UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Collection 'New Sources and New Technologies' Session 1

14-16 October 2019, Geneva, Switzerland 25 February 2020

Exploring precision farming data: a valuable new data source? A first orientation

Tim Punt, Ger Snijkers (Statistics Netherlands)

g.snijkers@cbs.nl

Abstract and Paper

Businesses are ever innovating. New business processes, depending on the industry, are heavily data-driven nowadays. A fine example of such a development is precision farming, where sensors aid farmers in their business operations. The generated data are full of information that might also be useful for National Statistical Institutes (NSI) and used instead of collecting data via survey questionnaires. In that case, response burden on the companies could be greatly reduced. Moreover, the sensor data could contain information that traditionally was too detailed and technical to ask for in questionnaires. Furthermore, these data may be used in timely benchmark indicators in complement to statistics. In theory sensor data thus might be a valuable new source for official statistics.

In order to study these research goals, a case study has been initiated in which CBS worked together with an innovative farmer and the Eindhoven University of Technology. The farmer has made a selection of their (sensor) data available to CBS with the purpose of exploring the data for overlap with relevant surveys and other promising aspects. First of all, the data generating process has been examined as well as the data storage systems in place for the industry. Next the overlap between the generated data and data asked for in questionnaires is studied to get an idea of the options. Also, the data stored in crop registration systems is taken into account.

The conclusion is that this data source may be valuable, but there is still a long way to go. Challenges that have surfaced during research include e.g. data quality issues and missing meta-information. These challenges will be discussed. In addition, a number of other criteria that will be paramount in the statistical use of these new data sources will be examined in the context of the case study. These include e.g. the ubiquity of these data among farmers, harmonisation of data definitions, data access, and stability of data definitions and delivery over time, to name a few. Even though it is still a long way to go, we feel that now is the time to start examining and discussing the challenges and options with regard to this new data source in order to be ready for the future.

Keywords

Precision farming, Sensor data, Response burden, Data quality, Metadata management, Data access

Statistics Netherlands

Exploring precision farming data: a valuable new data source? A first orientation

Tim Punt, Ger Snijkers, and Sofie de Broe (contact: g.snijkers@cbs.nl)[1]
Statistics Netherlands, Heerlen, 9 October 2019.
Paper presented at the 2019 UNECE Workshop on Statistical Data Collection 'New sources and New technologies', 14-16 October 2019, Geneva, Switzerland.[2]

1. Introduction

1.1. Evolution of business data collection

Historically, data for statistics were gathered via a census approach: every individual within a population was observed. Since this approach is very costly and time consuming, the method of sampling was suggested at the end of the 19th century by Anders Nicolai Kiær, director of the Norwegian statistical institute at the time. Sampling is a statistical technique where only a selection of the population is carefully made. This method was further developed by Arthur Bowley in 1906 and Jerzy Neyman in 1934. Bowley suggested implementing a lottery mechanism and Neyman was a pioneer in confidence intervals and hypothesis testing. The actual data are collected by means of questionnaires that the sampled businesses have to complete. In the second half the 20th century, sample surveys have become the main method to collect data for official statistics. Sampling has proven to result in both precise and representative statistics of the actual population, although error sources need to be taken into account (using the Total Survey Error framework), and surveys are costly both for the National Statistical Institute (NSI) and businesses (i.e. response burden) (Groves, 1989; Bethlehem, 2009; Snijkers et al., 2013; Snijkers, 2016).

In addition to surveys, the use of administrative sources emerged in the 1970's. These registers are maintained by other organisations such as the Tax Office or municipalities in support of their own tasks. Since they are not constructed with the initial intention of making statistics, they are called secondary sources, as opposed to survey data which are a primary source. The process of fully integrating administrative sources is still ongoing; nowadays (Buiten et al., 2018). Statistics Netherlands (CBS) can access over 200 registers. More recently an enormous amount of organic data is being generated. These data, also known as big data, are often characterised by the four V's: volume, velocity, variety and veracity. Some examples of this are social media messages and traffic loop data (Daas et al., 2015). Unlike surveys, statistical institutes have no control over the data generating process of these secondary type of data (Snijkers & de Broe, 2018). Therefore, survey data are also called 'designed data', in contrast to secondary data, especially big data, that are called 'found data'.

Currently, the general policy applied by NSIs to collect data is a two-step approach. First, administrative sources (registers) are utilised as much as possible. Only in case the registers do not apply or additional information is needed, surveys are conducted. A main reason for this way of working is cost reduction, both for NSIs and businesses.

---

[1] Acknowledgement: The work presented in this paper is an internship project (Maastricht University), and was carried out in collaboration with Jeldrik Bakker and Ralph Meijers (Statistics Netherlands). The authors are grateful to Prof. Dr. J. de Vlieg (Eindhoven University of Technology) for his support in this project, and an arable farmer for providing the data. The authors like to thank José Gómez (Statistics Netherlands) for reviewing a draft of this paper.
Disclaimer: The views expressed in this paper are those of the authors and do not necessarily reflect the official policy of Statistics Netherlands.
[2] https://statswiki.unece.org/display/Collection/2019+Data+Collection+Workshop

1.2. Innovations in the business world

In the above described evolution, developments in the business world are not considered. However, businesses are not standing still, but are ever innovating (Srinivasan, 2017; Thomas and McSharry, 2015). Nowadays, more and more business processes are heavily data-driven,, depending on the industry. The data that are being generated in these business processes could be valuable for statistical ends as well. It is possible that the information that is asked for in surveys, is already encompassed in the data a business has themselves. Even more so, the data might actually contain more information than currently is being collected. And in addition, the data might become available at a much more rapid pace than when using surveys. If these data can be used to create the same as well as well as additional statistics, then the current statistics could be published more frequently and with much more precision, while at the same time lowering the administrative burden on the companies and reducing costs for the statistical institute (Snijkers & de Broe, 2018).

A fine example of such a development is precision agriculture. In contrast with traditional ways of producing crops, where actions are taken on field-level, in precision agriculture actions are specifically tailored for a certain section within a field. This level of detail requires good insight into the fields, usually provided by sensors (see e.g. De Vlieg, 2018; Viviano, 2017; Wolfert et al., 2017). The generated data are full of information that might also be useful for NSIs, and used instead of collecting data via surveys (Snijkers et al., 2013). In that case, response burden on the companies could be greatly reduced. Moreover, the sensor data could contain information that traditionally was too detailed and technical to ask for in questionnaires, meaning that additional statistics could be produced from these data. Furthermore, these data may be used in timely benchmark indicators in complement to statistics. In theory sensor data thus might be a valuable new source for official statistics. In this paper we'll look into the 'fitness for use' of this new data source (Biemer and Lyberg, 2003).

1.3. Explorative case study

In order to study this research goal, a case study has been carried out in which Statistic Netherlands (CBS) worked together with an innovative arable farmer and the Eindhoven University of Technology (TUe). The farmer has made a selection of his (sensor) data available to CBS with the purpose of exploring the data for overlap with relevant surveys and other promising applications.

Three questions lie at the heart of this case study. Is this data source fit for use with regard to:

1. Replacing (parts of) surveys by these data?
2. Developing new statistics from these new sources?
3. Providing useful, more detailed information back to the farmer, i.e. closing the data cycle?

This paper summarises the main findings of this case study (Punt, 2019), and will focus on the first research aim by investigating the overlap between the generated data by the farmer and the data needs of Statistics Netherlands. First of all, Section 2 discusses the data at the supply side, i.e. the data generating process at the farm, as well as the data infrastructure, i.e. the data flow and the systems in place used for data storage. Here also quality of these data are discussed. Section 3 discusses the demand side: data that is asked for in agricultural surveys targeted at arable farmers. In Section 4 the data selection made available by the arable farmer is analysed with respect to the data asked for in surveys, i.e. the overlap with the surveys. All the challenges that were encountered will be discussed in Section 5. Here, criteria laid out in a number of quality frameworks for official statistics have been applied. Finally, Section 6 concludes this paper by summarising all results and discussing a way forward.

2. Case study: innovative potato farmer

The arable farmer in our study produces potatoes, sugar beets, and corn, of which potatoes are the main crop. No animals are housed, nor is horticulture practised. Furthermore, the business area is more than 500 hectares spread over more than 100 fields, located in both the Netherlands and Flanders (Dutch-speaking part of Belgium). On the farm many sensors are used to monitor and optimize the farming processes.

2.1. Data generating process: Precision agriculture cycle

2.1.1. Winter

First of all, in the winter field borders are drawn in. The goal is to determine the location and size of each field so that the employees know where the fields are located. Variables such as soil moisture and nutrients within

one field can vary greatly. To identify these differences the farmer does a soil scan using sensors that measure electrical conductivity, which results in a better understanding of the composition of the soil.

Moreover, tractor lanes are determined; these are the paths on which farm machines drive. Since this concerns heavy machinery, the soil will be affected, reducing the potential yield in these lanes. It is therefore important that these paths are chosen very carefully. In doing so, a number of constraints need to be considered: 1) the farmer must be able to operate with various tools that can vary in width; 2) sharp turns should preferably be avoided, because this will damage the surrounding plants when the tractor drives back and forth in order to make the turn, and 3) the farmer wants to minimize underlap and overlap of land within the field, to save costs and time. The most optimal lines are calculated using GPS systems.

2.1.2. Spring

In spring, the seed potatoes are planted. As discussed earlier, the yield potential within one field is not uniformly distributed; certain sections in the field may contain more or less nutrients, are more or less humid, and get more or less sunlight. Based on these parameters, the farmer varies the planting distance within a field.

Examples of section with a different yield potential are tractor lanes and shaded areas:

- Tractor lanes. Since heavy machines would destroy all plants in the lane, they are not sown. The total area of these paths amounts to 5.5% of the total cultivation area at this farm. The consequence of leaving these lanes open is two-sided: on the one hand, dropping the yield potential, because there are simply no crops planted in these lanes. On the other hand, it increases the yield potential of the directly adjacent rows to the tractor lanes. This is due to the fact that the extra free space makes it so that more nutrients per plant are present and that there is more sunlight incoming from the side onto these plants. To take advantage of this increased yield potential in the rows right next to a tractor lane, the planting distance is reduced by 10%.
- Shade areas. As the plants beside the tractor lanes get more light, there are also places, often at the edges of a field, which get less sunlight. Less light will result in a lower yield potential. The shade is often due to trees and shrubs next to the fields. By increasing the planting distance in the shadow, there will be more nutrients and sunlight per plant left over. This increases the yield per plant and supresses costs on seed potatoes.

2.1.3. Summer

In the summer, the potato plants are supplemented with water, nutrients and pesticides, if necessary. The fields worked by the farmer are characterized by sandy soils that are very sensitive to drought. The water deficit is supplemented by irrigation, so the plants can grow optimally. For this an irrigation management system is used. The system contains a number of sensors; one measures the soil humidity of a number of soil layers via electricity. The deepest layer is 60 cm under the ground, the other layers are incrementally 10 cm higher. The information that is gathered is then linked to a local weather station. On the basis of the measurements of the humidity sensors and the weather station, the system can decide whether water should be added or not and at what time. However, since the sensors only measures one spot on the field, the information might be flawed for big fields. The location where the sensors are placed, is determined by the earlier soil scans. Sections in the field with a higher conductivity can hold more water than zones with low conductivity. In this way the farmer can monitor critical sections in the field.

In addition, there is also data gathered on the condition of the plants. The collection can be divided into two categories: sensing and sampling.

- Sensing encompasses the data collection with the use of sensors. This allows, for example, on the basis of the reflected light of the plant different kinds of vegetation indexes to be derived. Other indices that are measures are the IBI and IMRI. The first measures the bio mass; the amount of green surface compared to the total. The second is a nitrogen index. This index can be checked whether the plants are getting enough nutrients; when this is not the case, the machine adds fertilizer on the spot.
- In addition to the use of sensors, the farmer also makes use of manual sampling. Around eight different samples are done per field in a given cycle. The sample consist of three plants that are uprooted and dissected further into different parts:
  - Haulm (the stems and leaves): the weight, the stem length the number of levels,
  - Tubers (potatoes): the size, the weight, the number of tubers, and
  - The roots: the length, the weight.

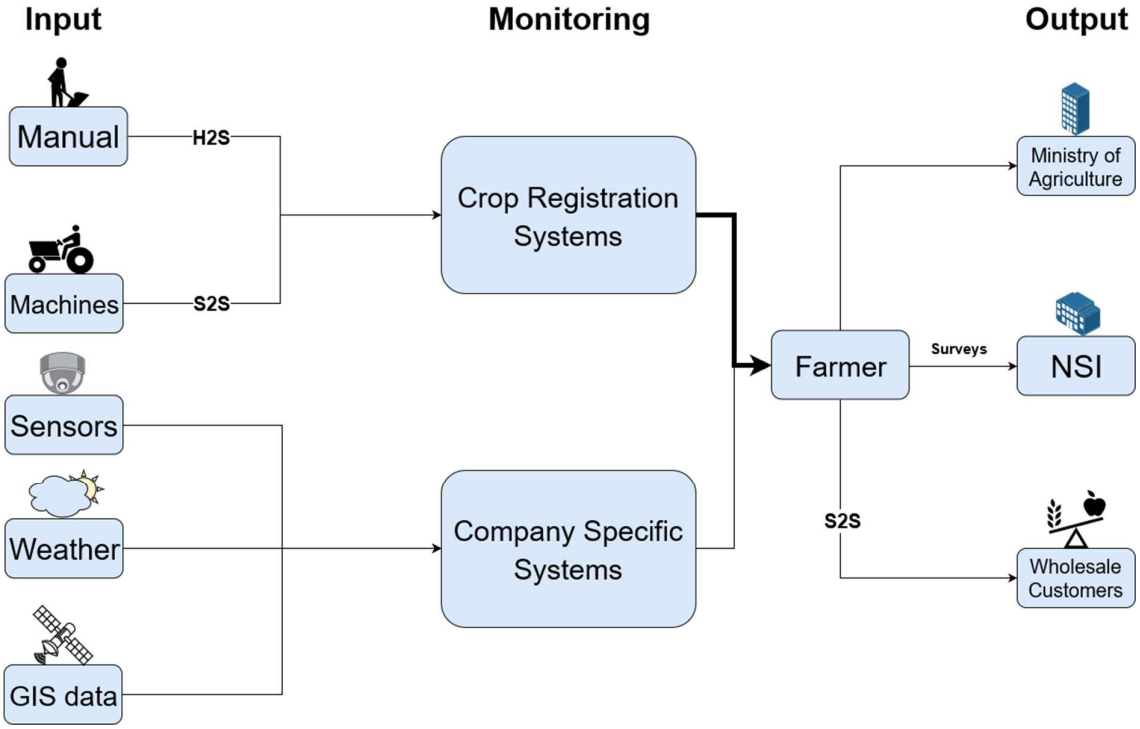
2.1.4. Autumn

Potatoes are harvested in the fall. Then, the farmer monitors the yield of the fields closely using a custom-made harvester. The machine combines three different sources of information: the weight of the potatoes, GPS information to determine the exact location within a field, and the speed of the tractor and treadmill on which the potatoes move towards the storage bunker. The last source corrects for the delay between the GPS information and the weighing sensor. By combining these three sources, it is possible to create insightful maps of fields, in which good and bad-performing sections can be identified. These maps are useful to evaluate different approaches in the context of precision agriculture, to see what works and what does not.

2.2. The farmer's data infrastructure

The farmer's data are stored in two different data platforms: a platform for precision farming data (called DIPPA: Data Integration Platform for Precision Farming; Enkhtaivan, 2018), and the so called 'crop registration system'. The DIPPA platform is specifically tailored to the farmer's needs and houses the sensor-generated data. Figure 1 depicts a simplified overview of these platforms, also showing data flows. This sub-section discusses input, monitoring and output of this data infrastructure.

Figure 1. The farmer's data infrastructure



2.2.1. Input: open source and private data

The input side of the data infrastructure is split into two categories: farmer specific data, and open data. First of all, we have the data that are generated by the farmer, both by sensors and by manual measurements (like the sampling of crops in summer). As we have seen above our farmer uses of a wide array of sensor-generated and manually collected data. But apart from these data, he also has data on e.g. use of manure (in the manure administration), financial data, and data on his staff (Hazim & van Melis, 2018).

Secondly, the farmer uses open data that are freely available to everyone. Van Dijk and Kempenaar (2016) describe various open data sources that can be used within precision farming. These data include satellite data, meteorological data, soil data:

- Satellite data. The Dutch government provides satellite data via the national satellite data portal, free of charge. The satellite images have many conceivable applications within the agricultural sector. Various indices can be calculated from the images, examples of which are: biomass, crop development and nitrogen content in plants. Based on this information, farm activities such as fertilization, watering and crop spraying

may be planned. However, application within the sector is still very limited due to the lack of delivery of high-quality images throughout the season and the knowledge to transform these data into practical actions.

- Meteorological data. The weather is one of the most impactful factors in the agricultural sector. Not only is it responsible for a large part of the crop yield, it is also leading in the daily actions of the farmer. For example, there are desired weather conditions for certain actions such as spraying and fertilization. In the Netherlands, weather data are provided by the Royal Netherlands Meteorological Institute (KNMI). KNMI has a network of measurement stations that observe various weather variables, such as temperature and solar energy. These data, often in time series based on the hourly, daily, monthly and annual levels, have been made available free of charge by the institute.
- Soil data. Next to the weather, the soil is also a crucial factor to take into account when it comes to crop cultivation. Different soil types are characterized by a substantial difference in the retention of nutrients and water. As a result, some types of soil have a higher yield potential than others. Wageningen University & Research has made a GIS file freely available providing a global overview of the soil of the Netherlands.

2.2.2. Monitoring: data storage platforms

After the data is generated or imported, the data are stored in two data platforms. One is widely used in the agricultural industry, the so-called 'crop registration system'. The other system was specifically designed for this company in collaboration with the University of Technology Eindhoven (TUe): the DIPPA system (Data Integration Platform for Precision Farming; Enkhtaivan, 2018) and will be referred to as 'company specific systems'.

First of all, let's look at the crop registration system. In the Netherlands, the number of generally used crop registration systems is limited to two: Dacom and AgroVision (Hazim and Van Melis, 2018). An educated guess is that 30-40% of all farmers use such a system, mainly the larger ones. These systems have the following characteristics:

- 'Dairy' of the parcels of the farm
- Administration about pesticides, manure, etc.
- Export functions to report to third parties
- Industry harmonised data formats (EDIteelt)

In a crop registration system it is possible to keep digital records of all farm activities. To start, a farmer enters the coordinates of his fields into the system. This can be done by uploading a Shapefile or by drawing the fields on a map using a selection tool. Afterwards, the farmer can assign crops to the fields with additional information such as crop strain and purpose of crop e.g. human consumption, fodder or bio-energy. Once the farm has been set-up, records can be added to the fields during the cultivation season. These activities usually fall into a category associated with standard farm practices, and include among others: planting, irrigating, fertilizing, crop protection and harvesting. These activities can be entered manually, i.e. human to system (H2S) data communication, or they can be imported directly from the farmer's machines, i.e. system to system (S2S) data communication. Accompanying the records are a timestamp, the field on which the activity takes place, and a number of (optional) activity-specific variables. Systematically entering all activities is a critical issue with regard to crop registration systems. If this is not done, the data in the system are incomplete.

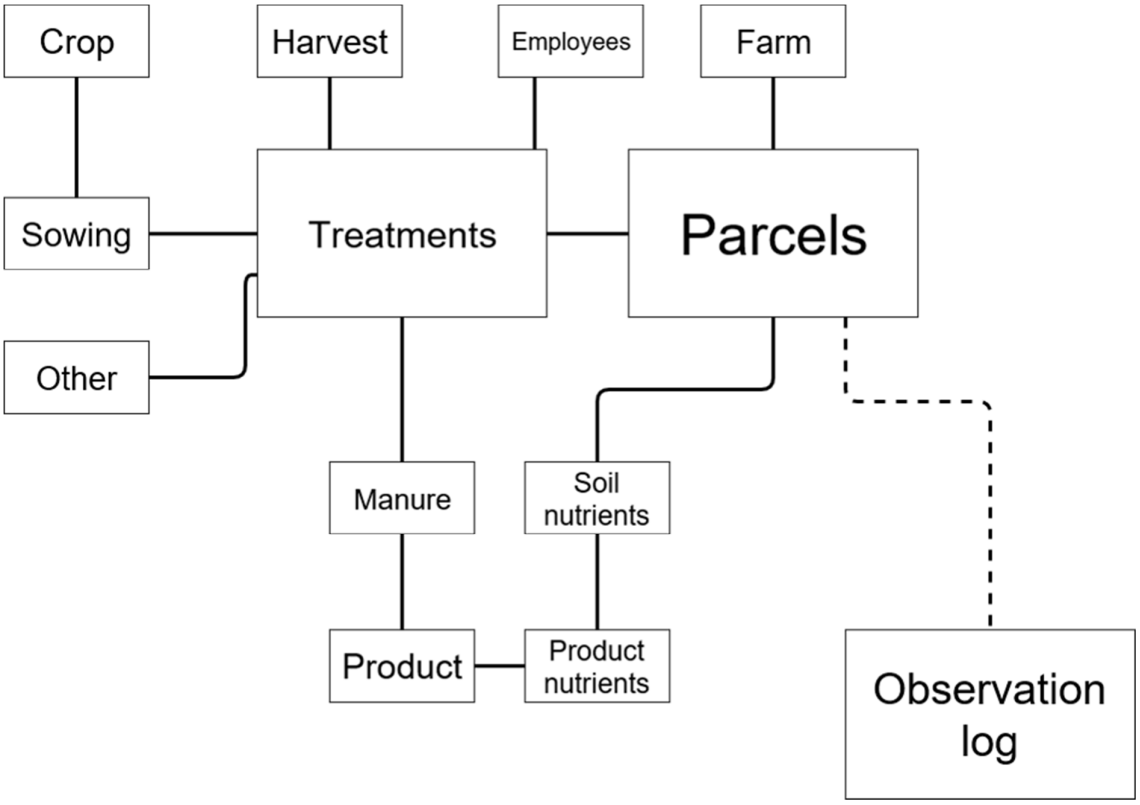
Another interesting feature of these crop registration systems is the export functionality. Data can be exported to multiple data formats such as Excel and PDF, but more importantly data can be transformed into a hamronised data format: Editeelt, making data communication with other systems easier. EDIteelt is e.g. used by wholesale buyers of the crops to check the quality products they buy.

Apart from the crop registration system, our farmer uses a data platform that was specifically designed for managing precision farming data, for which crop registration systems are insufficient. This platform, called DIPPA, was designed in close cooperation with the Eindhoven University of Technology (Enkhtaivan, 2018). The purpose of this collaboration is to develop a system that stores precision farming data and makes it easier to use for domain experts and data analysts. The propose of DIPPA is to satisfy the specific data management needs of our farmer, but with the aim of making it applicable to more farmer working with precision farming technology. This is the 'company specific system' in Figure 1, and is characterised by:

- Storing and managing 'new' data such as sensor data and other precision agricultural data
- Small scale
- Not standardised

The actual data is stored in a relational database. A selection has been made available to Statistics Netherlands as an SQL-dump. This can be seen as a recipe for building the database in the same state as when it was exported. After building the database, it turned out that the database consists of two separate structures. The first consists mainly of administrative information and the second houses various measurements and can be seen as an observation log. Figure 2 shows a simplified overview of this system. Central to the relational database are the tables *Parcels* (fields) and *Treatments* (farm activities). Every activity falls under a treatment category and is linked to a field. The data that is gathered in the context of precision agriculture is in the table *Observation log*; it consists of 26,000 individual observations spanning a 6-month period, these observations can again be linked to a field.

Figure 2. Relational structure of the company specific system, containing precision farming data



2.2.3. Output: agricultural surveys

These two systems contain the data that are needed for reports to be sent to other parties, like reports to the Ministry of Agriculture, to wholesale customers, the Tax office, the bank, but also to the Netherlands Enterprise Agency (RVO) and CBS. These surveys are discussed in more detail in section 3.

2.3. Exploration of the farmer's sensor data

When we started exploring the farmer's sensor data in DIPPA (eventually resulting in Figure 3), we realised that the data could not be analysed straight away. Before the data could be analysed two steps had to be taken. First of all, a complete and accurate description of the data was missing. We only had a database with variable labels at the top of each column and numbers in the cells.

The next step was data cleaning: the data suffered from e.g. missing values, measurement errors (e.g. incorrect values), and duplicate records (Enkhtaivan, 2018). Some errors could be corrected by checking the data, but in order to get a complete understanding of the content of the data the farmer's expertise about the metadata and the data generation process was indispensable.

One framework with which the quality of a secondary information sources can be determined is the checklist developed by Daas et al. (2009). The checklist consists of three hyperdimensions: the first deals with the source of the data; the second deals with the meta information i.e. the data about the data; the third hyperdimension discusses quality aspects of the data themselves. We applied these hyperdimensions to these data. Issues related to the source of the data would include e.g. ownership of the data, privacy issues, and costs for farmers to deliver the data. Here, we will focus on the second and third hyperdimension: the metadata and the data.

2.3.1. Assessing the metadata

The checklist on metadata deals with the clarity of the population, variables and time dimension. When applied to the DIPPA data, we come to the following assessments:

- Population unit. The population unit are the individual fields. These fields are then linked to one farm or company. In the current state of the database, this only concerns one farmer, but it is possible to expand this with other agricultural companies.
- Clarity of data definitions. The clarity of data definitions leaves much to be desired: documentation about the variables in the system (like a code book) is missing. There are only a few variables that are clearly defined. This mainly concerns qualitative variables such as *type of crop*. For many the others variables it was unclear what they are about: labels on top of the columns are acronyms, and hard to interpreter, or these labels were missing at all (having columns with no names at all) . Furthermore, for most quantitative variables the units of measurement (like kg, ha, kg/ha) is not specified. This means that this information has to be derived from context, domain knowledge or contact with the farmer, which requires a lot of extra effort.
- Time dimension. The time dimension of all measurements and variables are well documented, they are all shown in one date format, which makes it easy to use for analysis.
- Unique keys. A key enables individual records within a dataset to be linked to different tables or other data sources. In the case of the DIPPA database there are two keys that are essential for the interpretation of the data: a company identifier, and identifiers for the specific fields within the company. Statistics Netherlands is particularly interested in what happens at the company level, in connection to data from other sources. The keys used in DIPPA are specific to this database and cannot be used to link the data to other data sources, like e.g. a cadastre number.
- Data treatment by database administrator. Another dimension that requires attention are the actions that may or may not be performed by the database administrator. This allows data to be modified. It is therefore essential for an NSI to be aware of this and to know what happened to the data before it arrives (Daas et al., 2009): all changes in the data need to logged and described. Contact with TU/e showed that the original, raw data, as provided by the farmer's data manager in Excel, were cleaned before it was handed over to CBS (Enkhtaivan, 2018). These data quality issues are related to the data hyperdimention, and are discussed in the next subsection.

2.3.2. Assessing the data quality

The original data that TU/e received contained a number of problems and errors, as described by Enkhtaivan (2018). These include:

- Missing values and missing records. Depending on the period of the growing season, the data are entered manually. Some samplings and actions are recorded, but not systematically for all fields. Due to this, the database contains a lot of missings (null or zero values), either item missings, but also missings of records. Furthermore, the fact that cells either contain a null or zero or are empty, leaves us with the question whether data is actually missing, or if these zeros are correct values.
- Multi data formats. Data are written in inconsistent ways. E.g. dates are written in various ways, like 28 May: 28/mei, 28-05-16, or 28/5/2016. Another challenge is the use of Excel, which guesses the format when a dataset is opened. This can hide how data are represented in the original data files.
- Multiple data representations. Because of inconsistent spelling or spelling errors, data are represented in various ways, like for fertiliser products: 'zeugen mest' or 'zeugenmest'. Because of these spelling errors or slight differences in variable labels, the effect may also be that data are stored in different variables, while originally they are one and the same variable. Also some data is measured in ranges, like 50/60, 50 60, 50-60. It would be better to have two variables: one with the lower bound, and one for the upper bound.
- Duplicate records. In the original database records and columns were duplicated. This was done, e.g., to make comparisons on the outcomes of the previous/current year or weeks. One column was repeated as much as 60 times.

- Redundant data. In addition, the database contained redundant information. This applies e.g. to ratios, sums or differences of two variables that is also stored unnecessarily.

The result is that the data cannot be used straight away, and a lot of data inspection and cleaning was needed. As Enkhtaivan (2018) points out, the risk here is that this introduces errors because of the missing data documentation, and the required domain knowledge.

In this data cleaning process, outliers were not detected, and wrong values of data were not modified. When inspecting the data we found e.g. unrealistic values like freezing temperatures in August, or clearly wrong data like field coordinates in Kazakhstan.

To conclude this section, and leaving aside the question whether the data are relevant as a source for official statistics (which we will discuss in the next sections), there are still many challenges with regard to its data quality. That is also why the Eindhoven University of Technology (TUe) decided to improve the DIPPA system and develop DIPPA 2.0.

3. Data asked for in agricultural surveys

So far we have discussed the data available at the farmer's side, let's say the supply side. Now let's look into the demand side. As said in subsection 2.2.3, the data in the two systems are the basis for reports to third parties, like CBS. The reports to the Netherlands Enterprise Agency (RVO) and CBS include the following four surveys, specific for arable farming:

1. The combined survey (conducted by RVO, and sent out in May; data are forwarded to CBS)
2. Harvest estimate survey (CBS, sent out by the end of October)
3. Crop protection survey (CBS, sent in mid-December)
4. Grassland usage survey (CBS, sent in mid-December)

The Combined survey deals with all kinds of topics such as employees, farm area, manure, farm animals, horticulture, floriculture, mushrooms and subsidies. Whereas, the harvest estimate survey only deals with the (expected) harvest of a number of crops (like wheat, barley, rye, corn, and potatoes), and the sowing of winter crops (e.g. winter wheat, barley, and rye). The crop protection survey asks the farmer of their use of chemical, mechanical and biological measures taken in order to protect their crops. Finally, the grassland usage survey inquiries about the harvesting of grass on the farm: how much, how is it conserved and for what purpose?

All surveys, except grassland usage, have become mandatory since 2017; a drastic measure to increase the response rates which were around 45%. After 2017 response rates have increased to 70%, but the administrative burden on the businesses remain (Snijkers et al., 2018). For the combined survey, larger farms such as our farm are integrally observed every year. The harvest estimate survey only takes these larger farm companies into account, which are included in this survey at most once in every three years. The crop protection is being sent out every four years.

4. Data overlap between the farmer's data and surveys

We have seen that our farmer uses two different systems. Therefore in the analysis with respect to the overlap with the surveys, it was decided to include both systems. First, with respect to the company specific systems (Punt, 2019): the analysis of the sensor data showed that a lot of the questions asked in the surveys can either be directly answered or indirectly deduced from the information that is present in the database. In the case of the harvest estimation survey, the production part of the survey could be retrieved directly from the database of the case company. The more complex, combined survey, can be partly filled out. The part that can be filled out, deals with surface area and fertilization. The section that deals with employees can potentially be filled out using the database, however key details about the relationships to the farm owner and contract specification are still missing and is unlikely to be added. Figure 3 shows the subsections of all surveys and how well they overlap with the company specific system. It is colour coded where green indicates a strong overlap, yellow meaning partial overlap and red indicating no overlap at all.

Figure 3. Overlap between agricultural surveys and farm sensor data

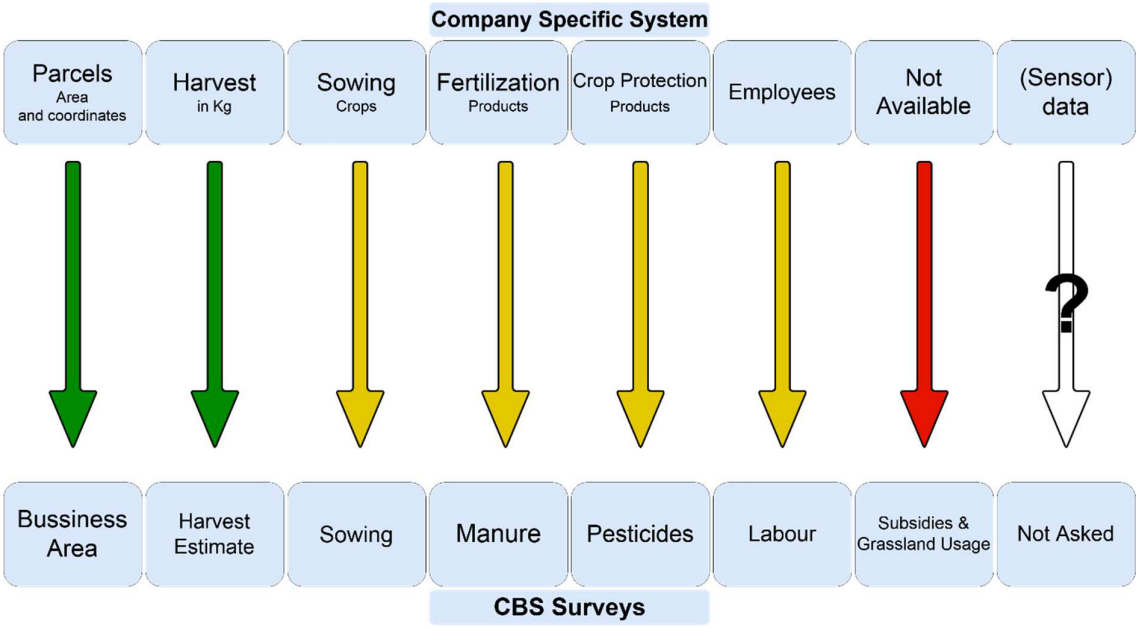
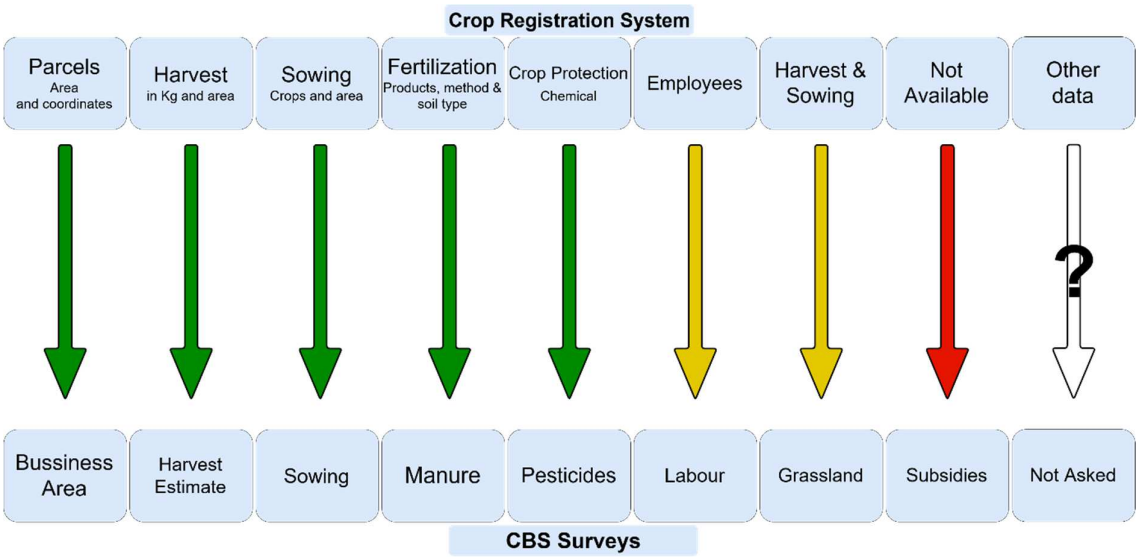


Figure 4. Overlap between agricultural surveys and crop registration systems.



Shifting the focus to the crop registration systems used by our farmer, a similar figure has been constructed, depicted in Figure 4. Major differences to the company specific system is that now also all variables related to pesticide usage are available, as well as the sowing of (winter) crops. Yet, both systems are not useful to gather information about subsidies. These questions often deal with possible event in the future and whether the company wants to utilise certain subsidies if applicable. The crop registration system houses the more traditional information about a farm and has therefore, unsurprisingly, more in common with the surveys sent out by Statistics Netherlands, whereas, the company specific system has very detailed information about the farm that currently only a handful of farmers measure.

5. Data challenges in the context of official statistics: fitness for use

So far we have discussed the data of one farmer in relation to the data needs of an NSI like Statistics Netherlands. We have concluded that there is some overlap, but also that the data quality needs to be improved, both at the metadata and the data level. However, going back to our first research question (Is this data source fit for use with regard to replacing (parts of) surveys by these data?), more issues need to considered in order to assess its fitness for use (Biemer and Lyberg, 2003). To identify these issues a number of quality frameworks have been applied. Since this data source uses sensors as measurement instrument, and combines characteristics of both primary and secondary data sources, the applied quality frameworks are associated with:

- the use of sensors as measurement instruments and the data generation process (see e.g. Rai et al., 2017),
- primary data sources like business survey data (Snijkers, 2016; Haraldsen, 2013),
- the mode of data collection used by NSIs, like System-to-System (S2S) data communication (see e.g. Buiten et al., 2018), and
- the use of secondary data sources, like registers (Daas et al., 2009).
- the data-user perspective, which in our case is the NSI perspective (Eurostat, 2017; Snijkers and De Broe, 2018; De Broe, Snijkers & Schouten, 2019).

Based on the results from the discussed case study, and having these frameworks in mind, we can define a number of data challenges that are relevant with regard to the use of sensor data collected directly from farmers with the aim of using these data as a source for official statistics. In this context, in the introduction we referred to the concept of 'fitness for use' (Biemer and Lyberg, 2003). Without trying to be complete, this list includes the following issues:

- Privacy, security, data-ownership, and data sharing

A first issue is: do we actually get access to the data? Some of the data of a farmer is very sensitive; agricultural businesses are very strictly regulated with hefty fines for exceeding the limits on e.g. pesticide and manure usage. While Statistics Netherlands will in no way ever share microdata with regulatory institutions, the farmer may remain suspicious. Trust is the magic word. Furthermore, a debate is going on with regard to data ownership: whether the data that are generated at the farm belongs to the farmer or the manufacturer of the sensor equipment (see e.g. CEMA, 2019). So, even if farmers trust Statistics Netherlands with data about their farms, they may not be able to share it because they have no control over the data. (Daas et al., 2009; see Olhede and Wolfe, 2018, for a discussion.)

- Metadata

Once the data can be accessed, good and complete metadata is vital for the interpretation of the data, as we concluded from our case study. Without it the transformation from data to information is not possible. It is therefore crucial that the data are described in a detailed and complete manner. (Daas et al., 2009.)

- Data quality issues

Once we know what the data are about, data inspection is needed to check for data quality. Data quality is a very broad concept and includes e.g. measurement errors, unit and item non-response, and representativeness (Snijkers, 2016). In terms of sensor data quality one can think of repeatability (the dispersion between consecutive measurement obtained from a given sensor), reproducibility of the measurements with different sensors, stability of the sensor, and data drift (the capability of the sensor to maintain its performance characteristics over a sufficiently long time) and limit of detection (the lowest concentration of a substance that can be significantly differentiated from zero concentration) (Rai et al., 2017). To check for errors in the data, data cleaning is an important part of the data process, as we have seen in our case. These issues are related to what in survey methodology is called bias and variance: systematic errors result in bias, random errors result in variance (Bethlehem, 2009).

- Conceptualisation

When using these data in the context of official statistics, a 'fit for use' issue that also needs to be considered is conceptualisation: we might have good data, but are the data actually correlated to the concepts for which statistics are to be produced? This relates to validity of the data (Snijkers, 2016). This touches on the requirement that the concepts of the data in the systems have to be (closely) related to the concepts for which statistics need to be produced, and in case of a survey are asked about in the questionnaire (Snijkers, 2016). This relates to the comparability of secondary data with NSI definitions (Daas et al., 2009). In our study we found

that the most promising data source was the crop registration system, since the concepts overlap the most with the questionnaires.

- Unit issues

Another 'fitness for use' issue is the unit. In case of collecting data from businesses, it is important to check whether the data are related to the correct, pre-defined units: do we get data about the correct unit (Snijkers, 2016). In case of our farmer we found that some of his fields are located in Belgium. When producing statistics about the Netherlands, data about these fields need to be excluded.

- System-to-system data collection and the use of sampling

In the 20st century surveys using sampling have proven to be to a cost-efficient method to get accurate statistics (Bethlehem, 2009). The data are collected by use of a questionnaire (Snijkers et al., 2013). With the possibility of connecting to a registration system, and collecting the data using System-2-System (S2S) data communication, the question arises whether it is necessary or even desirable to use sampling. Once a S2S has been set-up, it is more efficient to use this system all the time, instead of using only now and then based on whether a farm has been sampled. If the ubiquity of these new data is sufficiently large, the data collection process may turn into integral observation. (Buiten et at., 2018.)

- Ubiquity and market penetration, standardisation

In our case study we used data from just one farmer. Obviously this is not enough to produce statistics. When talking about these data as input for statistics, we need data from many farmers. This means that the data about the relevant topics are commonplace in the industry. More specifically considering the two data storage systems identified in the business case, it means that enough farming companies use similar systems, i.e. standardisation of systems (Buiten et al., 2018). When market penetration is sufficient enough, it becomes interesting for NSIs. Because of this issue, crop registration systems are more interesting for NSIs than precision farming systems.

- Data harmonisation

Suppose that many farmers use precision farming, and ubiquity is guaranteed, there probably will be a diversity in sensors and machines used. Each farmer will use its own equipment and machines produced by multiple manufactures. Consequently, the data that are being generated, may be very likely be defined differently. It is therefore important with the goal of statistics in mind that the data are harmonised. Insights from the business case indicate that the crop registration systems have developed their own data format and standard that facilitate easy data exchange. This speaks in favour use the systems becoming a potential data source for statistics in the future. In the Netherlands a harmonised data system called 'EDI-teelt' has been developed; on a European level an international standard called 'e-crop' is being developed. (Buiten et al., 2018.)

- Stability of (meta)data delivery

Once all the above has been dealt with, we also need to include the time dimension. An important aspect of official statistics is time series: statistics are especially useful when observed over time. Therefore, it is of the utmost importance that the data delivery is constant and consistent over time, so statistics may be produced for more than just a single time. This needs to be guaranteed. (Daas et al., 2009.)

It is obvious that having these quality issues in mind, we still have a long way to go.

6. Discussion and summary

The first research aim was to investigate whether sensor data could be used to (partly) substitute surveys. This was examined in the context of precision agriculture in collaboration with one innovative arable farmer that uses many sensors throughout the year. In the winter, he scans his fields for soil characteristics such as humidity and nutrients. In the spring, new potatoes are variably planted in such a way that it corresponds to the yield potential of the fields. In the summer, crop growth is carefully monitored by taking measurements, both with the aid of sensors and manually. When the crops are harvested in the fall, the potatoes are weighed on the spot, giving the farmer detailed insights into high performing and low performing sections in the fields.

What initially was a research project on the usefulness of sensor data as a source for statistics, quickly shifted to a broader case study on how the data infrastructure in the agricultural industry was situated. The sensor data may indeed be very interesting, but in order to be able to gather enough relevant data in a harmonised way, standardised systems have to be in place with harmonised data that an NSI can tap into.

It was found that the generated data during the season is stored in two different platforms: a company specific system and a so-called crop registration system. The first system is tailored to the needs of our farmer and houses most of the precision agricultural data. An initial exploratory analysis of this platform shows that parts of questionnaires can be completed using this data, while other parts still have to be completed manually. Questions about the area of the farm and the harvest estimate can be filled in to a large extent, but the sections on planting, fertilization and the number of employees can only be partially answered. Financial information requires a completely different data source. It was concluded that parts of the sensor data are associated with some concepts in the agricultural questionnaires.

Two problems became apparent at the start of the analysis. First, an accurate description of the farmer's sensor data (i.e. metadata) was missing. When interpreting the data, a combination of context, domain knowledge and contact with the farmer has proven to be essential. The second problem has to do with the quality of the data and the database. There are several challenges: different data formats are used, missing values and inconsistent naming. We concluded that this company-specific system is very interesting to gain insight into what farmers are measuring about their business processes, but currently it lacks the quality to produce high quality statistics.

In addition we looked into other quality issues that need to be taken into account when assessing the fitness for use of this data source in the context of official statistics. This includes getting access to the data, unbiquity and market penetration, standardisation and data harmonisation. Also here, we concluded that there is still a long to go: in theory the idea we started with looks promising, but in practice there are still many challenges. This can of course change over time as the industry innovates.

But even though, we feel that now is the time to start examining and discussing the challenges and data communication options with regard to this new data source in order to be ready for the future. Also more studies like the one described here are needed, to get a better picture of the data challenges and their fitness for use. E.g. looking into sensors that specifically are used for crop growth and crop protection. We may also look into other sectors of the farming industry. In horticulture and dairy farming, the use of sensor technology may be more ubiquitous than for arable farming. This may also be the case for other sectors in the economy, like the transport sector.

In addition to the company specific data system, we also investigated crop registration systems. These system house the more traditional information about farm activities and are used by many farms in the Netherlands. In the system it is possible to import, save, interpret and share data about the fields. These systems can be entered manually as well digitally via farm machines and contain the most overlap with the questionnaires that farmers can receive. These include sections about field and crop registration, pesticide use, fertilization and harvesting. The number of mainly used systems in the Netherlands is also limited, meaning that there is some standardisation; also some data harmonisation by use of EDIteelt is supported by these systems. For the short term, we therefore feel that these crop registrations may be a better source to investigate than farmer specific data systems.

Moving forward, the crop registration systems could be investigated in-depth as a source for official statistics. The data could be imputed in the questionnaires farmers receive, thus integrating the data that are already available and the questionnaire into one data collection system. Since farmers are using these systems and some level of harmonisation has been done by the industry already, integration with the questionnaires seems feasible. While the upsides are clear there are also some risks that exists with this:

1. Can the farmers use this new integrated method of data collection?
2. The quality of the data?
3. Will the farmers use it?

The first risk touches on both the ubiquity of the crop registration systems and the technical skills of the farmers: is the pool of farmers using these systems large enough to make this work? Moreover, can the farmer navigate their system to make an export and the subsequent tool created by Statistics Netherlands to automatically fill in the questionnaire? The second risk deals with the quality of the data in crop registration systems. This approach assumes that the data in this system are of high quality: that the data are correct, and no data is missing. This assumption has to be studied. The third risk deals with trust. Since some of the information is rather sensitive, e.g. pesticide and manure data, the farmers might not be willing to share their export from the registration system. Even though Statistics Netherlands does not share micro data with any institution whatsoever, the farmer might still be suspicious. Our experience so far is that is a high hurdle to take.

References

Bethlehem, J. (2009), *Applied survey methods: a statistical perspective*. Wiley, Hoboken.

Biemer, P., and L. Lyberg (2003), *Introduction to Survey Quality*. Wiley, Hoboken.

Buiten, G., G. Snijkers, P. Saraiva, J. Erikson, A.-G. Erikson, and A. Born (2018), Business data collection: Toward Electronic Data Interchange. Experiences in Portugal, Canada, Sweden, and the Netherlands with EDI. *Journal of Official Statistics*, Vol. 34, No. 2 (ICES-5 special issue).

CEMA (2019), EU code of conduct on agricultural data sharing by contractual agreement. European Agricultural Machinery Association, Brussels, Belgium. (https://www.cema-agri.org/publications/19-brochures-publications/37-eu-code-of-conduct-on-agricultural-data-sharing)

Daas, P., M. Puts, B. Buelens, and P. van den Hurk (2015), Big Data as a Source for Official Statistics. Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 249–262, http://dx.doi.org/10.1515/JOS-2015-0016.

Daas, P., S. Ossen, R. Vis-Visschers, and J. Arends-Tóth (2009), Checklist for the quality evaluation of administrative data sources. Discussion paper 09042, Statistics Netherlands, Heerlen/The Hague, The Netherlands.

De Broe, S., G. Snijkers, and B. Schouten (2019), Sensor data at the heart of innovation in official statistics. Paper presented at the 62nd ISI World Statistics Congress, Kuala Lumpur, Malaysia.

De Vlieg, J. (2018), 'A huge push in technology is coming' (in Dutch: 'Er is een enorme push in technologie in aantocht'). Boerderij, 03: 19 (februari 2018).

Enkhtaivan, B. (2018), DIPPA – Data integration Platform for Precision Farming. SAI Technical report 2018/104, Eindhoven University of technology, Eindhoven, The Netherlands.

Eurostat (2017), European Statistics Code of Practice. European Union, Luxembourg.

Groves, R.M. (1989), *Survey errors and survey costs*. Wiley, New York.

Haraldsen, G. (2013), Quality issues in business surveys. In: Snijkers, G., G. Haraldsen, J. Jones, and D. Willimack, *Designing and Conducting Business Surveys*, pp. 83-125. Wiley, Hoboken.

Hazim, J., and R. van Melis (2018), Crop Registration Systems 2.0; Business information data flows (in Dutch: Bedrijfsregistratie 2.0: Datastromen naar bedrijfsinformatie). Internship report GiM-2017-009814 / 18200146, HAS (higher vocational education), 's Hertogenbosch, The Netherlands.

Olhede S.C., and P.J. Wolfe (2018), The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A* **376**: 20170364. http://dx.doi.org/10.1098/rsta.2017.0364

Punt, T., 2019, (Sensor) data in precision farming; exploring one potential data source for official statistics (in Dutch: (Sensor)data in de precisielandbouw; Verkenning van één potentiéle bron voor statistiek). Internship Report, Statistics Netherlands, Heerlen.

Rai, A., P. Kumar, F. Pilla, A. Skouloudis, S. Di Sabatino, C. Ratti, A. Yasar, and D. Rickerby (2017), End-user Perspective of Low-cost Sensors for Outdoor Air Pollution Monitoring. *Science of The Total Environment*, 607-608: 691-705.

Snijkers, G. (2016), Achieving Quality in Organizational Surveys: An Holistic Approach. In: *Methodische Probleme in der empirischen Organisationsforschung*, Liebig, S., and W. Matiaske (eds.), pp. 33-59. Springer, Wiesbaden.

Snijkers, G., M. Geurden-Slis, G. Goris, J. Burger, and L. van den Hombergh (2018), The effect of response measures in business surveys. Paper presented at the 2018 UNECE Workshop on Statistical data Collection "Resourceful data Acquisition", 10-12 October 2018, Geneva, Switzerland.

Snijkers, G., and S. De Broe (2018), Smart business statistics: how to integrate technology and official statistics. Paper presented at the 2018 European Conference on Quality in Official Statistics, 26-29 June 2018, Krakow, Poland.

Snijkers, G., G. Haraldsen, J. Jones, and D.K. Willimack (2013), *Designing and Conducting Business Surveys*. Wiley, Hoboken.

Srinivasan, V. (2017), *The Intelligent Enterprise in the Era of Big Data*. Hoboken, NJ: Wiley.

Thomas, R., and P. McSharry (2015), *Big Data Revolution: What Farmers, Doctors, and Insurance Agents teach us about discovering Big Data Patterns*. Wiley, Chichester, West Sussex, UK.

Van Dijk, C., and C. Kempenaar (2016), Open data for precision farming in the Netherlands (in Dutch: Open data voor precisielandbouw in Nederland). Wageningen Univeristy & Research, report 662, Wageningen, Netherlands.

Wolfert, S., L. Ge, C. Verbouw, M.-J. Bogaardt (2017), Big Data in Smart Farming – A review. *Agricultural Systems* 153: 69-80.

Viviano, F. (2017), This tiny country feeds the world. National Geographic, September 2017. (www.nationalgeographic.com/magazine/2017/09/holland-agriculture-sustainable-farming/).