, which took approximately 24 full days to process, so we terminated that experiment and made a subselection.)
- The results were outputted to a .csv file, enabling us to do further analysis in Microsoft Excel.
- The input combination with the lowest RMSerror was found to be the bottom row of Table 5.5.

5.4.7 Input investigations

The data coming from section 5.4.6, step 2 was used to filter out certain inputs, so we could test constraints coming from the domain experts (in other words: intentionally leaving out predetermined inputs). We calculated the performance of our model for all possible remaining input combinations for 4 possible constraints:

1. The best model without concentrates feed intake
This experiment was performed to see how much influence the concentrate feed intake has on the total feed intake. This resulted in a high RMS error, meaning this input is of high importance.
2. The best model without concentrates feed intake, and without milk yield
To see how much effect the milk yield has.
3. The best model without the cow's life number
This experiment was performed to see how much effect the 'genotype' has on the feed intake. As expected leaving this factor out resulted in a 1.8% drop in RMS error
4. The best model without constraints
The best performance was found when no constraints were imposed. This resulted in an RMS error of 1.726 (approximately 7.58%), see the bottom row of Table 5..

The output was put in a Table 5.. The red fields show the inputs we intentionally skip.

Table 5.5 Showing the results coming from brute force calculating all possible input combinations. An 'x' means the input is used, an 'o' means it is omitted. The results shown are the best results for our testing hypothesis. The red fields are fields we intentionally left out.

Levens nummer	Melkgift	Pariteit	Gewicht	Kracht Voer	Sum Herkouw minuten	LactDag	Fatperc	Eiwperc	TG	UG	Rms error	% RMS error
o	x	o	x	o	x	x	x	x	x	o	5,230	22,97%
x	o	x	x	o	x	x	x	o	x	x	7,330	32,19%
o	x	x	o	x	o	x	o	x	x	x	2,118	9,30%
x	o	x	x	x	o	x	o	x	x	x	1,726	7,58%

We performed a sensitivity study where we adjust every input parameter slightly up and downwards. The measured RMSerror did not show any significant difference between inputs, this may be due to the physical phenomenon were modelling is complex (the inputs are dependent on one another). The output is not purely correlated to the change of one input, but rather the combination of inputs. This is what a Neural Network can model very well, but makes it hard for us to get insight into the model. There are techniques that allow us to look into the 'black box' and know exactly what the underlying relationships are, however they require further research.

5.5 Results

After the initial setup we conducted numerous experiments, trying to find the optimal settings, inputs and change filtering methods based on the domain experts. Finally we settled for a model using the following 8 inputs:

1. Lactation day
2. Animal number
3. Parity
4. Weight
5. Concentrate feed intake
6. Protein percentage
7. Temperature
8. Humidity

This setup resulted in a RMSerror of 1.726 kg, or 7.58% based on an average feed intake of 26 kg per cow per day. Some farms do not have a concentrate feed intake sensor, therefore we also tested our model without this sensor as input. Resulting in only 7 inputs. The best model achieved a RMSerror of 5.23 kg, or 22.97% on average.

For display purposes we intentionally left out two random cows, and see how our model would perform on them. As it turns out these two cows both had two lactations in the dataset resulting in four comparison plots (Figure 5.18 till Figure 7).

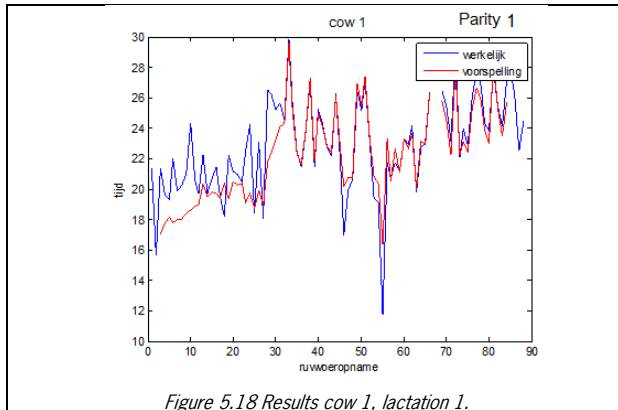


Figure 5.18 Results cow 1, lactation 1.

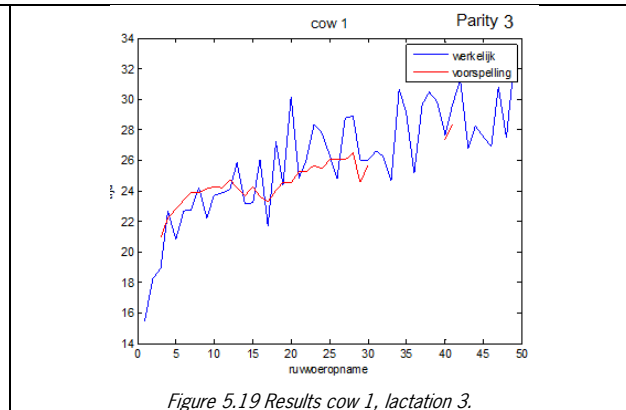


Figure 5.19 Results cow 1, lactation 3.

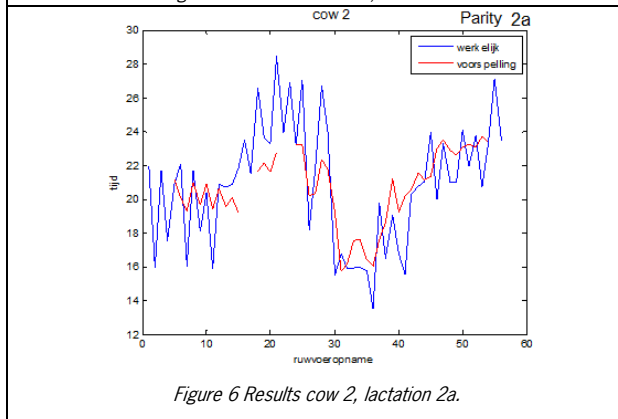


Figure 6 Results cow 2, lactation 2a.

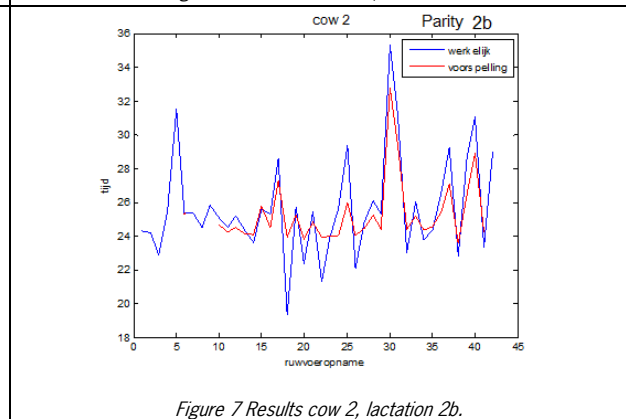


Figure 7 Results cow 2, lactation 2b.

When we exclude concentrates intake from the inputs we see that the model has more trouble following the peaks, resulting in a smoother curve (but it still follows the general trend, see Figure 5.22 till Figure 8).

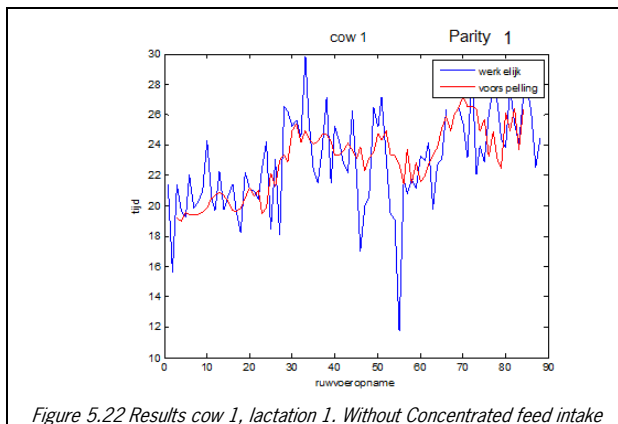


Figure 5.22 Results cow 1, lactation 1. Without Concentrated feed intake

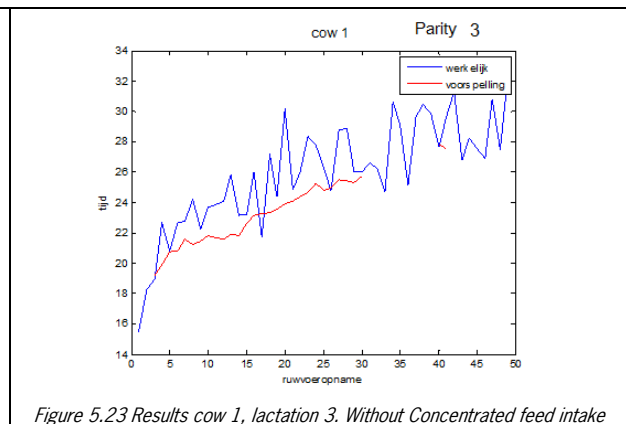
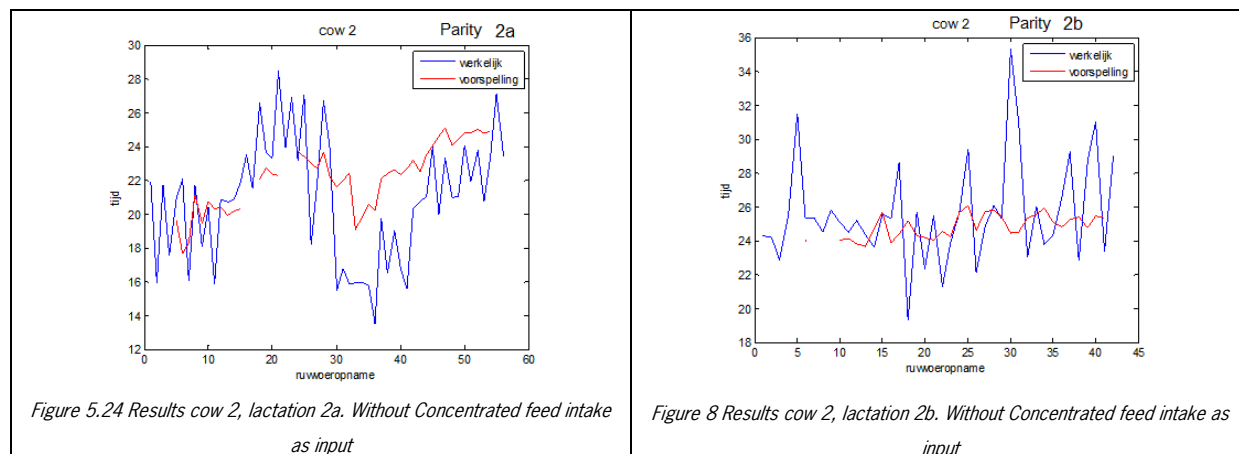


Figure 5.23 Results cow 1, lactation 3. Without Concentrated feed intake



5.6 Insights in machine learning techniques for analysis

We were able to construct an effective machine learning model to predict the roughage intake of cows. We are able to predict this with a precision of approximately 7.6%.

This means that we are able to predict rather accurately the feed intake of cows that do not have labour expensive systems to monitor the roughage intake. These models form a good basis to develop proxies for daily roughage intake of individual dairy cows, based on routinely available data on the dairy farm. These proxies for daily roughage intake can be used in management modules to optimise the feed management of individual or groups of dairy cows or be aggregate to predict full lactation intake, when no or partial feed intake records are available. To evaluate the value for genetic improvement needs further investigation, since also the predictor trait might be used directly as selection criteria.

6. Remote sensing based grass growth analysis

6.1 Introduction

In dairy farming grass forms a significant element in the food management and milk production process. The volume and quality of the grass production and consumption (preferably per individual cow) are relevant for the insight in and optimization of the milk production process, but also for the grass production process and in the end, given the context of this project, also for the investigation of the genotype and phenotype information in the framework of food efficiency and genomic selection.

Technology for sensing, location specific treatment and data processing and analysis provide new opportunities for the optimization of the grass production process. In this project research has been done on establishing grass quantity by making use of remote sensing observations. Insight in the grass quantity is relevant both in relation to the grass growing speed and the grazing progress.

Use has been made of observations from both satellites and drones. For this project a link could be made to a grazing experiment at the Dairy Campus in Leeuwarden. In the framework of this experiment a parcel was subdivided into multiple sub-parcels with varying grass stages. The parcels were observed by a drone on a weekly basis and field measurements of the grass length were collected on a daily basis. Also satellite observations were available at irregular intervals of one to four weeks (depending on the weather circumstances).

Big data techniques are relevant as the volume of remote sensing data tend to be large. Remote sensing data consist of a raster of observations covering the whole area. For each pixel reflections in multiple spectral bands may be recorded. For drone observations photos are collected with resolutions of some centimetres and with 60 to 80% overlap. As a consequence in this project the data volume of one drone flight over the Dairy Campus is in the order of 1.5Gb. Big Data techniques can help in the management and processing of these data volumes.

Secondly Big Data techniques can help in the combined analysis of multiple types of data: satellite and drone raster observations, field point measurements for height and dry weight, weather conditions, information on the grazing characteristics and cow situation. Relations between these features can be investigated.

6.2 Source data

6.2.1 Dairy Campus grazing experiment

The Dairy Campus (DC) experimental farm is situated directly south-west of the town of Leeuwarden and encompasses about 54 ha of grassland. The grassland parcels are used in various ways, most of them are used for regular mowing and some of the parcels are used for grazing experiments. For both sorts of use, the DC collects data about the amount of grass before and after actions employed (like grazing, mowing). In Figure 6.1 an overview map of the Dairy Campus parcels is shown.

The parcels that are coded A until C, D3 until D5 and E are being mowed three to four times a year. At the point of mowing, data is collected about:

- Area mowed in ha;
- Total weight (kg);
- Percentage dry matter;
- Weight dry matter (kgds);
- Weight dry matter per ha (kgds/ha).

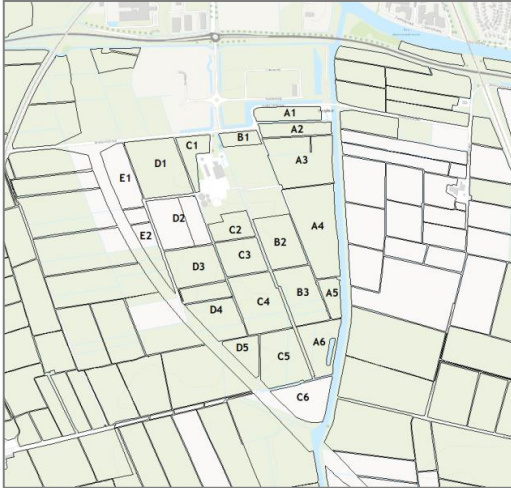


Figure 6.1 Dairy Campus parcels

At the parcels D1 and D2, grazing experiments have been conducted. Three regimes of grazing were applied:

1. 'Stripgrazen' (SG): cows get the extent of the parcel as much as needed (variable fences). In this system the area is defined by the amount of grass. The fence is moved every day (or within a few days);
2. 'Standweiden' (SW): cows get a fixed area of about 0.55 ha (6 blocks in total) and graze until the grass is shortened to a low level (5 to 10 cm height);
3. 'Blokbeveiding' (BB): the same group of 20 cows graze on one of the 24 sub-parcels available and migrate to a next one every day.

For all regimes the grass heights are measured before- and after grazing. See the Figure 6.2 for a detailed map.



Figure 6.2 Three different grazing regimes at the Dairy Campus

The figures from grazing activities and grass height in cm have been gathered in order to relate them to the NDVI figures that have been collected from satellite- and UAS data.

6.2.2 Satellite data

Satellite data have been collected from the DMC and SPOT satellites. This data is systematically made available through the satellite data portal of the Netherlands Space Organisation (NSO). NSO took the initiative to make medium resolution satellite data freely available with support of the Ministry of Economic Affairs during a 4-year period in preparation of the

arrival of the Sentinel optical satellites. From the portal, raw data can be gathered. The satellite data has to be pre-processed, namely compute geometric- and atmospheric corrections in order to get the data calibrated and enable mutual comparisons. The SPOT data from the Dairy Campus (during 2015) have been treated this way by NLR, an atmospheric correction (including darkest pixel - to be discussed later) and a (geometric) ortho-correction were applied to the SPOT images. The DMC data, also available in the NSO portal, have been picked up by Alterra, Wageningen for use in the Groenmonitor (see www.groenmonitor.nl). In the Groenmonitor, NDVI values are computed from the NIR and red spectral bands from the DMC images. The DMC data are geometrically corrected to the Dutch RD map projection and have had no atmospheric correction. Further details on atmospheric correction can be found in paragraph 6.3.1. In total, 7 SPOT images and 16 DMC images have been used for the DC during the year 2015.

6.2.3 UAS recordings

UAS recordings of the Dairy Campus have been conducted in a frequent way: about one recording every two weeks. The system used is de eBee which is a so-called wing. The eBee is operated commercially, meaning that the recorded imagery is sent to a central processing server where firstly the photogrammetric processing is performed and secondly dedicated products are manufactured like NDVI maps and Digital Surface Models (DSM). These products are made available to the customer very fast: mostly within a day. Apart from the down looking camera system the eBee instrument also contains a photometer which looks upwards and measures incoming light. These measurements are applied to perform corrections for atmospheric circumstances, such as the amount of light (clouds, haze, shadow). No specific information is available about the eBee mechanism of atmospheric correction. For the year 2015, in total 19 eBee images were recorded.

6.2.4 Field measurements

An overview of the field measurements is shown in the data matrix of figure 6.4. Data about grass height are not shown in this matrix but are available as well. The grass height data were especially gathered for the two parcels with the three grazing regimes, the parcels D1 and D2. The height is established by a so-called grass height meter that consists of a light-weight polystyrene or plastic disk with a (movable) measuring rod in the middle. During one measurement the disk is placed on the grass while the rod is put down to the ground. This action is accompanied with a counter that generates so-called clicks measuring the height. Both mechanical and digital systems are available, see also Figure 6.3.

The DC uses a standard formula to translate grass height measurements to kilograms dry matter. This formula has been derived empirically, the amount of dry matter (ds) is established to be: $ds = 110 \cdot \text{clickaverage} + 800$ while $\text{clickaverage} = \text{total nr of clicks} / \text{nr of measurements}$. One click represents a height difference of 0.5 cm. A number of measurements of about 30 per parcel is considered to be a representative amount.



Figure 6.3 Grass height meters used at Dairy Campus for field measurements

Besides the grass height measurements as collected for the grazing experiment also dry matter measurements at the moment of mowing were available for a number of parcels of the Dairy Campus. These measurements were not collected systematically, sometimes they were obtained by weighing the collected grass, sometimes by counting the number of loads and sometimes by a simple estimate.

6.2.5 Overview of all measurements

In the schedule in Figure 6.4 all gathered data from the Dairy Campus during 2015 are showed (green blocks). On the data matrix vertical axis (top-down): 15 DC parcels, SPOT and DMC satellite images, UAS images. On the horizontal axis: the time period from 17 January 2015 until 10 October 2015. The DC mowing dates can be deduced from the groups of vertical clusters of green blocks in the matrix: here data are available about the dry matter weight per hectare. The mowing dates were 27 May (and 4 June), 10 July, 21 August and 30 September. Obviously it is of most interest when acquisition dates of satellite- or UAS data coincide with the dates of mowing in order to make best comparisons.

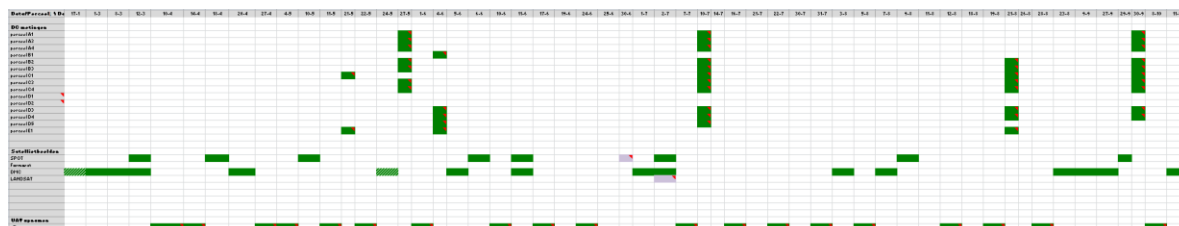


Figure 6.4 Data matrix of all remote sensing related collected data at the Dairy Campus in 2015.

6.3 Remote sensing based grass growth monitoring

6.3.1 Principle

First objective of this part of the study is to make an analysis of the relation between grass growth and remote sensing observations and define a practical methodology. Second objective is to apply big data processing and analysis technology to operationalize and enhance the methodology by fast processing of the huge amounts of collected remote sensing data and by possibly also including other information in the analysis, like weather information and grass growth characteristics.

For relating the remote sensing observations to grass growth the main idea at the start of the project was to use a vegetation index (NDVI) obtained from the multispectral observations. This index can be computed both from the satellite and UAS observations. As the spatial resolutions of the systems used in the project are quite different (satellite observations used 22m (DMC) and 6m (SPOT), eBee observations 0.12m) two experiments have been done:

- Evaluate the vegetation index at parcel level with the dry matter values available from the moment of harvesting. So only during one moment when the grass has maximal length the relation is evaluated (as no dry matter numbers are available during the growth period);
- Evaluate the vegetation index at sub-parcel level for the grazing experiment (UAS based). For this experiment different grass growth stages are available within one observation, for all these stages field measurements are available and a relatively dense sequence of weekly observations is made.

During the project it turned out that also detailed height information from the stereo UAS observations could be computed. This resulted in a third experiment:

- Evaluate the UAS based height measurements at sub-parcel level for the grazing experiment. A comparison can be made between the UAS and field based height measurements.

6.3.2 Vegetation index measurement

Before describing the results of experiments A and B, first the principles and artefacts of satellite and UAS based vegetation index is discussed. The normalized difference vegetation index NDVI is derived from the red and near infrared spectral bands that are recorded by satellite or UAS. The formula is $NDVI = \frac{(NIR - R)}{(NIR + R)}$.

To set up a univocal series of NDVI measurements, through a growing season, through several years and from various sensors it is of importance to calibrate all data from which the NDVI is calculated. The extent of calibration often depends on balanced choice and/or the application. The values that are used for the NDVI calculation can either be, in an increasing degree of detail, digital numbers (DN), radiance values or reflection values:

- The digital numbers are the values that appear in the raw recorded data. These are translations of the energy received in the sensor instrument to a digital range or pixel depth (typically an 8-bit, 12-bit or 16-bit range).
- Radiance is the amount of physical radiation sensed by the satellite sensor (per spectral band). Radiance values (L) are computed from the gain- and offset factors (per spectral band). The radiance level is measured by the satellite at sensor level, top of atmosphere (TOA). In general one is interested however in the radiation values at ground level. For obtaining this a correction is needed for the influence of the atmosphere, a combination of several absorption and reflection effects. A relatively straightforward method for atmospheric correction of the radiation level is the so-called darkest pixel correction method. In this method it is assumed that for the darkest pixel in the image (in general water with a very low reflection) all radiation is originating from the atmosphere between the satellite and the earth surface and nothing from the object itself. Based on the radiance value of this darkest pixel all other pixels in the image are corrected for the atmosphere contribution.
- Reflection is the physical property of an object stating which percentage of the light in a specific spectral band is reflected. Reflection values are computed basically from the ratio radiance/irradiance, thus the relation between the reflected radiance on the ground and the incoming solar flux. The formula takes into account the distance between the earth and the sun (AU, varies within the year), the exo-atmospheric irradiance and the sun zenith angle. Reflectance values can be computed from radiance and irradiance at sensor level (top of atmosphere TOA reflectance). Reflectance values at ground level (bottom of atmosphere BOA) can be computed from radiance and irradiance extended with a darkest pixel atmospheric correction.

When a NDVI time series is calculated from the same satellite and all factors mentioned above are considered as constants or insignificant, thus no changes in gain, offset, AU, irradiance, sun angle or atmospheric circumstances, then the DN values can be used. This is the approach in the 'Groenmonitor' where the DN values of the DMC satellite are directly used for the NDVI calculation. But as soon as another sensor is added it will have to be examined if its DN values can be used straightforward in the same way for calculating NDVI or if relative or absolute calibration is required. That will depend on sensor sensitivity, calibration, pixel depth and the factors of AU, irradiance and sun angle and the potential tendency that those factors may change in time.

The UAS data (eBee) that were collected from the Dairy Campus are processed in an automated way via Agrometius / Sensefly with the help of the pix4D processing software. This photogrammetric software treats the individual photo's geometrically and radiometric. How exactly the corrections are performed is not revealed. It is known that the eBee carries an upward looking photometer in order to measure the column of atmosphere above the wing. It is also known that grey-values calibration panels on the ground are sampled in order to perform calibrations. But until now the user gets no inspection possibilities. Some issues appear from the processed imagery, for instance the fact that the four spectral channels are recorded with four different cameras which demands extra attention for the co-registration of the spectral bands. Also the way in which ortho-mosaics are manufactured (which parts of which photos prevail) is not very clear.

In Figure 6.5 an overview is shown of NDVI measurements for parcel C3 and surroundings as obtained with eBee, SPOT and DMC at 10th of June. Clear is the difference in spatial resolution. For the evaluation of the grass growth in the parcels as a whole the average NDVI is processed for each parcel. In that case the lower spatial resolution is less critical (note that pixels at the border of the parcel should not be included as they also contain information on neighbouring roads, ditches or parcels). What can be noticed is that in the high resolution eBee images seams are present (especially in the parcels around C3, see nearby the red arrows). This is caused by the fact the NDVI image is composed of a mosaic of multiple photos, each with own light conditions and calibration factors. Apparently this calibration and mosaicking process is not optimal yet. In case of detailed analysis of variations within a field the errors are significant and will disturb the result.

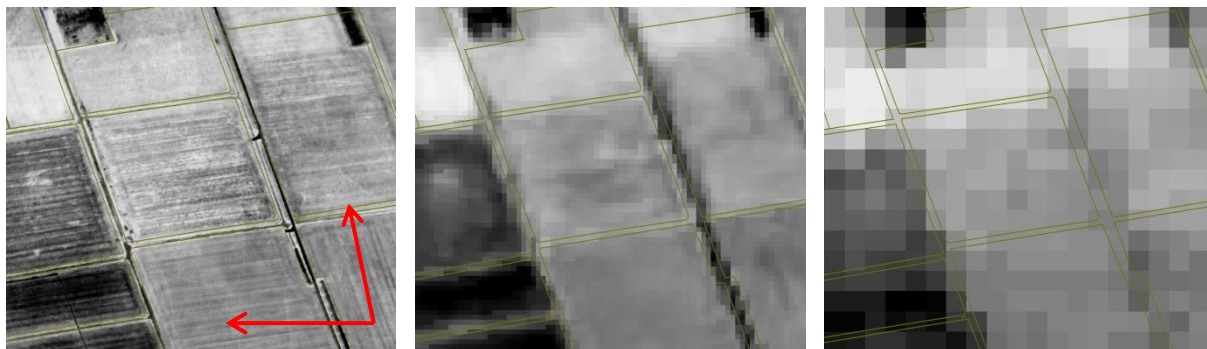


Figure 6.5 NDVI images of parcel C3 from eBee (left), SPOT (middle) and DMC (right), with radiometric artefacts in the eBee image marked.

The quantitative mean NDVI values for parcel C3 as obtained with DMC, SPOT and eBee are shown in Figure 6.6. The graph represents all recordings from parcel C3 at the Dairy Campus. From each recording the average NDVI for the area of parcel C3 was computed. The DMC data did not have an atmospheric correction, The SPOT data were corrected to reflection values including darkest pixel and the eBee data have been corrected with pix4D as described above. It is clear that significant differences can occur in the obtained NDVI when sensor calibration is not tuned (in various ways).

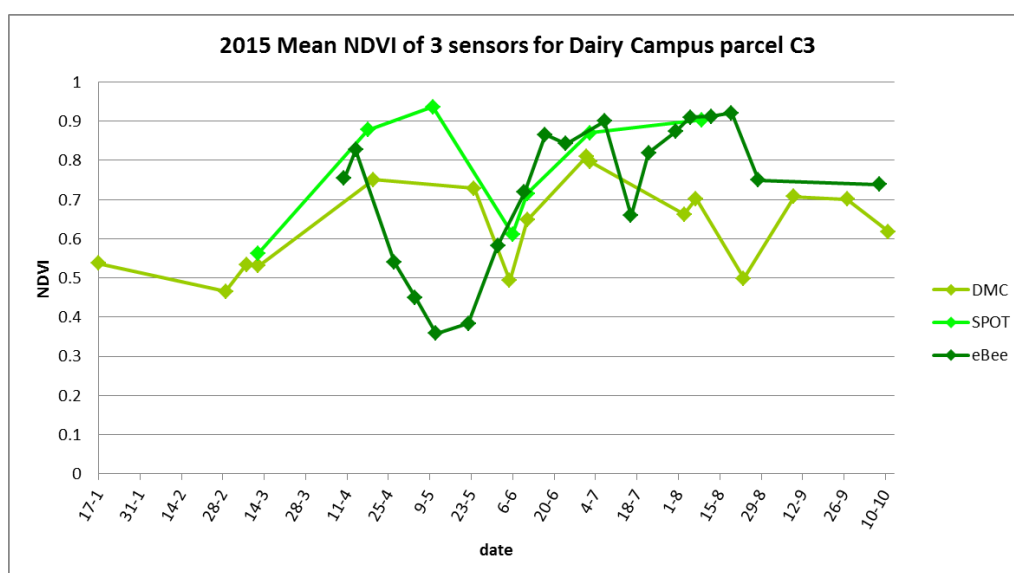


Figure 6.6 Mean NDVI values of DMC, SPOT and eBee for Dairy Campus parcel C3

The eBee recordings started at 10 April 2015. Although the first 2 images are about in the same NDVI range as DMC and SPOT, the following 4 recordings are not at all. Upon inquiry it appeared that for those acquisition dates the automated pix4D processing has not performed. Another outlier (compared to DMC and SPOT) was observed in the recording of 16 July 2015. But in this case the eBee data are probably correct, it just lacks SPOT and DMC data at the same acquisition date otherwise those values were probably lower as well (parcels were mowed on 10 July). The rest of the eBee recordings seem to coincide reasonably well with the SPOT imagery. SPOT have had atmospheric corrections including darkest pixel correction. As UASs fly much lower than satellites, 0.2 km height versus some 700 km height, almost no atmosphere is present between the sensor and the ground level and thus no atmospheric correction is required. Nevertheless, it is important to calibrate the UAS sensor data, so that calibrated radiances or reflections can be computed. The eBee calibration is done by taking a photograph of grey-level samples before a flight and by using a brightness sensor that measures the incoming sunlight.

The NDVI values from DMC, not atmospherically corrected, have a tendency of being lower than the ones from eBee and SPOT. That is also the conclusion from background research on the three different atmospheric correction methods (DN, radiance and reflectance); corrections on DN only always lead to lower NDVI's compared to corrections for radiance or reflectance (including darkest pixel). Variations in NDVI up to 0.2 can occur between the three approaches (when the same pixel value is computed by the three methods). This graph seems to show that differences in NDVI are even higher than 0.2. This will be due to whether or not applying an atmospheric correction but can also be influenced by differences in the specific atmospheric conditions as the acquisition dates of DMC, SPOT and eBee are almost never the same.

Concluding on NDVI measurement:

- Good calibration and atmospheric correction are essential in case of quantitative usage of the data. This is even more important when observations from different sensors (different satellites or satellites and UASs) or from temporal range during a season or sequence of years need to be combined.
- For satellite observations a TOA radiometric correction, extended with a darkest pixel atmosphere correction, is probably the best start approach. An atmospheric correction that also applies corrections for local variations within a satellite- or UAS scene, like clouds, haze and shadows, would be an ideal next step. Anyway, common understanding about the best way to calibrate satellite- and UAS data in the Netherlands would be a good thing to pay attention to.
- For UAS observations better calibration and mosaicking routines and procedures are required, also taking into account camera light fall-off and proper selection of photos (limiting shadowed areas and optimizing the usage of ground reference and light sensors).

6.3.3 Vegetation index in relation to grass harvest weight

An analysis has been carried out to investigate the degree of correlation between NDVI and dry matter weight per hectare (kgds/ha) for several Dairy Campus parcels just before the moment of mowing. The parcels used for the analysis are shown in Figure 6.7.

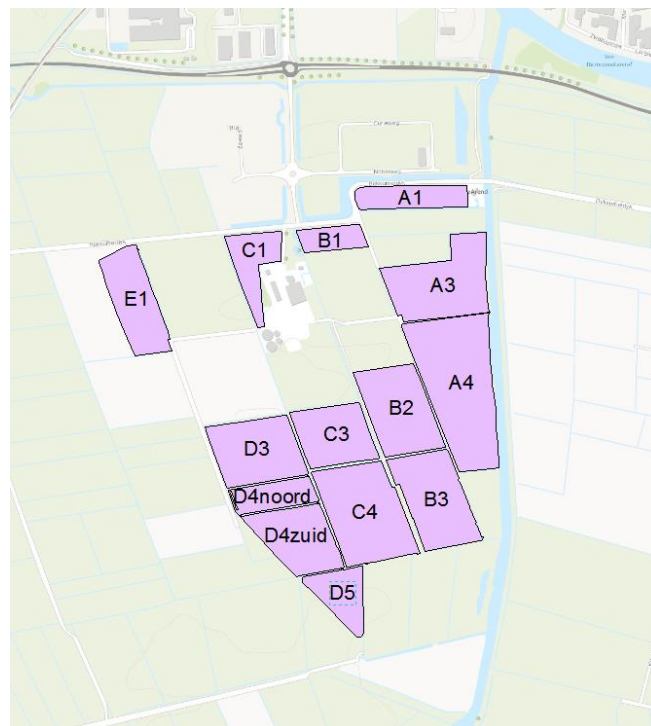


Figure 6.7 Dairy Campus parcels used for the analysis of the relation between NDVI and dry weight just for the moment of mowing

The grass mowing dates are leading for this analysis as these deliver the kgds/ha. For these dates available remote sensing observations have been selected that are gathered closest possible before the mowing dates of 27-5, 10-7, 21-8 and 30-9. The NDVI data may have been derived from several sensors (DMC, SPOT or eBee) but in practice, mostly it will be only one sensor that delivers NDVI data close to a mowing date.

From the remote sensing based NDVI images mean NDVI values per parcel were computed by averaging all pixels within the polygon of each parcel. The parcel polygons and parcel ha figures are obtained from the Dutch parcel registration (BRP) that is generated on a yearly basis. Sometimes polygons of the BRP 2014 were taken and sometimes parcels of the BRP 2015, that had to do with building activities during the two years. The parcel surface information was also compared with the surface information as obtained by the Dairy Campus.

In Table 6.1 the grass dry matter and NDVI data gathered for all parcels are listed. The dry matter numbers and NDVI values are also plotted against each other, as shown in Figure 6.8.

Table 6.1 Grass dry matter and NDVI data gathered for Dairy Campus parcels at mowing time.

parcels	kgds/ha - mowing dates					NDVI - acquisition dates					
	27-5	4-6	10-7	21-8	30-9	24-5 DMC	1-6 eBee	2-7 DMC+SPOT	7-7 eBee	19-8 eBee	27-9 DMC
A1	4766		2757		1283	0.5362		0.6847+0.8457	0.8689		0.5906
A3	4992		2431		1348	0.5286		0.6393+0.7423	0.8499		0.7287
A4	5062		2587		1434	0.7649		0.8001+0.8428	0.8788		0.7488
B1		4594					0.8516	0.5906+0.8151			
B2	5120		3320	1950	1380	0.7589		0.7931+0.8668	0.8979	0.9243	0.7424
B3	4830		3730	1632	1399	0.7636		0.7971+0.8620	0.8862	0.9292	0.7168
C1			2576	2656	1200			0.6237+0.8152	0.8489	0.9086	0.6391
C3	5338		3730	1974	1283	0.7290		0.7974+0.8704	0.9003	0.9210	0.7012
C4	5798		3730	1585	1359	0.7644		0.7946+0.8586	0.8912	0.9314	0.7127
D3		5655	2935	2089	1332		0.9047	0.7938+0.8661	0.9031	0.9166	0.7229
D4noord		6354	2957		1398		0.8865	0.7444+0.8141	0.8777		0.7291
D4zuid				2410						0.8933	
D5		4986	1935					0.7561+0.8364			
E1		4849		917			0.8072	0.6351+0.6870		0.8856	

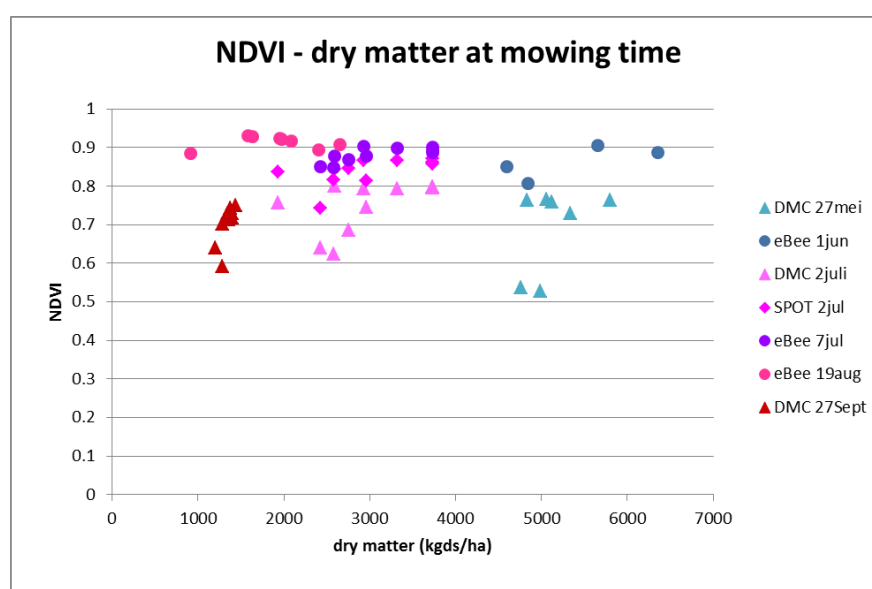


Figure 6.8 Relation between grass dry matter and NDVI values plotted

From the table and figure a number of conclusions can be drawn:

- No clear relation exists between the NDVI value and the amount of grass dry matter at time of mowing. At a specific date the found NDVI values have almost the same level for all parcels, while the amount of measured dry weight varies over the parcels. Also over the different grass cuts there is almost no variation between the NDVI values, while for the dry weight values a clear decrease is measured over the sequential cuts (as may be expected). An explanation for this is that the vegetation index satisfies at a certain moment during the grass growth, after which the grass growth continues, but the vegetation index does not increase further. All the measurements are taken at the moment of mowing, so with maximum grass volume.
- There are differences in the NDVI values for DMC and eBee. This is in line with the analysis on the calibration and NDVI computation for the different sensors as described in Section 6.3.2. The mean NDVI values for SPOT and eBee are in the same order, while the DMC NDVI values are systematically lower. The spreading in NDVI values for the DMC data seems also to be larger than for the eBee and SPOT data. This can be the effect of the larger pixels (25m instead of 6 and 0.12m) leading to a relative small amount of representative pixels within a parcel and consequently leading to potential larger variations in those pixels. Besides the sensitivity of the DMC sensor is lower and the information is quantized with only 6 bits.

6.3.4 Vegetation index in relation to grass growth

As described in Section 6.2.1 for the grazing experiment at the Dairy Campus 24 sub-parcels have been defined on the parcels D1 and D2 on which a block grazing regime was applied, see Figure 6.9. This means that a group of 20 cows graze on one of the 24 sub-parcels for a day, eat all the grass and migrate to a next one every day. The idea is that after 24 days the cows arrive at the first sub-parcel again and ideally the grass on this sub-parcel has grown to full length in the meantime. This means that the 24 sub-parcels show all growing stages between grazed and fully grown grass. As eBee remote sensing observations were made every week, also the grass growth activity at a specific parcel can be followed. For all sub-parcels also field measurements of the grass height were collected with grass height meters (as described in section 6.2.4.) that can be used as ground truth.



Figure 6.9 Overview of the 24 sub-parcels at the Dairy Campus as used for block grazing.

In Figure 6.10 at the left side an overview is shown of the NDVI vegetation index image of the sub-parcels, taken at 16th of July, and at the right side the grass height values as measured in the field with the grass height meter two days earlier. In the left image it can be seen that there is a clear variation in the NDVI values. Light areas have a relatively high NDVI value, corresponding with more grass biomass and thus with grass having a larger height. The yellow arrow shows the sub-parcel where the cows are grazing. North of this parcel the grass is still high with relatively high NDVI values, south the grass is short with relatively low NDVI values. The grass height values shown right are measured in the field with the grass height meter at 14th of July, so two days earlier. With the red arrow the field is shown where the cows are grazing at that day, two fields south from the field in the NDVI image. The colours show that north from this

field the grass is long, while south of it the length is shorter. The main pattern between both images looks similar, although there are smaller differences in the variations between sub-parcels.

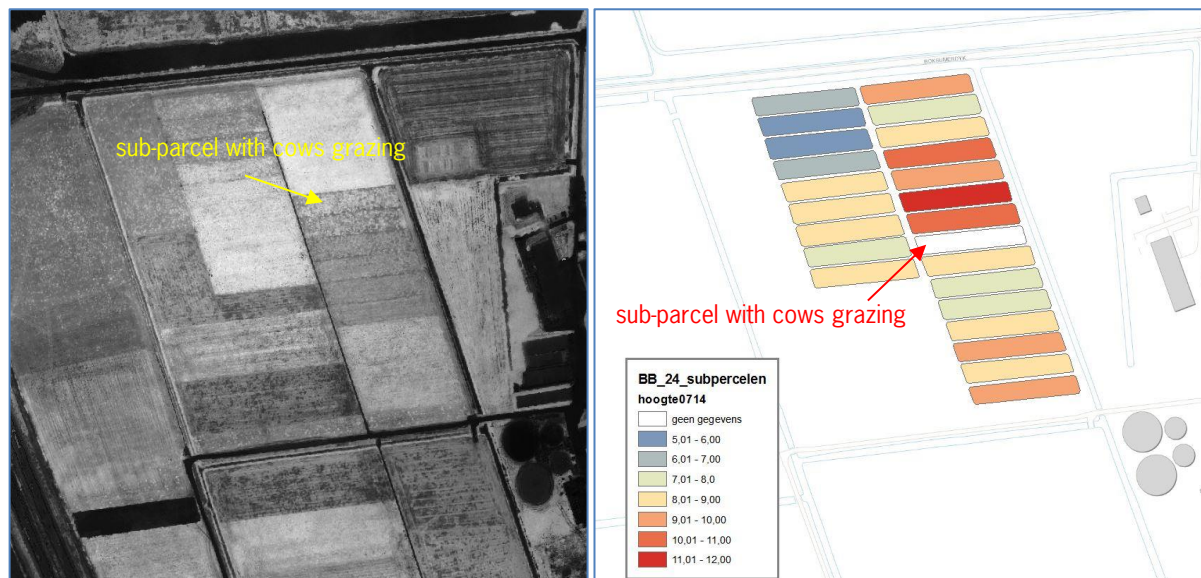


Figure 6.10 Overview of eBee NDVI image of the 24 sub-parcels at 16th of July (left) and the in the field measured grass height values two days earlier at 14th of July (right).

In Figure 6.11 the measured NDVI and grass heights values have been plotted against each other. From this plot it can be seen that there is a relation between the NDVI measurements and grass height, although this relation is not very strong. The R^2 value for a linear regression is 0.20.

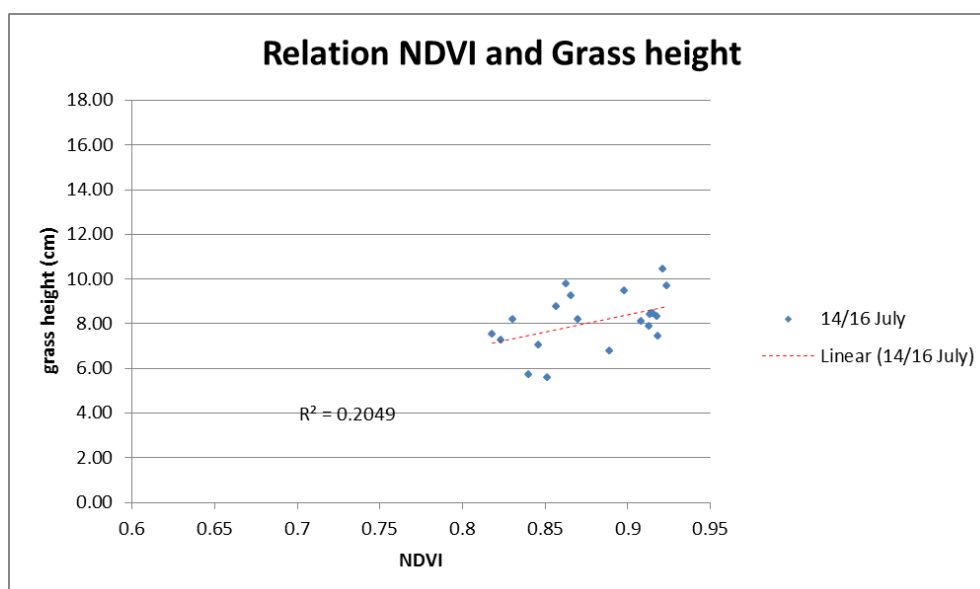


Figure 6.11 Values of eBee NDVI measurements and grass height measurements for all sub-parcels at 14th of July plotted against each other.

Also for 10 other observation dates the same plots have been made, as shown in Figure 6.12. In addition in Figure 6.13 the found height ranges, NDVI ranges and R^2 values are summarized.

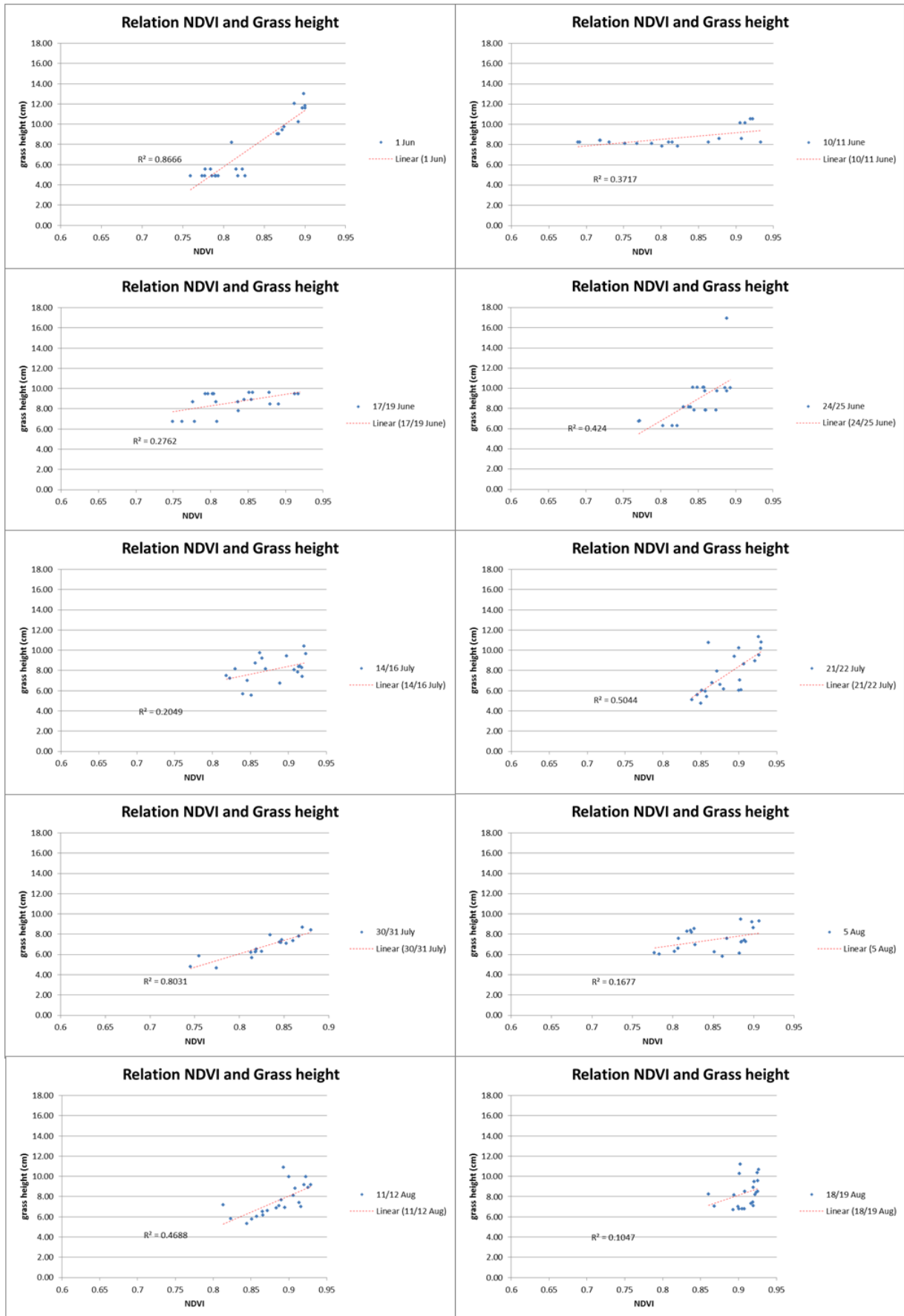


Figure 6.12 Relation between eBee NDVI grass height meter values for all sub-parcels at 10 different observation dates.

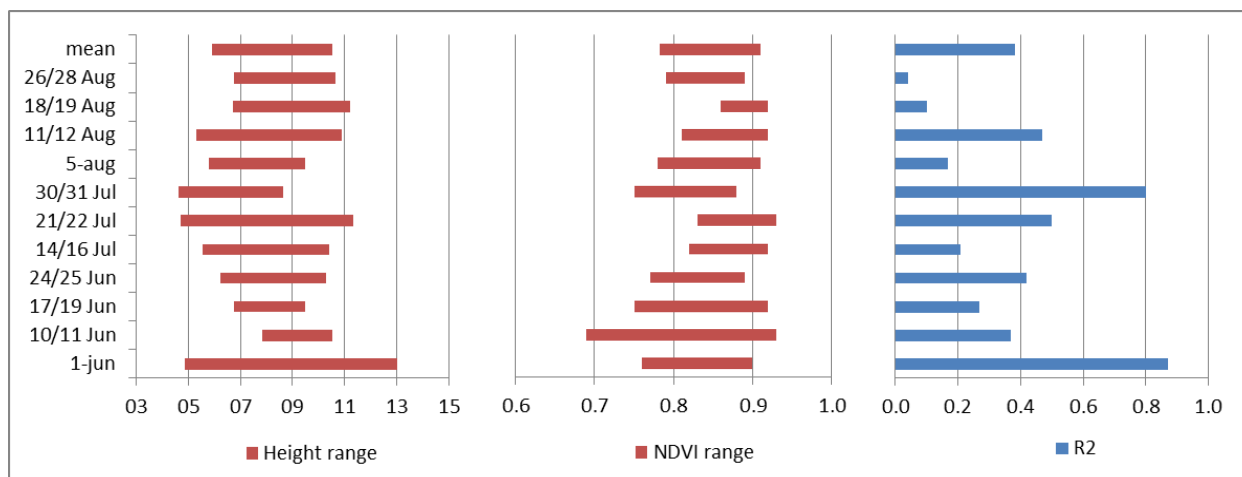


Figure 6.13 Overview of height ranges, NDVI ranges and R^2 values for the different observation dates.

From these results a number of conclusions can be drawn:

- There is a large variation in the strength of the correlation between NDVI and grass height. Figure 6.13 shows that R^2 varies between 0.04 and 0.87. This means that for some dates the NDVI value represents the grass height very well, while for others this relation is very weak.
- The variation in NDVI values is not very large. The NDVI varies between 0.69 and 0.93 over the whole period, but the variation at a specific date the NDVI variation is in the order of 0.13 (ranging from 0.06 to 0.24). The variation as found for the complete parcel C3 was larger, with a range of about 0.25 (see Figure 6.6). A cause for this might be the fact that for this parcel is mowed and not grazed, leading to shorter mowing and higher grass at mowing time.
- The variation in grass height is relatively limited. The height varies between 4.6 and 13.0cm over the whole period, while at a specific date the height variation between all sub-parcels is in the order of 4cm (varying from 2.7-8.1cm). As mentioned above it is expected that the height variation will be larger in case of mowing than in this case of grazing.
- From the set of observations it is not clear what the cause is for the large variation in the correlation strength R^2 between the NDVI and grass height. In Figure 6.13 the R^2 values and the ranges of NDVI and grass height are shown for the different observation dates. Different R^2 values are found for comparable ranges of NDVI and grass height.
- The accuracy of the grass height measurements in the field seems to have its limitations. For a number of cases the measured heights for several parcels are comparable, while from the relative NDVI values it can be seen that there should be more variation. See for example the plot for 24/25 June in Figure 6.12.

The current analysis involves measurements of different parcels at a specific observation date. Also an analysis can be made of measurements of a specific parcel at different dates. Here also the relative calibration of all measurements over time plays a role. As this is problematic for this dataset (see Section 6.3.2) it is not expected that the relations found for this case will be better. This is confirmed by the results found for a number of parcels that were evaluated.

Summarizing, it can be concluded that for the determination of the grass height/biomass with remote sensing observations it is essential that much attention is paid to the proper calibration of all observations. The limitations in calibration of the acquired UAV remote sensing observations were already mentioned in Section 6.3.2. This is even more important as the variation in grass height for grazed parcels is smaller than for mowed parcels and as the variation in NDVI for different grass heights is only small. For determining 10% of the grass height variation (10% of the range of 4cm is 0.4cm) also an accuracy of 10% in the NDVI variation is required (10% of the range of 0.13 is 0.01). The calibration aspect was not a major point of attention during the Dairy Campus experiment. In this framework it is good to realise that the NDVI is essentially more related to the biomass than to the grass height. The biomass in the current project is directly derived from the grass height with a straightforward linear formula. It would be good to determine the biomass for grass samples also in order to verify the proper derivation of biomass from grass height measurements.

6.3.5 UAS based height measurement

It is possible to compute detailed elevation models based on stereo photogrammetric techniques, as for all UAS eBee flights the photos were collected with 75% overlap. The standard processing of the data as applied by Thijssen (or actually Agrometius) with the Pix4D software already lead to Digital Surface Models (DSM). However, the detail and accuracy of these DSM was only of limited quality as these models were in fact only created as a half-product for supporting the image mosaicking process.

NLR used Agisoft PhotoScan Professional software to compute more detailed DSM's from the photo-sets. Objective was to arrive at very accurate DSM's that provide information on the grass height differences. The photos were collected with 12cm spatial resolution. In practice the accuracy of DSM's computed can be in the same order as the spatial resolution. This would be too coarse as the variation in grass length of the test parcels roughly varies from 6 to 10 cm. Nevertheless, in the DSM's obtained relative height (grass length) variations can be observed, see Figure 6.14 left down.

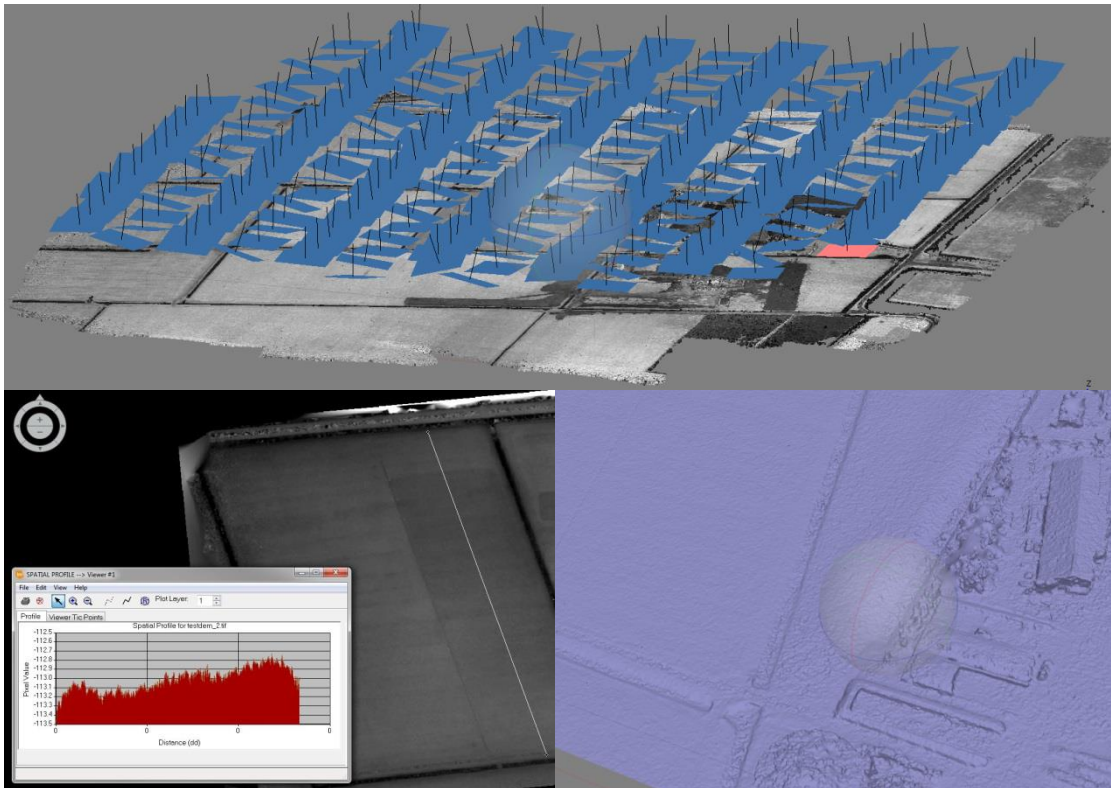


Figure 6.14 Overview of eBee DSM computed with PhotoScan. Up the result of the photo alignment. Left part of the elevation model and a height profile over the subparcels with varying grass heights. Right a shaded relief of a part of the model.

The DSM's as computed with PhotoScan were more accurate than the ones that were delivered by the standard dataset of Geometius, see the two profiles in Figure 6.15.

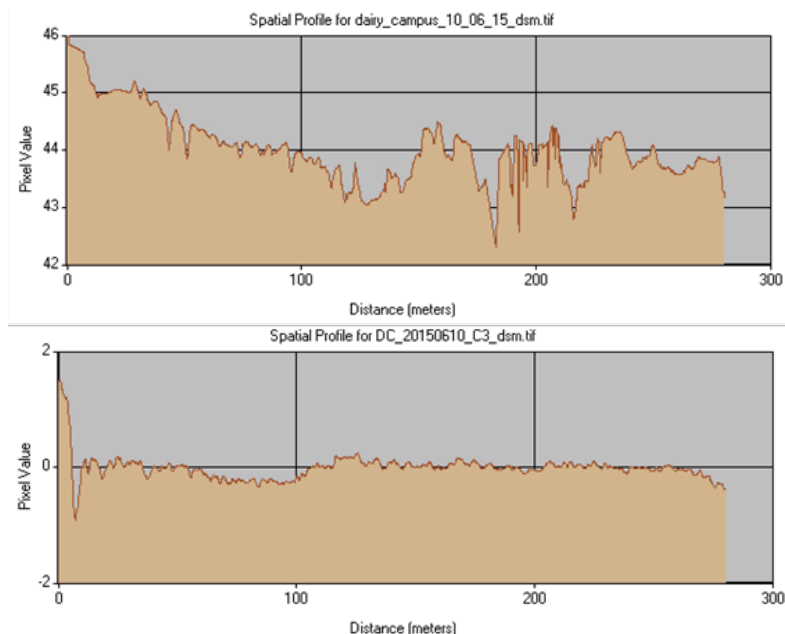


Figure 6.15 Height profiles of the same areas, up as delivered by Agrometius, down as computed with PhotoScan.

It was decided to make a further analysis of the height information as obtained from the eBee datasets. First elevation models were computed for the flight of 10 June. Because the eBee uses a separate lens for each spectral band, in fact four independent photosets are collected during one flight. For three of these photosets three independent DSM's were computed. Already from the residuals found for the gcp's in the photogrammetric model it could be seen that errors in the order of 10 to 40cm occurred. An analysis was made of the height profile along the sub-parcels BB1-15 (see Figure 6.9). Figure 6.16 shows the heights along the profile as obtained from the three individual photo sets. It can be seen that especially at the start there are large variations.

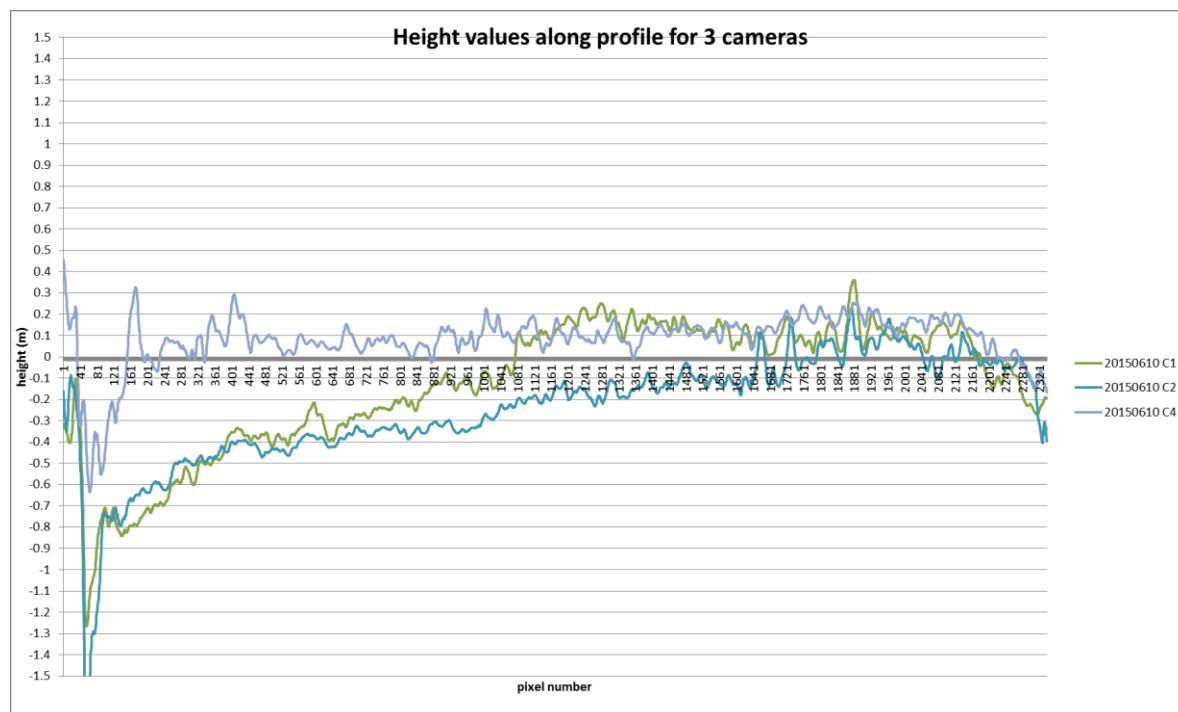


Figure 6.16 Height profiles obtained from camera C1, C2 and C4

For better co-registration of the three datasets, the geo-referencing of the photo-sets was enhanced by measuring 13 ground control points spread over the model. The ground control points were selected on roads or bridges that could be observed in the photos and from which coordinates could be determined in Top10NL reference maps and the AHN2 elevation model.

Secondly an additional height correction was given to the height profile over the sub-parcels to correct for the large differences between the three profiles. Because it is known that the first points and the last points of the profile are located on a road and should have the same height values, the profiles were recalibrated by a linear correction of all points with respect to the proper start and end values.

The result is shown in Figure 6.17. It can be seen that the profiles fit much better with respect to each other, but at the same time that deviations in the order of 10-20cm remain visible between C2 and C4, while the behaviour of C1 clearly shows large deviations in the order of 30cm in the middle of the profile.

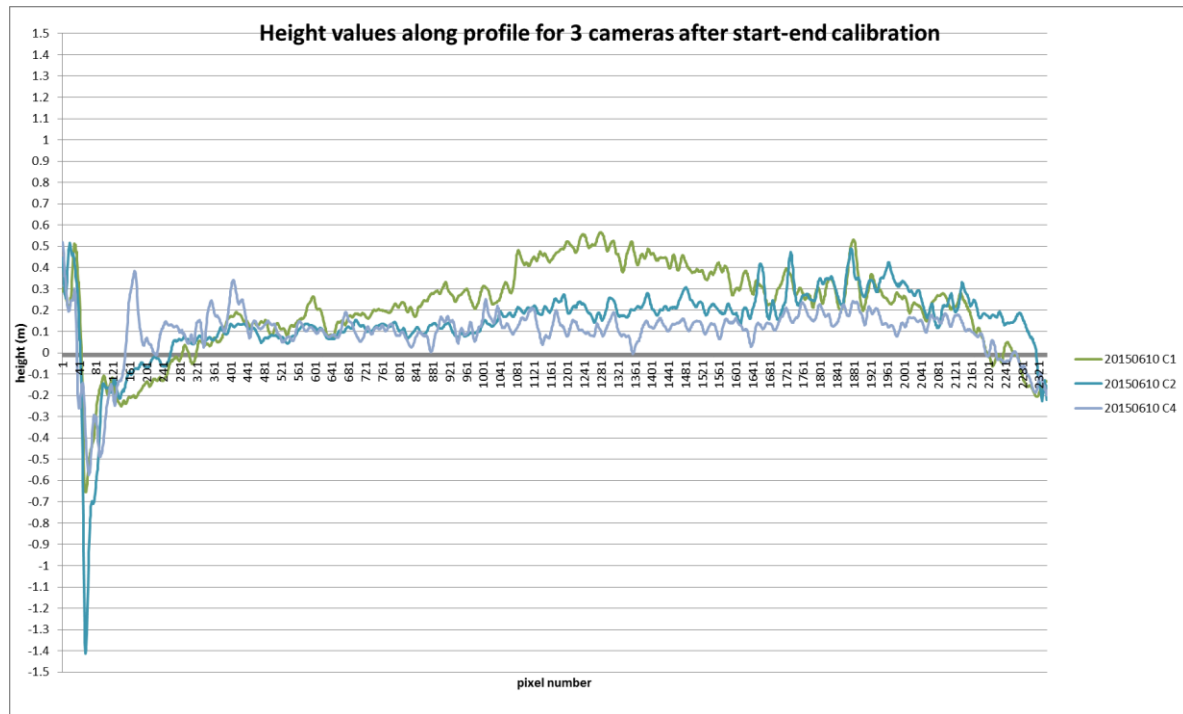


Figure 6.17 Height profiles obtained from camera C1, C2 and C4 after start-end calibration

In a next step the mean heights were computed based on all profile pixels belonging to a sub-parcel. The results are shown in Figure 6.18. It can be seen that the differences between the height values obtained with C2 and C4 are for most parcels in the order of 10cm, which is in line with the expectations. For parts of C1 the deviations are much larger, up to 40cm. These inaccuracies can be explained by the fact that half of the parcels are located at the border of the stereo model, where less overlapping photos are present and thus the geometry is weak. Also it is clear that the accuracy and positions of the ground control points used are not optimal.

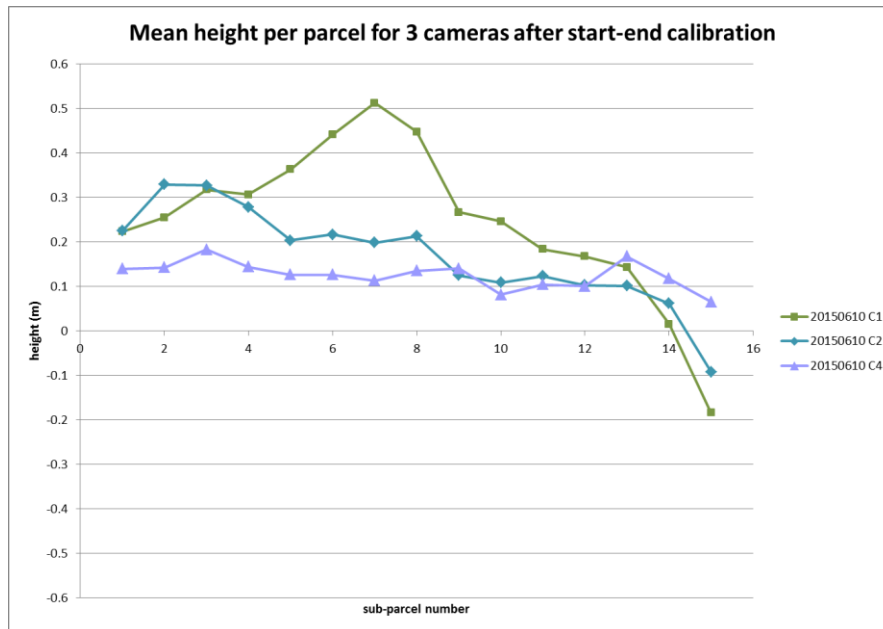


Figure 6.18 Mean heights for sub-parcel 1 to 15 as obtained with camera C1, C2 and C4.

Based on the described analysis a number of conclusions can be drawn.

- Relative variations in grass height can be observed in optimized height models from stereo UAS photosets with 10cm spatial resolution. Quantitative analysis learns that the accuracies of the height information is in the order 10 to 20cm with outliers that are much larger.
- It is expected that when higher detailed photos are collected with 1cm detail much more accurate results can be obtained in the order of 1 cm. This can be achieved by applying a lower flying height, higher zoom factor and or better cameras.
- Usage of proper geometric ground reference is required to obtain high absolute and relative geometric accuracy. This can be achieved by using accurate ground reference points and or by acquiring accurate camera positioning and orientation parameters (dGPS and high quality IMS sensors on board of the UAS).
- A point of attention is to take sufficient photo's beyond the borders of the area of interest to ensure consistent accuracy through the entire geometric model.
- Of interest might be the fact that the detailed grass height information can be used to provide additional information on local grass height variations around excreta.

6.3.6 UAS based height in relation to grass growth

Unless the weak accuracy of the obtained height profiles, height profiles from eBee data of five other dates were processed. DSM's, profiles and mean sub-parcel values were computed. The results are shown in Figure 6.19.

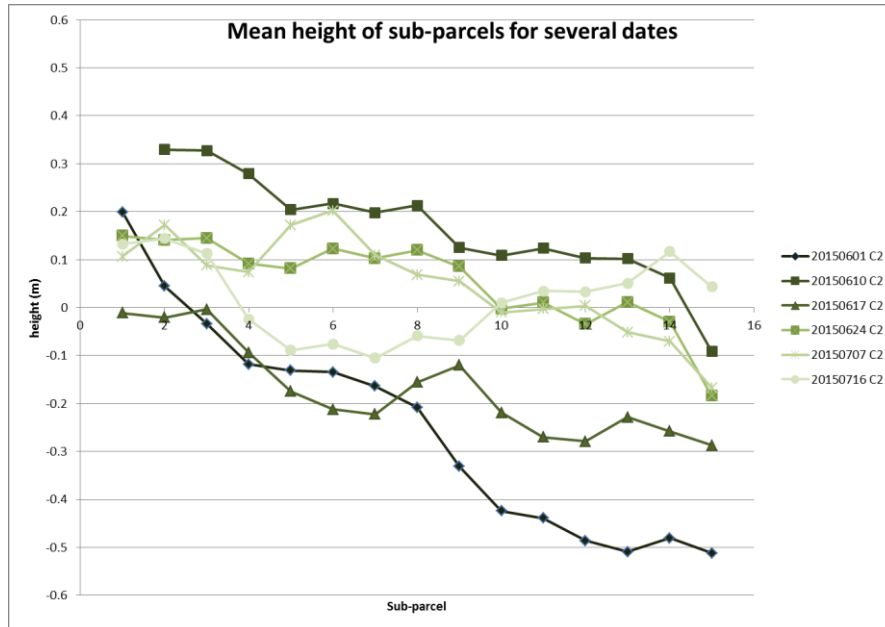


Figure 6.19 Mean heights per sub-parcel for several dates

The obtained sub-parcel elevations were also compared with the grass heights as measured with the grass height meter in the field. See Figure 6.20 where both heights are plotted against each other.

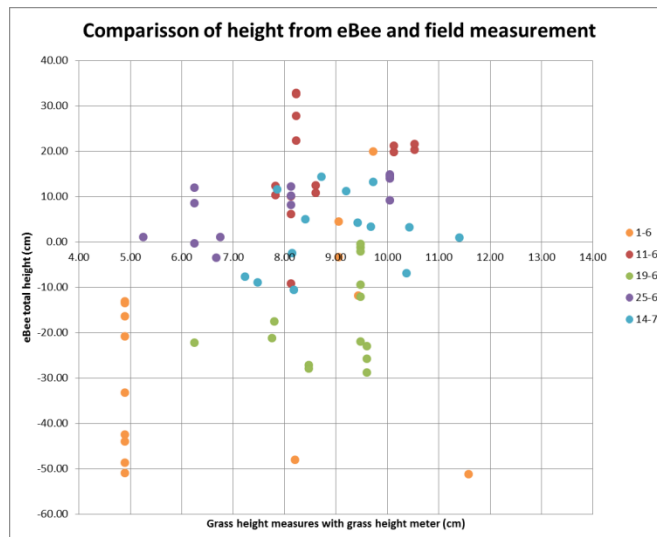


Figure 6.20 Relation between eBee height measurements and grass height meter values obtained in the field. With the used photo resolution the eBee measurements lack the required accuracy.

The result is comparable with the result from the comparison of the three profiles. No reliable relation is found between the field measurements and eBee measurements. It can be seen that the actual grass height varies between 5 and 11 cm, while the eBee height values vary between -30 and 20cm with outliers of -50cm. As described in Section 6.3.5 it is expected however that when higher detailed photos are collected with 1cm detail and a better definition of ground control points much more accurate results can be obtained.

6.4 Big Data technologies

In the framework of this project, Big Data technologies were finally applied only in a limited way. This, because much effort was first required to determine the basic relation between vegetation index and grass volume and pre-process all raw data to accurate grass height and vegetation index values. This concentrated on a small number of parcels.

As soon as this methodology is one step further, it is clear that the application of big data technology is relevant. At the one side big data technology can be applied for the handling and processing of the large remote sensing data volumes. Clear is that the data volume quickly increases when the methodology is expanded to more parcels, farms and when spatial resolutions of the observations need to increase. Already now a single drone observation involved several Gigabytes of data and for the derivation of heights it was concluded that the spatial resolution should be a factor 10 enhanced. On the other hand also big data analytics can be applied for the analysis of the relations between the grass properties and other parameters related to weather, soil, grass quality for the grass growing process, and parameters related to cow milk production and quality for the milk production process.

In the context of this project the Hadoop framework was used for the computation of mean NDVI values per parcel from all the pixels belonging to the parcel. In another project (inside NLR) the same routines were used to process satellite observations of three years for all parcels of the entire province of Flevoland (some 140.000 ha), see Figure 6.21. A Spark/Hadoop cluster was used to manage the raster data and parallelise the processing, showing that the processing time could be reduced significantly when adding more processing nodes. For this dataset also steps in big data analytics were set by correlating the remote sensing NDVI measurements with information on the vegetation types and local weather information in order to come to a predictive vegetation development model.

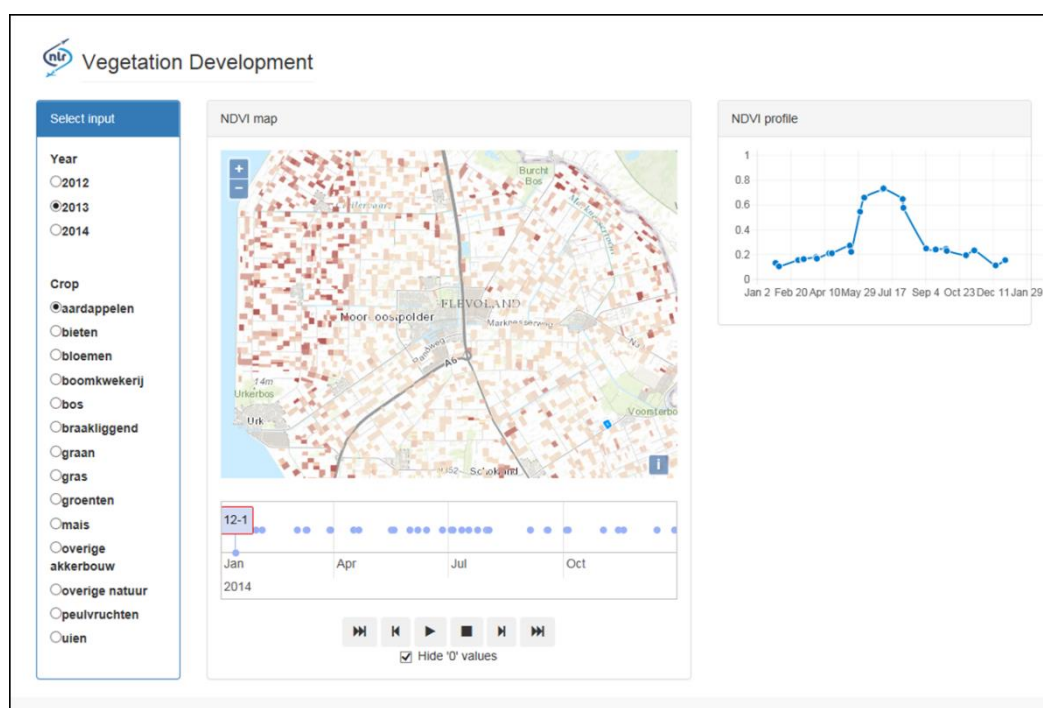


Figure 6.21 Overview of NDVI values for Flevoland parcels processed and analysed using big data technologies.

6.5 Conclusions and recommendations

Multispectral photography, either from satellite or from UAS, can be a valuable instrument for the evaluation of the normalized vegetation index (NDVI) as an indicator for the grass biomass. From this study it can be concluded however that for doing this the radiometric and atmospheric calibration of the sensor data is of big importance. Especially

because the NDVI variation for the different grass growth stages is only limited (0.6-0.8). Special points of attention for the calibration are the combination of observations from different sensors (relative calibration of different satellites and UAS) and the radiometric calibration and mosaicking of UAS data, especially when flying under cloudy circumstances. Strategies for calibration and usage of ground truth should be worked out and tested further.

Stereo photography from UAS can be a valuable instrument for the measurement of the actual grass heights. Important however is that the photos will be acquired with sufficient spatial resolution (cm level) and that attention is paid to accurate georeferencing (either by using accurate ground control points and/or by using RTK GPS positioning of the UAS). With the current 12cm photos relative variations in grass height could be observed, but the resolution was too coarse to extract accurate and reliable grass height information. As grass height variations in the order of 1cm are relevant, photos with a pixel size of 1cm need to be acquired which is well possible with current technology.

An interesting and promising aspect is that with a single UAS stereo photo flight both information can be obtained on grass NDVI and grass height. The combination of these can be input to obtain more reliable values on the grass biomass.

Big data will offer capabilities for the processing of large amounts of data and for the analysis of relations between the measured and other available parameters. The current experiment showed that the amount of remote sensing data easily is in the order of Gigabytes per observation day. When the methodology is expanded to more parcels, farms and higher spatial detail this is scaled up further. Experiments with using Spark/Hadoop technology for raster data management and processing showed that scaling of the processing times is well possible.

It is recommended to set up a next experiment based on the insights gained during this project, with higher detail UAS measurements, better collection of radiometric and geometric ground truth and good reference measurements for the grass height and weight. More systematic processing and analysis of all (big) data and relating the obtained values to other parameters (weather, fertilisation) can also be part of this.

7. Dairy Campus ICT infrastructure for the dairy farm case

7.1 Introduction

The structure of the dairy farm case is described in chapter 3. The goal of this case is to explore the possibilities to provide genotypic and phenotypic data to estimate the overall feed efficiency of dairy cows. Genotypic data are available for these cows. The availability of phenotypic data varies: data on milk yield are available per cow and milking, data on roughage and grass intake can only be estimated on an individual base. These estimations are based on an assignment of the group data to individual data. In this chapter the infrastructure of the experimental dairy farm Dairy Campus will be discussed. The current infrastructure is described and possibilities for improvement are discussed.

7.2 Data infrastructure Dairy Campus

The current implementation of data storage on Dairy Campus is depicted in Figure 7.1.

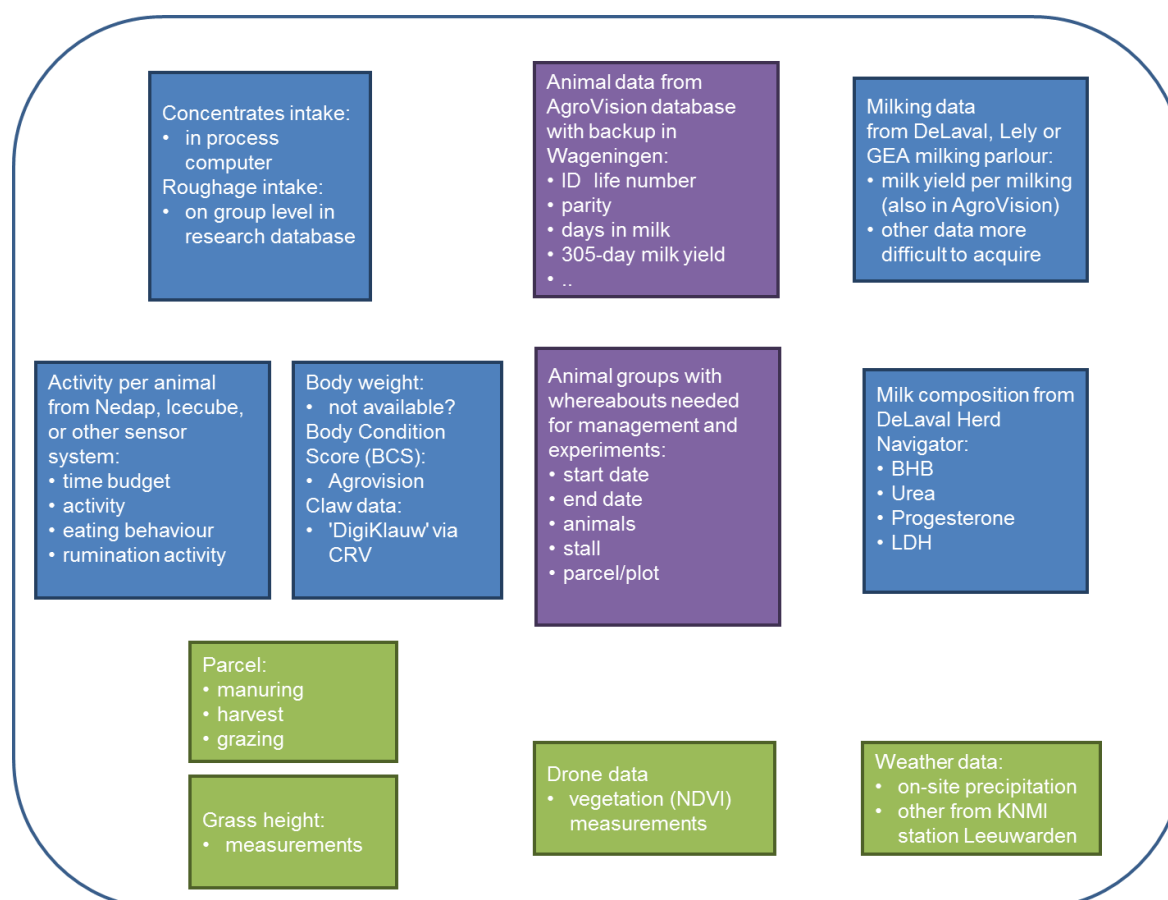


Figure 7.1 Current implementation of data storage on Dairy Campus

As depicted in figure 7.1 data are stored in different systems:

- The AgroVision database is the main system for animal and management data, it also included fertilization and harvest per parcel.
- Process computers for concentrates feeding and milking have their own storage system. Some of the data is also stored in the management system Agrovision.

- Data of specific sensor systems, e.g. for activity, are stored by researchers involved (with back-up facilities at Dairy Campus).
- Data from grassland experiments are also stored by the researchers (with backup facilities at Dairy Campus).
- The whereabouts of cows (which cows are when in which group) are recorded by the experimental farm, both for management purposes and for experiments.
- The backup database in Wageningen contains data from the AgroVision database (as well as data from process computers as imported in the AgroVision database) and the whereabouts.

A typical characteristic of the current situation is the distributed data storage. Not all data is available in one storage system. The backup database includes most management data but experimental data are not always included. There is a data co-ordinator that helps combining data needed in experiments. Most data are not accessible by a web interface, at the moment only the 'Akkerweb' system makes some data available in this way. Internet access to other data is planned for the near future. Internet access will never be public but restricted to researchers involved.

7.3 Data for the dairy farm case

The dairy cow case is described in chapter 3; the availability of the needed data is elaborated here for the different types of data.

7.3.1 Animal selection

It is possible to make a selection of lactating cows for a defined period. For the year 2015 cows involved in grazing experiments are selected. This selection is further restricted to cows for which also production and feeding data are available over the preceding years. The parity of these cows is known as well as the standard production level, days in milk. The body weights are not routinely recorded; these might only be available in experiments.

The genotypes of all cows at Dairy Campus are available in a dedicated database in Wageningen. This database is being installed and will be available in the near future. It should be checked whether access by a web interface is available. The whereabouts of the cows are available, these are recorded routinely both for farm management and experiments. Occurrences of diseases and treatments, moments in heat and inseminations are recorded in the farm management system and also available in the backup database.

7.3.2 Output

Milk yields (kg) are recorded per milking; these can be transformed to daily milk yields.

Data on the milk composition is available for test day milkings and from the Herd Navigator system. On test days not only the milk yield is recorded but also milk and protein content, cell counts and urea content. The Herd Navigator can measure progesterone, urea, LDH and BHB in milk, the measuring frequency is determined by the Herd Navigator system based on a bio model.

7.3.3 Input

Concentrates feedings are recorded by the process computer; daily totals are available in the AgroVision database (and backup database). The composition of the different types of concentrates is also available.

Roughage recordings are only available at group level. Analysis results are available per batch.

Grass intake recordings are not available, these may be estimated based on grazing data, measured grass height and analysis results for fresh grass.

For all input components, concentrates, roughage and grass, whereabouts are needed to come to individual feed intake data. Sensor data (eating, ruminating or behaviour data) may be useful to improve the redistribution of group data to individual data.

7.3.4 Data availability

The Dairy Campus situation with respect to the case and the availability of the data has been analysed for 2014 and 2015. Results are summarised in table 7.1. Per variable also the level of aggregation was mentioned.

Table 7.1 Survey of availability of animal data at Dairy Campus Leeuwarden in 2014/2015 (see explanation in text for a description of the data)

Feature	animal					group					farm				
	within day	day	week	month	year	within day	day	week	month	year	within day	day	week	month	year
<i>1 performance dairy cattle</i>															
1.1a milk yield															
1.1b additional milk measurements															
1.1c test milk data															
1.1d Herd Navigator data															
1.2a concentrates intake															
1.2b roughage intake															
1.2c grass intake															
1.2d predicted feed intake															
1.3 water intake															
1.4 body weight															
1.5a body condition scores															
1.5b locomotion scores															
<i>2 behaviour dairy cattle</i>															
2.1a activity Nedap															
2.1b behaviour IceQubes															
2.1c behaviour SensOor															
2.2 visits feeding station/milking robot															
2.3 temperature SensOor															
2.4 location Nedap															
2.5 behaviour studies															
2.6 webcam/video recordings															
2.7 sound recordings in cowhouse															
<i>3 performance young stock</i>															
3.1 milk intake															
3.2 water intake															
3.3 body weight															
3.4a concentrates intake															
3.4b roughage intake															
3.4c grass intake															
<i>4 behaviour young stock</i>															
4.1 activity Nedap															
4.2 visit milk/water supply															
4.3 location Nedap															
4.4 behaviour studies															
<i>5 other</i>															
5.1a grouping															
5.2a cow calendar data															
5.2b inseminations, oestrus cases															
5.2c health data															
5.2d medicine use															
5.3 weather data															
5.4 prices of feed, milk															

legend:

	data available		data estimated or incomplete		data not available
--	----------------	--	------------------------------	--	--------------------

Based on table 7.1 one can see that there are only 7 variables available. These are also important for normal daily management of the herd. Based on the structure of working with projects one can see that there is a big number of variables that are incomplete or missing at all. This is also the reason why we could not perform a proper calculation of the feed efficiency, the goal of our case, with actual data of the Leeuwarden Dairy campus data.

In order to be able to interpret these data and to understand the context per variable a more detailed description will be given. This can also be input for an ontology, as described in chapter 4.

Performance dairy cattle

1.1a milk yield:	is recorded for every milking; also total deliveries to the dairy factory are known
1.1b additional milk measurements:	no additional measurements in the milking parlour; in milking robot per quarter and milking: milk yield, electrical conductivity, blood, peak milk flow and average milk flow
1.1c test milk data:	fat and protein content, cell counts and urea content on day level; test milking's are once a month and sometimes more frequent during experiments
1.1d Herd Navigator data:	additional measurements in the milking robot for progesterone (an indicator for oestrus), LDH (lactate dehydrogenase, an indicator for mastitis), BHB (beta-hydroxy butyrate, an indicator for metabolic disorders) and urea (an indicator for feeding problems); measurement frequency determined by robot
1.2a concentrates intake:	is recorded in both milk robot and concentrates feeding station; intake per visit is known in the milking robot; only daily intake is known for concentrates feeder.
1.2b roughage intake:	is recorded per group per day
1.2c grass intake:	is sometimes estimated based on grass heights during grazing experiments; is also estimated based on manure sample during an alkane experiment
1.2d predicted feed intake:	is calculated by the management system
1.3 water intake:	is not recorded per drinking visit, only total water consumption on farm level is known
1.4 body weight:	is recorded incidentally
1.5a body condition scores:	are recorded by CRV on test milk days and incidentally during experiments
1.5b locomotion scores:	are recorded incidentally during experiments

Behaviour dairy cattle

2.1a activity Nedap:	activity and eating time per 15 minutes
2.1b behaviour IceQubes:	lying time, standing time, number of lying bouts and number of steps per 15 minutes; data are incidentally available during experiments
2.1c behaviour SensOor:	time spending (rumination, eating, active, highly active, not active) per hour; data are incidentally available during experiments
2.2 visits feeding station/milking robot:	feeding station visits are not recorded, visit to milking parlour and milking robot (with or without milking) are recorded
2.3 temperature SensOor:	ear temperature per hour
2.4 location Nedap:	location sensor has been used for youngstock, but not yet for dairy cattle, availability of data is a point of concern as data storage is not automatically included
2.5 behaviour studies:	incidentally during experiments
2.6 webcam/video recordings:	incidentally during experiments
2.7 sound recordings in cowhouse:	not yet available

Performance young stock

3.1 milk intake:	is recorded per visit (also unrewarded visits are recorded)
3.2 water intake:	has been recorded in some pens during the Smart Dairy Farming project (by combining water ticks of 0.25 litre with visit recordings)
3.3 body weight:	has been recorded in some pens during the Smart Dairy Farming project (by combining 5-seconds measurements with visit recordings)
3.4a concentrates intake:	is only recorded at group level incidentally during experiments

- 3.4b roughage intake: is only recorded at group level incidentally during experiments
 3.4c grass intake: not relevant as there is no grazing of young stock

Behaviour young stock

- 4.1 activity Nedap: can be recorded per quarter with the Velos system of Nedap for older young stock (especially for oestrus detection)
 4.2 visit milk/water supply: have been recorded in some pens during the Smart Dairy Farming project
 4.3 location Nedap: location sensor has been used for young stock during the Smart Dairy Farming project
 4.4 behaviour studies: incidentally during experiments

Other

- 5.1a grouping: every movement of cows between groups has been recorded; these whereabouts are essential for the analysis of sensor data
 5.2a cow calendar data: calving dates, dry-off dates and the like are recorded in the management system
 5.2b inseminations, oestrus cases: inseminations are always recorded in the management system; observed cases of oestrus might also be recorded.
 5.2c health data: occurrences of disease and treatments (including medicine use) are recorded in the management system; claw disorders are recorded in the DigiKlauw system of CRV
 5.3 weather data: there is no local weather station, but data from the nearby airport Leeuwarden are available.
 5.4 prices of feed, milk: are known in the bookkeeping system

7.4 Internet access and expected infrastructure

The data on animal selection, output and input as described in the previous paragraphs is available, but weak point may be the whereabouts (should be checked on completeness), the differences in aggregation level (it is not always evident how to transform group data to an individual level) and the accessibility of data stored in different databases. Data storage may be improved by facilitating internet access. Some possibilities like Akkerweb and InfoBroker are described here.

- **Akkerweb** is a platform for geo-based data, developed for arable farmers. Akkerweb contains data of plot level and various apps are available to handle these data. Web services are available and apps can co-operate in combined different type of data (plot info, satellite maps and weather data). Akkerweb is being used by 3400, mostly arable, farms. Scalability and performance is not an issue yet but the number of farms increases and geographical data becomes available at a higher resolution. Most users are arable farmers, but Akkerweb is also applicable for grassland and roughage (maize and others) applications in dairy farming. Akkerweb provides possibilities to share data on an existing platform and to develop and test apps, especially for geo-based data like data on grazing and grass growth. This also includes drone data for grass and maize measurements, but might also be applicable for location data of cows measured with location sensors.
- **InfoBroker** is a platform where sensor data can be exchanged based on supply and demand. Data are cow-centric (and not sensor-centric as often occurs in practice). InfoBroker is scalable and flexible in the available data and participants. It is developed in the Smart Dairy Farming project and runs on the seven involved practical farms; there is a direct connection to the databases of CRV (breeding company), TNO (applied research institute) and Gallagher (weighing equipment). Currently sensor data from different providers are made accessible for users to develop SOPs (Standard Operating Procedures) to assist the farmers in their operational management. Sensor data are diverse: activity, ruminating, behaviour, body weight, concentrates supply and intake, milk yield, milk composition, visits to milking robot and concentrates feeder, milk and water intake of calves, video observation of dry cows, pH and temperature of rumen bolus. In addition to these sensor data also cow calendar data and recorded events like occurrences of diseases or oestrus are available

in the InfoBroker. Authorization should be arranged in an agreement with the InfoBroker foundation. An interface should be developed for every new sensor type; this can be a task for the sensor supplier.

The genotype database is in development, the possibilities for internet access (e.g. by using the InfoBroker) are not clear yet.

Some concluding remarks for feasibility of internet access:

- Internet access to management and experimental data of Dairy Campus is wanted. The existing management database (AgroVision) will remain the core of the system. In the near future data from AgroVision will be available via CowVision which is running as an Internet platform.
- Akkerweb is a ready-to-use platform; advantages are possibilities for visualisation of geo-based data and the availability of an app store.
- InfoBroker might be useful for data exchange between Dairy Campus and Wageningen UR in Wageningen. Authorization is needed and interfaces should be developed if not yet available. Cooperation with manufacturers is needed.
- Preferably also the genotype database should include internet access.

Figure 7.2 shows the potential situation for Dairy campus when the Infobroker and Akkerweb platform will be adopted.

★ = InfoBroker compliant

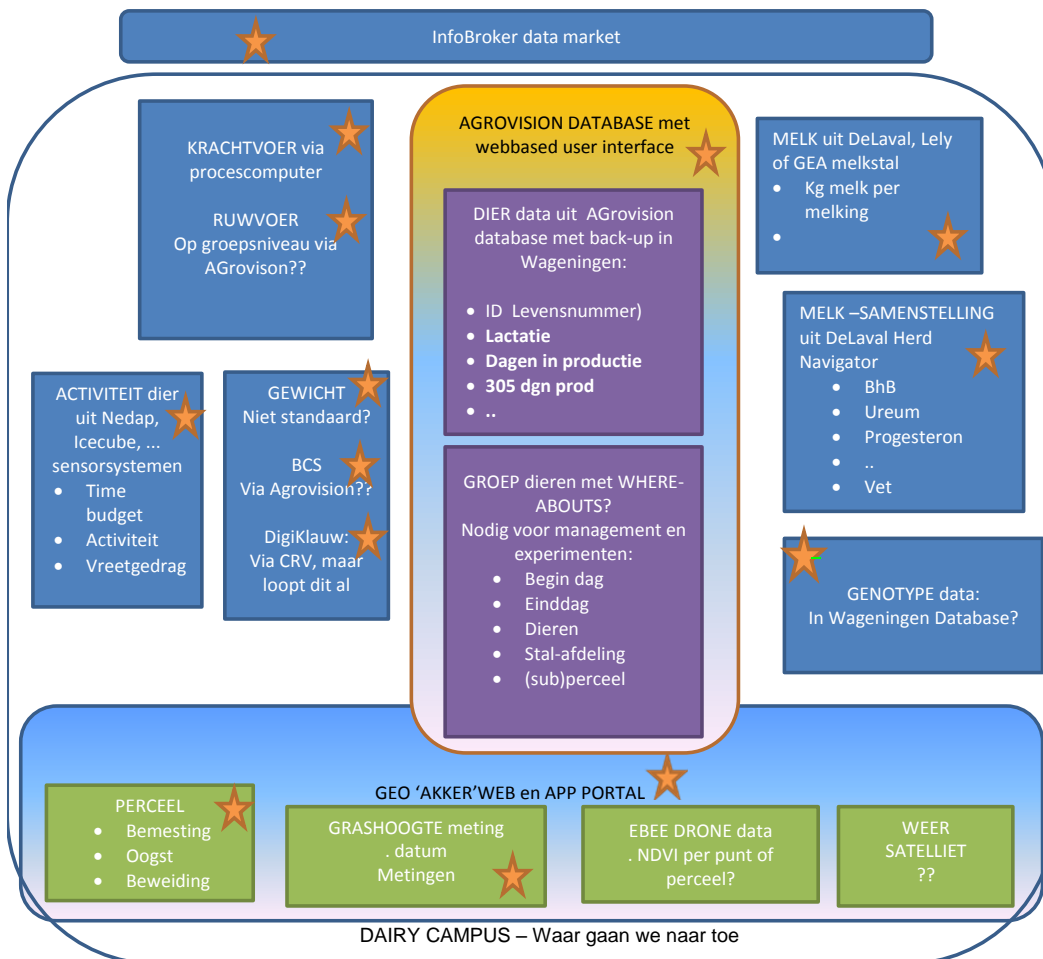


Figure 7.2 Potential data infrastructure for Dairy Campus (in Dutch).

8. General discussion and insights

In this chapter we integrate the work done on specific issues and topics as discussed in the previous chapters. The overall objective will be discussed, the generic integral framework and the specific use case.

Answer to objective:

The goal of this project is to create an integration platform for big data analysis for smart farming and to develop a show case. This includes both technical (hard /software) and organizational integration (developing business ecosystem) and combining and linking of data and models. DLO, NLR and TNO work together in 2015 towards the realization of an IT data infrastructure that makes it possible to solve data to connect from different sources and models in an effective and safe way, ontology problems, specific analysis tools develop, opportunities and risks to identify and assess the acquired knowledge and experience and present it in a smart farming show case, from 'grass to glass'.

The creation of an integration platform for big data analysis for smart farming was far too ambitious for a one year activity with emphasis only on one use case. However, by doing this project we brought different kind of expertise and background together to explore parts of the solution. As described in the chapters it is much more complex than expected.

Smart farming puts emphasis on the ICT- and decision making part of precision farming. It looks like a buzzword used to describe the concept of intelligent use of data-rich ICT-services and ICT-applications. It is presented as an extra on top of the concepts of precision agriculture and precision livestock farming. Smart farming has the potential to contribute to more sustainable agriculture. And big data use, if established, will support smart decisions and management.

Organisations are evaluating to invest in big data technology and use? Big data technology represents a disruptive innovation that market orientated organisations will use to drive competitive advantage and governmental bodies to set and reach policy targets. The value of big data lies in the information and insight that organisations can draw from it, rather than in the data itself. Linking physical and socio-economic data, for example, may generate entirely new insights and market opportunities. So, the impact of big data for smart farming outreaches the impact of a single farmer and his processes, although at farm level maybe most of the big data will be created by implementing new sensing techniques that are high data sensitive.

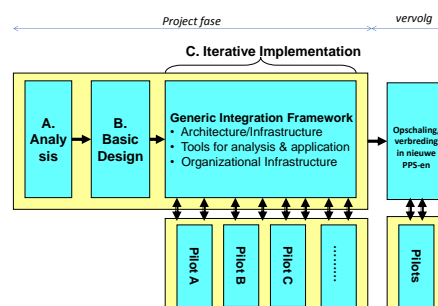
The perspective arises that use of big data technology has the potential to dramatically change organisations. It will alter data availability, knowledge creation, decision making, production optimization and competitiveness. A fair question is: "Will big data bring real benefits to organisations and society, or will it end up a hype with only one or two companies benefiting from it?". We can answer this question within ca. 20 years from now. Presently, the authors of this report have positive expectations of big data applications in the field of smart farming.

Generic Integral Framework:

In the generic integration framework three important aspects can be discussed. These are 1) the architecture and infrastructure, 2) the tools for analysis and application and 3) the organizational infrastructure.

Architecture and infrastructure

In chapter 3 and 7 the architecture and infrastructure of the dairy farm case are discussed. Basically the actual architecture is a mix form of island automation. There is no common architecture available. Beside architecture there is also limited effort put into alignment of terminology. Data are harvested for either managing the herd or for dedicated projects. This goes so far that also data availability follows the same trend. So for Big data analysis a huge effort is needed in streamlining architecture and infrastructure. In chapter 6 proposals were made for a more web-based architecture by using Infobroker and Akkerweb platforms. In chapter 4 alternatives were looked at. Linked data concepts can also be used to exchange data more easily data. However, there is a lack of uniformity in practical definitions to be part of an ontology and the performance in big networks is still unpredictable.



Tools for analysis and application

In chapter 5 historical cow production data were combined with additional data from weather information. This to create a more big data like dataset with variable data sources. From the potential set of machine learning techniques the Artificial Neural Networks were used because this technique fits best to the data. In itself it took already quite some energy to reconstruct data so that they could be analysed. Also the choice for the used technique seems to be rather intuitively. The lesson from this part is that there is still a big effort needed for data inspection, data cleaning and model selection. This looks very much like 'traditional' statistical analysis and urges the need for more alignment in data architecture and semantics. To elaborate more on the feed intake of grass in chapter 6 data from remote sensing were used. It is quite easy to increase the amount of data when you work with remote sensing camera's from satellites and drones. In the end it became clear that these data still do not have the proper resolution and the knowledge is still too limited to estimate properly grass growth.

Organisational infrastructure

In chapter 2 the organisational infrastructure is addressed. In itself dealing with BigData needs cooperation between all kind of different organisations. In the literature review the change in relations between venture capitalist, tech and data start-ups, AgTech and AgBusiness multinationals, farmers, and tech Companies is addressed. Most examples come from arable farming and food processing. There is a big gap between these developments and the use case we addressed in this report. The use case and organisations in this project have a research background. In the literature review the role of research organisations and the role of education is not addressed separately. The use case was complex and challenging enough to stimulate the discussion between the different research institutes and multidisciplinary scientist, and also to get some extra experience in specific topics, but in the end the use case was too complicated to bring data together, analyse them and give a proper answer on feed efficiency.

Insights:

From the different chapters some insights can be obtained.

- To be active as research institute and owner of research data in the field of bigdata you must realise that also data infrastructure and data policy should be adapted.
- When data have to be connected and analysed by people with different background there is a need for adequate description of the data and terminology. There should be initiative to work on a smart farming ontology.
- Big data applications in farming are not strictly about primary production, but play a major role in improving the efficiency of the entire supply chain and alleviating food security concerns.
- Expected is that big data has focus on in-depth, advanced, game-changing business analytics, at a scale and speed that the old approach of copying and cleansing all of it into a data warehouse is no longer appropriate, but experience on use cases is that presently exchange of data through new platforms or open data is not feasible yet.
- While in the past advisory services were based on general knowledge that once was derived from research experiments, there is an increasing need for information and knowledge that is generated on-farm in its local-specific context. It is expected that Big Data technologies help to achieve these goals in a better way (Pope et al., 2015; Sonka, 2015), but in this project the use case was still too difficult to get a proper answer.
- We support the remark of Lesser (2014) and Orts & Spigonardo (2014) that the availability and quality of the data is often poor and needs to be ensured before you can make use of it.
- It was possible to create and verify a common ontology for the domain and mapped it to the existing ontologies of the different accessible data sources. However, not all tools were yet able to find the relation between "parcel", "perceel" and "plot".
- An expected advantage of machine learning techniques, like the used Artificial Neural Network, is that the amount of domain knowledge required is minimal and that the resulting models (when set up and validated correctly) can model very complex (nonlinear) systems. However, interpretation by domain experts still was needed in the use of ANN.
- Using images for drones drastically increased volume of data and resolution of vegetation indexes, but still were not capable for an adequate estimation of grass biomass in the field. In the (limited) framework of this project it turned out that Big Data technologies were applied only marginally. This, because much effort was required to pre-process all raw data to accurate grass height and vegetation index values and determine the basic relation between vegetation index and grass volume.

- It is recommended to put attention to the common understanding about the best way to calibrate satellite- and UAV data in the Netherlands.
- The data infrastructure and policy of the Dairy Campus has to be reconsidered in such a way that more project related data should be harvested routinely in order to be an interesting partner in BigData research.

From the literature review in chapter 2 and supported by the experiments in the other chapters the following challenges need to be addressed:

- *Data ownership* and related *privacy* and *security* issues – these issues have to be properly addressed, but when this is applied too strictly it can also slow down innovations;
- *Data quality* - which has always been a key issue in farm management information systems, but is more challenging with big, real-time data;
- *Intelligent processing* and *analytics* – for Big Data this is also more challenging because of the large amount of often unstructured, heterogeneous data which requires a smart interplay between skilled data scientists and domain experts;
- *Sustainable integration* of Big Data sources – integration of many different data sources is challenging but because this is crucial for your business model this has to be done in a sustainable manner;
- *Business models* that are attractive enough for solution providers but that also enable a fair share between the different stakeholders;
- *Openness of platforms* that will accelerate solution development and innovation in general but also empower farmers in their position in supply chains.

The promise of Big Data in agriculture is alluring, but the challenges above have to be addressed for increased uptake of Big Data applications. Although there are certainly technical issues to be resolved we recommend to focus first on the governance issues that were identified and design suitable business models because these are currently the most inhibiting factors.

Literature

Related literature and references chapter 1

- DLO, 2015. Big data technologies and methodologies. Chapter from strategic plan Wageningen UR. 14 p.
- FAO, 2009. Feeding the world. World summit on Food Security, FAO, Rome, 16-18 Nov. 2009.
<ftp://ftp.fao.org/docrep/fao/meeting/018/k6021e.pdf>
- Kempenaar, C., Kocks, C.G., 2013. Van Precisielandbouw naar Smart Farming Technology. Inaugural Speech, Lectoraat Precisielandbouw. CAH Viltentum, Dronten. 29 May 2013.
- Lokhorst, C. and A.H. Ipema, 2010, Precision livestock farming for operational management support in livestock production chains. In: Towards effective food chains: models and applications, edited by Jacques Trienekens, Jan Top, Jack van der Vorst and Adrie Beulens, published by Wageningen Academic Publishers, p 293-308.
- Needle, D., 2015. Big Data or Big Disappointment? Experts Debate Hype Versus Reality.
<http://www.eweek.com/database/big-data-or-big-disappointment-experts-debate-hype-versus-reality.html>.
- TNO, 2015. TNO ICT Roadmap excerpts. TNO. 8.p.
- Top, J., Wigham, M., 2015. The role of e-science in agriculture. A study of how e-science technology assists participation in agricultural research. Wageningen UR Food & Biobased Research.
https://ec.europa.eu/eip/agriculture/sites/agri-eip/files/field_core_attachments/scar_swg_akis_3_-_final_report.pdf
- Trienekens J, J. Top, J. van der Vorst, A. Beulens, 2010. Towards effective food chains: models and applications, published by Wageningen Academic Publishers, ISBN 978-90-8686—148-4, 320 pp.
- Voskamp-Harkema, W., Lokhorst, C, 2015. Duurzame en zorgvuldige melkveehouderij met Smart Dairy Farming. Inaugural Speech, Duurzame en zorgvuldige melkveehouderij. VHL, Leeuwarden. 15 January 2015.
- Wolfert, S., Kempenaar, C., 2012. The role of ICT for Future Agriculture and the role of Agriculture for Future ICT. Abstract and presentation at 6th International Weed Science Congress, Hangzhou, China, June 17-22, 2012.
www.iwss.info/downloads/files/n50d33368b217e.pdf.

Related literature and references chapter 2

- AgGateway, 2014. Data Privacy and Use White Paper.
<http://www.aggateway.org/WorkingGroups/Committees/DataPrivacySecurity.aspx>. Accessed: 7 May 2015
- Al-Debei, M.M., Avison, D., 2010. Developing a unified framework of the business model concept. Eur J Inf Syst 19, 359-376.
- Anonymous, 2012. Dairy industry in era of big data, in: Western Farm Press,
<http://westernfarmpress.com/markets/dairy-industry-era-big-data>. Accessed: 7 May 2015.
- Anonymous, 2014a. Coalition of Ag Groups, Companies Reach Data Agreement, in: Farm Industry News,
<http://farmindustrynews.com/precision-farming/coalition-ag-groups-companies-reach-data-agreement>. Accessed: 7 May 2015.
- Anonymous, 2014b. The Digital Transformation of Row Crop Agriculture: A report to the Iowa AgState Group. The Hale Group, Ltd; LSC international, Inc.
- Anonymous, 2014c. Technology helps farmers to cater climate changes effects, in: Flare, <http://www.flare.pk>. Accessed: 7 May 2015.
- Ault, A.C., Krogmeier, J.V., Buckmaster, D., 2013. Mobile, Cloud-Based Farm Management: A Case Study with Trello on My Farm, 2013 Kansas City, Missouri, July 21-July 24, 2013. American Society of Agricultural and Biological Engineers.
- Barmponakis, S., Kaloxylas, A., Groumas, A., Katsikas, L., Sarris, V., Dimtsa, K., Fournier, F., Antoniou, E., Alonistioti, N., Wolfert, S., 2015. Management and control applications in Agriculture domain via a Future Internet Business-to-Business platform. Information Processing in Agriculture 2, 51-63.
- Beer, S., 1981. Brain of the Firm: Second Edition. John Wiley, London and New York.

- Bennett, J.M., 2015. Agricultural big data: utilisation to discover the unknown and instigate practice change. *Farm Policy Journal* 12, 43-50.
- Burrus, D., 2014. Who Owns Your Data?, in: Business2Community.com, <http://www.business2community.com/big-data/owns-data-0793456#fX2ep06aoxZHQZdk.97>. Accessed: 7 May 2015.
- Carlson, C., 2012. Fujitsu rolls out cloud-based Big Data platform for farmers, in: FierceCIO, <http://www.fiercecio.com/story/fujitsu-rolls-out-cloud-based-big-data-platform-farmers/2012-07-19>. Accessed: 7 May 2015.
- Chen, M., Mao, S., Liu, Y., 2014a. Big Data: A Survey. *Mobile Netw Appl* 19, 171–209.
- Chen, M., Mao, S., Liu, Y., 2014b. Big Data: A Survey. *Mobile Networks and Applications* 19, 171-209.
- Data FAIRport, 2014. Find Access Interoperate Re-use data. <http://www.datafairport.org/>. Accessed: 2 August 2016
- Davenport, T.H., 1993. *Process Innovation: Reengineering Work through Information Technology*. Harvard Business School Press, Boston, Massachusetts.
- Davenport, T.H., Short, J.E., 1990. The New Industrial-Engineering - Information Technology and Business Process Redesign. *Sloan Management Review* 31, 11-27.
- De Mauro, A., Greco, M., Grimaldi, M., 2016. A formal definition of Big Data based on its essential features. *Library Review* 65, 122-135.
- Devlin, B., 2012. The Big Data Zoo—Taming the Beasts: The need for an integrated platform for enterprise information. 9sight Consulting.
- Drucker, V., 2014. Agriculture springs into the digital age, in: Fund Strategy, <https://www.fundstrategy.co.uk/issues/fund-strategy-sept-2014/agriculture-springs-into-the-digital-age/>. Accessed: 7 May 2015.
- Dumbill, E., 2014. Understanding the Data Value Chain, in: Big Data & Analysis Hub, <http://www.ibmbigdatahub.com/blog/understanding-data-value-chain>. Accessed: 02 August 2016.
- Esmeyjer, J., Bakker, T., Ooms, M., Kotterink, B., 2015. Data-driven innovation in agriculture: Case study for the OECD KBC2-programme. TNO report TNO 2015 R10154.
- Faulkner, A., Cebul, K., 2014. Agriculture gets smart: the rise of data and robotics, Cleantech Agriculture Report. Cleantech Group.
- Fenn, J., LeHong, H., 2011. Hype cycle for emerging technologies, 2011. Gartner, July.
- Gilpin, L., 2015. How big data is going to help feed nine billion people by 2050, in: TechRepublic, <http://www.techrepublic.com/article/how-big-data-is-going-to-help-feed-9-billion-people-by-2050/>. Accessed: 7 May 2015.
- Guild, M., 2014. Big Data Comes to the Farm, in: Financial Sense, <http://www.financialsense.com/contributors/guild/big-data-farm>. Accessed: 7 May 2015.
- Haire, B., 2014. Ag data: its value, who owns it and where's it going?, in: Southeast Farm Press, <http://southeastfarmpress.com/cotton/ag-data-its-value-who-owns-it-and-where-s-it-going>. Accessed: 7 May 2015.
- Hardy, Q., 2014. A Low-Cost Alternative to Pricy Big Data on the Farm, in: The New York Times Blogs, http://bits.blogs.nytimes.com/2014/12/01/a-low-cost-alternative-to-pricy-big-data-on-the-farm/?_r=0. Accessed: 7 May 2015.
- Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Ullah Khan, S., 2015. The rise of “big data” on cloud computing: Review and open research issues. *Information Systems* 47, 98-115.
- Holmes, M., 2014. Different Industries Debate the Potential of UAVs and the Need for Satellite, in: Via Satellite Integrating SatelliteToday.com, <http://www.satellitetoday.com/technology/2014/10/24/different-industries-debate-the-potential-of-uavs-and-the-need-for-satellite/>. Accessed: 7 May 2015.
- in 't Veld, J., 2002. *Analyse van organisatieproblemen*, 8th edition ed. Stenfert Kroese.
- Ishii, K., 2014. Big data analysis in medicine, agriculture and environmental sciences. *Seibutsu kogaku Kaishi* 92, 92-93.
- Kaloxylou, A., Eigenmann, R., Teye, F., Politopoulou, Z., Wolfert, S., Schrank, C., Dillinger, M., Lampropoulou, I., Antoniou, E., Pesonen, L., Huether, N., Floerchinger, T., Alonistioti, N., Kormentzas, G., 2012. Farm management systems and the Future Internet era. *Computers and electronics in agriculture* 89, 130-144.
- Kaloxylou, A., Groumas, A., Sarris, V., Katsikas, L., Magdalinos, P., Antoniou, E., Politopoulou, Z., Wolfert, S., Brewster, C., Eigenmann, R., Maestre Terol, C., 2014. A cloud-based Farm Management System: Architecture and implementation. *Computers and Electronics in Agriculture* 100, 168-179.
- Kshetri, N., 2014. The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns. *Big Data & Society* 1.
- Lambert, D.M., Cooper, M.C., 2000. Issues in supply chain management. *Industrial Marketing Management* 29, 65-83.

- Lane, J., 2015. Digital Soil: the Four Secrets of the New Agriculture, in: Biofuels Digest, <http://www.biofuelsdigest.com/bdigest/2015/03/09/the-four-secrets-of-the-new-agriculture/>. Accessed: 7 May 2015.
- Layton, A.W., Balmos, A.D., Sabpisa, S., Ault, A., Krogmeier, J.V., Buckmaster, D., 2014. ISOBlue: An Open Source Project to Bring Agricultural Machinery Data into the Cloud, 2014 Montreal, Quebec Canada July 13–July 16, 2014. American Society of Agricultural and Biological Engineers.
- Lazzarini, S.G., Chaddad, F.R., Cook, M.L., 2001. Integrating supply chain and network analyses: The study of networks. *Chain and network science* 1, 7-22.
- Lesser, A., 2014. Big Data and Big Agriculture. Gigaom Research, p. 11.
- Li, X., Chen, S., Guo, L., 2014. Technological innovation of agricultural information service in the age of big data. *Journal of Agricultural Science and Technology* 16, 10-15.
- NEC, 2014. NEC and Dacom Collaborate on Precision Farming Solution to Maximize Yields and Reduce Costs. <https://www.nec-enterprise.com/news/Latest-press-releases/NEC-and-Dacom-collaborate-on-precision-farming-solution-to-maximize-yields-and-reduce-costs-791>. Accessed: 7 May 2015
- Needle, D., 2015. Big Data or Big Disappointment? Experts debate hype versus reality, in: eWeek, <http://www.eweek.com/database/big-data-or-big-disappointment-experts-debate-hype-versus-reality.html>. Accessed: 2 August 2016.
- Noyes, K., 2014. Big data poised to change the face of agriculture, in: Fortune Data, <http://fortune.com/2014/05/30/cropping-up-on-every-farm-big-data-technology/>. Accessed: 7 May 2015.
- Orts, E., Spigonardo, J., 2014. Sustainability in the age of Big Data. IGEL/Wharton, University of Pennsylvania, Pennsylvania, US, p. 16.
- Osterwalder, A., 2004. The Business Model Ontology: A Proposition in a Design Science Approach, l'Ecole des Hautes Etudes Commerciales. Université de Lausanne.
- Plume, K., 2014. The Big Data bounty: U.S. startups challenge agribusiness giants, in: CloudTweaks, <http://www.reuters.com/article/us-usa-farming-startups-idUSKCN0HX0C620141008>. Accessed: 7 May 2015.
- Poppe, K.J., Wolfert, J., Verdouw, C.N., Renwick, A., 2015. A European perspective on the economics of Big Data. *Farm Policy Journal* 12, 11-19.
- Porter, M.E., 1985. *Competitive Advantage: Creating and Sustaining Superior Performance*. The Free Press, New York.
- Porter, M.E., Heppelmann, J.E., 2014. How Smart, Connected Products are transforming competition. *Harvard Business Review* November 2014, 65-88.
- Provan, K.G., Kenis, P., 2008. Modes of network governance: Structure, management, and effectiveness. *Journal of public administration research and theory* 18, 229-252.
- Royse, R., 2014. The growing field of agriculture technology, in: Vator News, <http://vator.tv/news/2014-05-14-the-growing-field-of-agriculture-technology>. Accessed: 7 May 2015.
- Semantic Community, 2015. Big Data Science for Precision Farming Business. http://semanticcommunity.info/Data_Science/Big_Data_Science_for_Precision_Farming_Business. Accessed: 2 August 2016
- Sonka, S., 2015. Big Data: From Hype to Agricultural Tool. *Farm Policy Journal* 12, 1-9.
- Sonka, S., IFAMR, I., 2014. Big Data and the Ag sector: More than lots of numbers. *International Food and Agribusiness Management Review* 17, 1.
- Sun, Z., Du, K., Zheng, F., Yin, S., 2013. Perspectives of research and application of big data on smart agriculture. *Journal of Agricultural Science and Technology* 15, 63-71.
- Sundmaeker, H., Verdouw, C., Wolfert, S., Pérez Freire, L., 2016. Internet of Food and Farm 2020, in: Vermesan, O., Friess, P. (Eds.), *Digitising the Industry - Internet of Things connecting physical, digital and virtual worlds*. River Publishers, Gistrup/Delft, pp. 129-151.
- Tien, J.M., 2013. Big Data: Unleashing information. *J. Syst. Sci. Syst. Eng.* 22, 127-151.
- Tong, L., Hong, T., JingHua, Z., 2015. Research on the big data-based government decision and public information service model of food safety and nutrition industry. *Journal of Food Safety and Quality* 6, 366-371.
- UNECE, 2013. Classification of types of Big Data. <http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data>. Accessed: 2 August 2016
- Van 't Spijker, A., 2014. *The new oil - using innovative business models to turn data into profit*. Technics Publications, Basking Ridge.

- van der Vorst, J., Beulens, A., Beek, P.v., 2005. Innovations in Logistics and ICT in Food Supply Chain Networks, in: Jongen, W.M.F., Meulenbergh, M.T.G. (Eds.), Innovations in agri-food systems: Product quality and consumer acceptance. Wageningen Academic Publishers, 2005.
- Verdouw, C.N., Beulens, A.J.M., Reijers, H.A., van der Vorst, J.G.A.J., 2015. A Control Model for Object Virtualization in Supply Chain Management. *Computers in Industry* 68, 116–131.
- Verhoosel, J., van Bekkum, M., Verwaart, T., 2016. HortiCube: A Platform for Transparent, Trusted Data Sharing in the Food Supply Chain. *Proceedings in Food System Dynamics*, 384-388.
- Vogt, W., 2013. Looking at Big Data one plant at a time, in: *Farm Industry News*, <http://farministrynews.com/blog/looking-big-data-one-plant-time>. Accessed: 7 May 2015.
- Welte, J.T., Ault, A.C., Bowman, C., Ellis, S., Buckmaster, D.R., Ess, D., Krogmeier, J.V., 2013. An Approach to Farm Management Information Systems Using Task-Specific, Collaborative Mobile Apps and Cloud Storage Services, 2013 Kansas City, Missouri, July 21-July 24, 2013. American Society of Agricultural and Biological Engineers, p. 1.
- Wikipedia, 2016. Big Data. https://en.wikipedia.org/wiki/Big_data. Accessed: 2 August 2016
- Williamson, O.E., 1996. *The Mechanisms of Governance*. Oxford University Press.
- Wolfert, J., Sørensen, C.G., Goense, D., 2014. A future internet collaboration platform for safe and healthy food from farm to fork, Global Conference (SRII), 2014 Annual SRII. IEEE, San Jose, CA, USA, pp. 266 - 273.
- Yang, C., 2014. Big data and its potential applications on agricultural production. *Crop, Environment & Bioinformatics* 11, 51-56.
- Young, L., 2016. Farmobile takes ag data to the next level, in: *Agrinews.com*, http://www.agrinews.com/news/minnesota_news/farmobile-takes-ag-data-to-the-next-level/article_5074de27-ecdf-5dd3-a89e-e3b28815475c.html. Accessed: 7 May 2015.

Related literature and references chapter 4

- [1] Otero-Cerdeira, L., Rodriguez-Martinez, F.J., Gomez-Rodriguez, A.: Ontology matching: A literature review. *Journal on Expert Systems with Applications*, 949-971 (2015).
- [2] Ontology matchings tool overview: www.mkbergman.com/1769/50-ontology-mapping-and-alignment-tools/
- [3] Hertling, S., 2012. Hertuda results for OAEI 2012. In *Ontology Matching 2012 workshop proceedings*, 141-144 (2012).
- [4] HerTUDA download: www.ke.tu-darmstadt.de/resources/ontology-matching/hertuda
- [5] AgreementMakerLight website: somer.fc.ul.pt/aml.php
- [6] LogMap website: www.cs.ox.ac.uk/isg/tools/LogMap/
- [7] YAM++ website: www.lirmm.fr/yam-plus-plus

