2023

A Big Data Smart Agricultural System: Recommending Optimum Fertilisers For Crops

Vuong Ngo
Technological University Dublin, Ireland, vuong.ngo@tudublin.ie

Thuy-Van T. Duong
Ton Duc Thang University, Ho Chi Minh City, Vietnam

Nguyen Nguyen
University of Information Technology, Ho Chi Minh City, Vietnam

See next page for additional authors

Funder: Open Access funding provided by the IReL Consortium.

Int. j. inf. tecnol. (January 2023) 15(1):249–265
https://doi.org/10.1007/s41870-022-01150-1

ORIGINAL RESEARCH

A big data smart agricultural system: recommending optimum fertilisers for crops

Vuong M. Ngo[1,2] · **Thuy-Van T. Duong**[3] · **Tat-Bao-Thien Nguyen**[4,5] · **Cach N. Dang**[6] · **Owen Conlan**[7]

Received: 18 March 2022 / Accepted: 13 December 2022 / Published online: 10 January 2023

Abstract Nutrients are important to promote plant growth and nutrient deficiency is the primary factor limiting crop production. However, excess fertilisers can also have a negative impact on crop quality and yield, cause an increase in pollution and decrease producer profit. Hence, determining the suitable quantities of fertiliser for every crop is very useful. Currently, the agricultural systems with internet of things make very large data volumes. Exploiting agricultural Big Data will help to extract valuable information. However, designing and implementing a large scale agricultural data warehouse are very challenging. The data warehouse is a key module to build a smart crop system to make proficient agronomy recommendations. In our paper, an electronic agricultural record (EAR) is proposed to integrate many separate datasets into a unified dataset. Then, to store and manage the agricultural Big Data, we built an agricultural data warehouse based on Hive and Elasticsearch. Finally, we applied some statistical methods based on our data warehouse to extract fertiliser information such as a case study. These statistical methods propose the recommended quantities of fertiliser components across a wide range of environmental and crop management conditions, such as nitrogen (*N*), phosphorus (*P*) and potassium (*K*) for the top ten most popular crops in EU.

Keywords Electronic agricultural record · Data warehouse · Nutrient · Crop yield

1 Introduction

In the last few years, there have been approximately 124 million people in 53 countries experiencing acute food insecurity. Besides, in another 42 countries, there are additional 143 million people were at the edge of facing acute hunger [12]. The world population in 2021 is 7.9 billion [41] and will increase to 8.6 billion by 2030 and 9.8 billion in 2050 [38]. So, the major urgent challenge for humans is the growing food demands of the annually increasing population [8, 42]. The problem is exacerbated by resources for crop production which are really limited, such as available freshwater and cropland [13]. There is an urgent need to increase crop yield by using new agricultural technologies, such as smart farming also called digital agriculture.

✉ Vuong M. Ngo
vuong.ngo@tudublin.ie; vuong.nm@ou.edu.vn

Thuy-Van T. Duong
duongthithuyvan@tdtu.edu.vn

Tat-Bao-Thien Nguyen
thienntb@uit.edu.vn

Cach N. Dang
cach@ut.edu.vn

Owen Conlan
Owen.Conlan@scss.tcd.ie

[1] School of Computer Science, Technological University Dublin, Dublin, Ireland

[2] Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam

[3] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[4] University of Information Technology, Ho Chi Minh City, Vietnam

[5] Vietnam National University, Ho Chi Minh City, Vietnam

[6] Science and Technology Application for Sustainable Development Research Group, Ho Chi Minh City University of Transport, Ho Chi Minh City, Vietnam

[7] School of Computer Science, Trinity College Dublin, Dublin, Ireland

Today, farmers often enhance soil nutrients through fertilisers to improve crop yields. The fertiliser application is crucial to protect global food and enhance the yield of cereals [15]. However, excess fertilisers will make toxic and negative impacts to crop quality and yield [7, 43]. Besides, fertilisers are very expensive and fertiliser waste will reduce producer incomes. Further, redundancy of fertilisers makes pollution of air, soil and water [34]. It can create high salt concentration to hurt beneficial soil microorganisms. It also makes eutrophication of surface water and groundwater, and greenhouse gas (GHG). Agricultural production contributes to global climate change. In 2008, it accounted for about 25% of global anthropogenic GHG emissions, such as carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O). Fertiliser production is the main source because it not only consumes large amounts of energy, but also produces N_2O and CH_4 in the manufacture of nitrate (NO_3^-) and ammonia (NH_3) fertilisers [39, 40]. So, fertiliser inputs need to be optimised to increase crop yield, farmer income and environmental quality [9, 36]. Fertilisers are composed of many elements, such as nitrogen (N), phosphorus (P), potassium (K), sulfur trioxide (SO_3) and magnesium oxide (MgO). An excess element can affect uptake other elements and cause both redundancy and deficiency of fertilisers. So the recommended quantity of very element in fertilisers also needs to be determined.

Smart farming applies statistical algorithms and data mining methods on historical data to discover new agricultural knowledge or build expert systems for improving farm productivity or being used tools for farmers [1, 24]. The global agricultural analytic market will increase more than 110% from $580 million in 2018 to $1.236 million in 2023 [22]. For example, the Bayer company collects data from farms, processes and analyses it, and then sells it back to the producers [28]. In this paper, firstly, we analyse many separated agricultural datasets by describing their Entity Relationship Diagrams (ERDs) to determine useful attributes, entities and objects. Secondly, a constellation schema, called Electronic Agricultural Record (EAR), is modeled and built. The EAR is adjustable to combine any agricultural data in making a united representation and used to build an agricultural data warehouse (DW). Thirdly, the information of the separated datasets is standardized, extracted, transferred and loaded into the EAR schema to make our EAR dataset which is an agricultural Big Data dataset. Fourthly, to store, process and manage the EAR dataset, we propose and implement an agricultural Big Data system on top of Hive and Elasticsearch. Finally, the proposed analytic methods use data about crops, their yields and their used fertiliser elements which are extracted from the unified EAR dataset. The methods are applied to discover the suitable quantities of fertiliser components for adapting criteria, such as quality advancing, yield increasing, profit improvement and environment protection. For example, we extract the suitable quantities of fertilisers N–P–K being 100–92–123 (kg/ha) for Spring Dried Beans and 126–106–112 (kg/ha) for Winter Oats. We study the three most popular fertiliser components (i.e. N, P, K) to adapt efficiently nutrients for the top ten most popular crops in EU (i.e. Spring Barley, Winter Barley, Spring Dried Beans, Winter Dried Beans, Spring Linseed, Forage Maize, Winter Oats, Winter Rape, Winter Rye, and Winter Wheat).

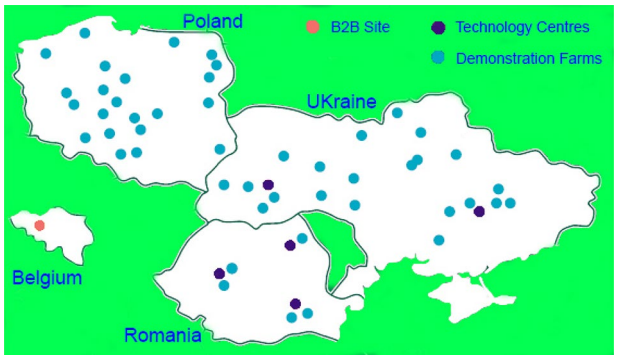


Fig. 1 Data sources in Belgium, Poland, Romania and Ukraine

The rest of this paper is organised as follows: in the next section, we reviewed related work on agricultural systems which propose useful information about fertilisers. In Sect. 3, the original datasets and their ERDs are presented and analysed. Specially, we build and propose an electronic agricultural record for agricultural data integration. In Sect. 4, the agricultural Big Data system is designed and implemented through the data warehouse. Section 5 presents a statistical methodology about fertiliser based on agricultural Big Data system as a case study. It proposes suitable quantities of fertiliser components for ten crops cross a wide range of environmental and crop management conditions. Finally, we present conclusion and future work in Sect. 6.

2 Related work

Many research papers used machine learning or analysis methods for fertiliser optimisation in digital agriculture. For instance, Barrett et al. [4] applied regression analysis algorithm to determine a suitable quantity fertilise N for cabbage. Cambouris [5] investigated the effect of soil texture and fertiliser N on corn yield. In [9], the back propagation neural network model was proposed to determine the recommended quantity of fertiliser N for maize. In [34], the authors experimented and analysed on a trial dataset about wheat to evaluate a long-term N management strategy which maintains a base level of fertiliser N rather than attempting

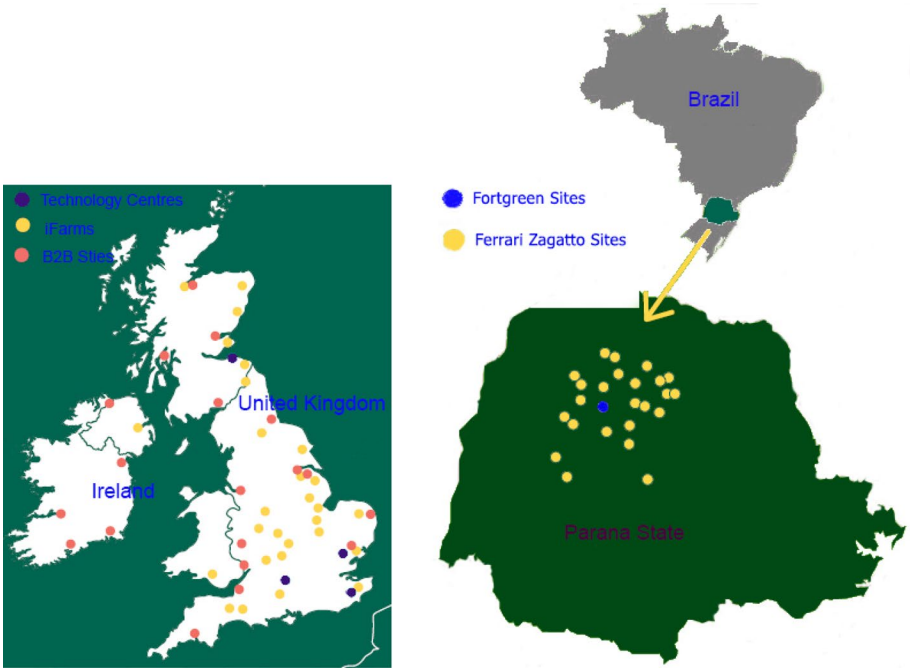


Fig. 2 Data sources in Ireland, United Kingdom and Brazil

to match *N* inputs to seasonal conditions. They concluded that a long-term *N* management strategy was potential to increase wheat yields, improve soil reserves and decrease environmental damage. In [43], a randomized complete block design on foxtail millet was conducted with four different rates of fertilisers *N* and *P*: no fertiliser, low, medium and high. The authors discovered fertiliser application at a medium rate (i.e. 180 kg/ha for *N* and 120 kg/ha for *P*) which would be suitable to improve yield and water use efficiency of foxtail millet in the semiarid regions. However, these papers only used and analysed trial data which is not real data collected from different farms. Especially, they did not adapt to Big Data in agriculture, where diverse external and internal factors have been combined, analysed and exploited together to give exact information or decisions to farmers or companies. Besides, they just researched fertiliser *N*, fertiliser *P* and one crop.

Moreover, some other papers analysed their datasets to support some decisions about fertilisers. For instance, Kaizzi et al. [19] developed a fertiliser optimization tool by the linear programming algorithm. The tool was used by smallholders to select fertiliser and the amount of each nutrient for their crops. In [37], the authors implemented an E-Water system used a multi-objective genetic algorithm. The system provided efficient management functions to find suitable amount of fertiliser and water. In [17], a system was built for developing optimum crop plan which exploited data regarding cropping pattern, rainfall, water status, land use etc. In [31], the authors presented a soil nitrate sensor technology based on spectroscopy analysis to manage and improve the used fertiliser *N*. In [16], the authors implemented the smart weather prediction using the internet of things and statistical models. The algorithm used data about temperature, rainfall, humidity, soil moisture and air pressure. However, the papers did not adapt Big Data exploitation and integration. Their datasets just include a few of agricultural information, similar to trial datasets of the papers mentioned in the above paragraph. Besides, they used available agricultural knowledge to decide on fertilisers; they did not propose recommended fertiliser quantities from agricultural Big Data.

Hence, to understand the importance of the data scale for agricultural analysis, the authors in [2, 18] and [33] analysed pretty large datasets. The paper [18] contains information about 10 crops and 5 cultivation strategies. The information was collected from crawling webs and interviewing selected farmers by face-to-face. The authors proposed suitable planting strategies for some crops to get high economic benefit. In [2], the authors presented some steps to build a data warehouse in agriculture based on Microsoft SQL to facilitate accessibility and explorations of open datasets. While, the paper [33] monitored N performance for different 7 crops across different arable farms and over time in Dutch. However, the papers did not present how to organise agricultural Big Data to deal with large scale and high performance in data analysis. In addition, their datasets are not considered as real Big Data.

Finally, to fix restrictions of all the above papers, a data warehouse constellation schema is used in [24, 26, 27] to

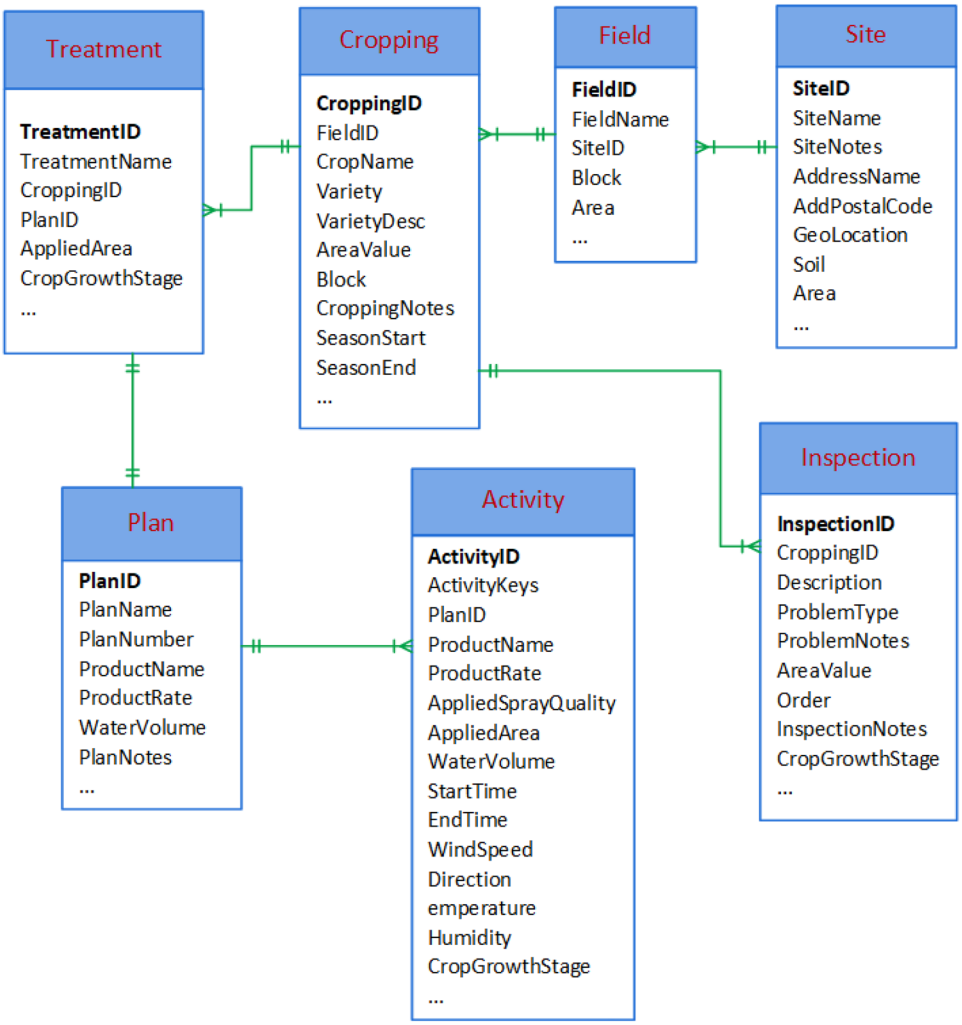


Fig. 3 A part of the ERD of Dataset 1

combine diverse datasets and comply with standards of agriculture Big Data. Nevertheless, there is not much important information about farming operations in the schema of [26], e.g. testing of crops, soils and nutrients, and management actions of treatments, spray, fertilisers, nutrients and inspection. The papers [26, 27] are about data warehouse design and implementation. While, in [24], the authors did not build an agricultural data warehouse. They just proposed information about insecticides, herbicides, soil properties and soil pH. Specially, the three papers did not present how to design a suitable schema, and use data mining algorithms to extract recommended quantities of fertilisers.

3 Electronic agricultural record and data integration

3.1 Original datasets

We study and assess 29 datasets supplied by a leading agronomy company from 2014 to 2018. In that, each dataset is about 1.4 gigabytes in textual format and has 18 tables of records on average. The agronomy company collected these datasets from its technology centres, operational systems, iFarms, field trials and research results [29]. The company has real agricultural data in 103 distribution centres, 70 demonstration farms, 12 million hectares of direct farm customer footprints, 34 input formulation and processing facilities, 45, 000 trial units and 800 sale forces at 7 countries being Belgium, Poland, Romania, Ukraine, Ireland United Kingdom and Brazil (Figs. 1 and 2)

Each dataset just contains a few of farming information. For example, the information in the crop dataset is almost about crops, such as crop name, season, crop condition, estimated yield, diameter, height and crop coverage percent and BBCH growth stage index. Besides, in the treatment dataset, there is treatment information for crop diseases, such as treatment name, form type, lot code, rate, applied date, description and comment.

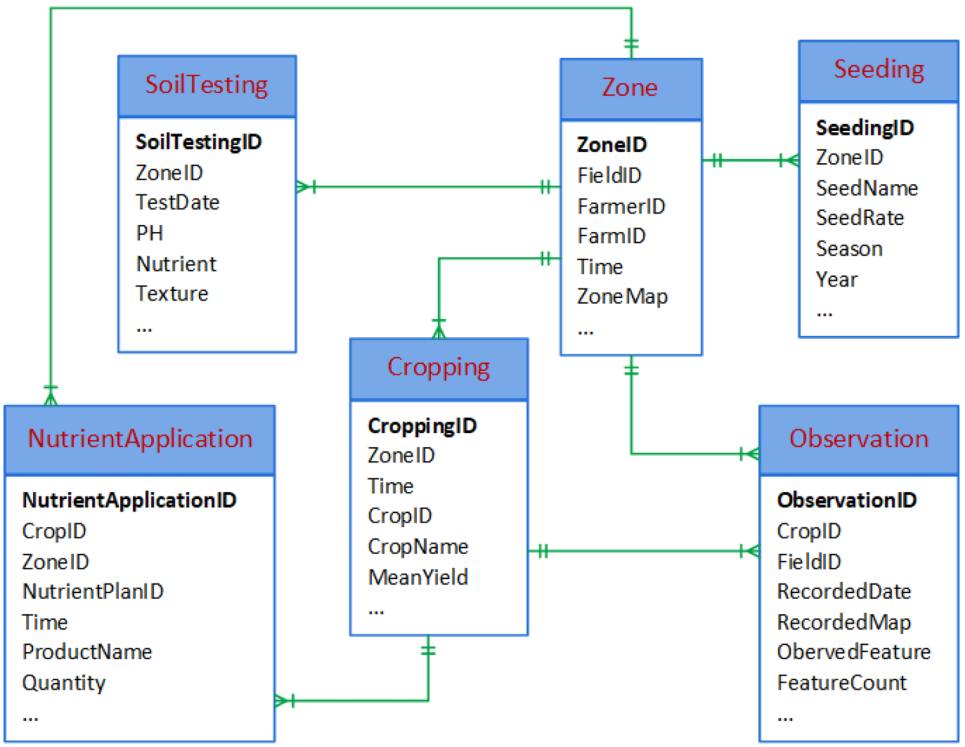


Fig. 4 A part of the ERD of Dataset 2

Fig. 5 Examples about a field divided into zones

3.2 Data standardization and analysis

Data standardization and integration are important tasks and need to be built in large scientific projects by the big enterprises which have various data sources. The above separate datasets will be cleaned, standardized and combined into a united dataset to analyse information and extract agricultural significant knowledge. To do this, studying and understanding the datasets are necessary and useful for not only designing a suitable unified schema, but also extracting, loading and transferring information from the separate datasets into a unified dataset. So, the ERDs of the complex original datasets need to be designed and explored.

A part of ERD of Dataset 1 is presented in Fig. 3 which has seven main entities being `Activity`, `Cropping`, `Field`, `Inspection`, `Plan`, `Site` and `Treatment`. The dataset contains information about farming operations on fields with detail plans, such as inspections, sprays and treatments for each crop. While, a part of ERD of Dataset 2 is presented in Fig. 4 which has six main entities being `Cropping`, `NutrientApplication`, `Observation`, `Seeding`, `SoilTesting` and `Zone`. It focuses on information about crops, soils and nutrients on zones. There are relationships between `Field`, `Site` in Dataset 1 and `Zone` in Dataset 2. Because, in agriculture, a site has some fields and a field has a few of zones. For example, Fig. 5 presents a field divided into 5 zones.

In Fig. 6, a part of ERD of Dataset 3 is presented. It has seven main entities being `Company`, `Cropping`, `DiaryNote`, `Fertiliser`, `Operation`, `Order` and `SoilSampling`. Among them, the `Company` and `Cropping` entities have the most relationships, i.e., 5 for every entity. It contains mainly information about agricultural companies/farmers and their operations on farms such as cropping, fertilisers, soils and orders. Finally, Fig. 7 shows a part of ERD of Dataset 4. It has six main entities being `Crop`, `CropState`, `Field`, `Pest`, `SoilAirStation` and `Treatment`. The dataset contains information about pests on fields

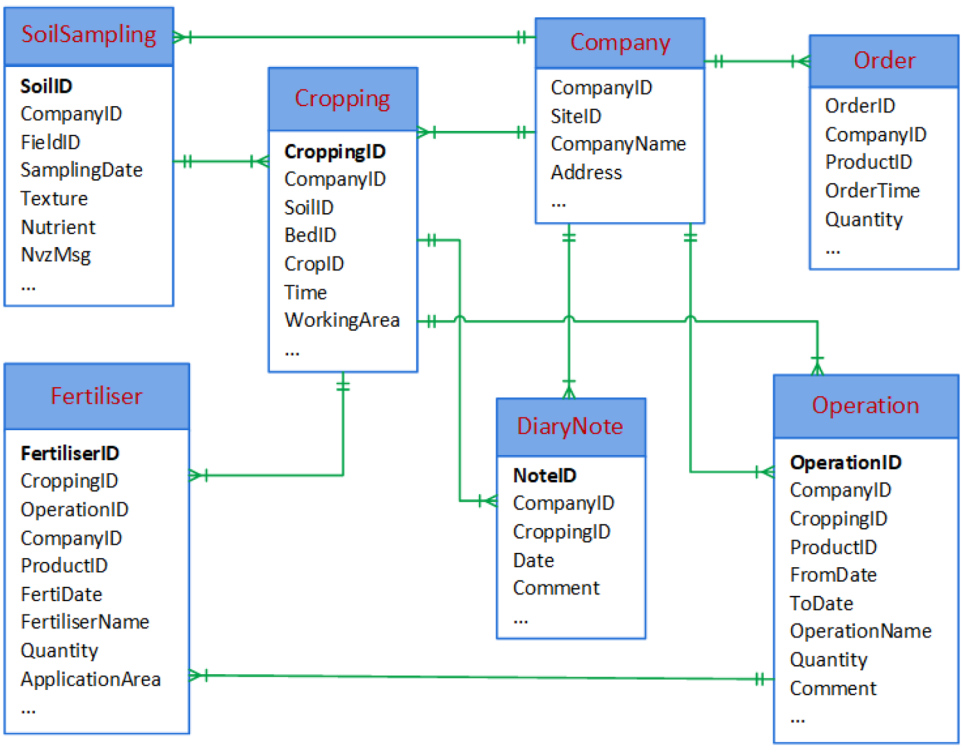


Fig. 6 A part of the ERD of Dataset 3

and how to treatment for crops. Besides, it also has information about soils and air temperatures.

In the original datasets, there is a lot of overlapped information which needs to be integrated carefully. For example, all four datasets contain information about crops through entities `Cropping`, `Crop` and `CropState`. While, datasets 1 & 4 contain information about fields which need to connect to information about sites and zones in datasets 1 & 2. Treatment information for crops is contained in datasets 1 & 4. Further, there are some overlapped information between the activity table in Dataset 1 and the operation table in Dataset 3. So, an agricultural constellation schema is designed and proposed to integrate information from our 29 separate datasets as below.

3.3 Electronic agricultural record

Our original datasets are collected from different sources and almost raw and semi-processed data. Specifically, the data is very complex, diverse, large, unstructured, conflicting and non-standardised. So, data in agriculture has all the attributes and criteria of Big Data [25]: (1) Volume: The quantity of agricultural data is fast increasing and is explosively made by external and internal sources, such as sensors, farming company operations, retail agronomists, satellites, intelligent equipment, government agencies, research centres and farmers. The external sources can help to supply information about local market accessing, pest and disease outbreak tracking, treatment and food price; (2) Variety: The data in agriculture has various formats and types which are structured data, text, imagery, multimedia, video, equations, metrics and models; (3) Velocity: The data in agriculture is being generated, collected and stored at very high rates. Because the sensing and mobile devices become cheaper and more efficient; (4) Veracity: the characteristics of agricultural data are inaccuracy, ambiguous, uncertain and inconsistent. Because the data is collected from various systems, sensors, operations and manual processes. Hence, agronomic Big Data harmonisation and integration are very difficult and challenge missions.

We need to propose and implement a suitable schema for integrating various separate datasets. Specially, this schema must adapt the criteria of data warehouse and the analysis on agricultural Big Data. So, firstly, in three kinds of DW schema models (i.e. star, snowflake and constellation), we select constellation schema for our agricultural enterprise DW which needs many fact tables and their dimension tables. Secondly, the ideas of agronomists and the ERDs of original agricultural datasets are reviewed and selected carefully to choose suitable attributes, entities and subjects for the schema.

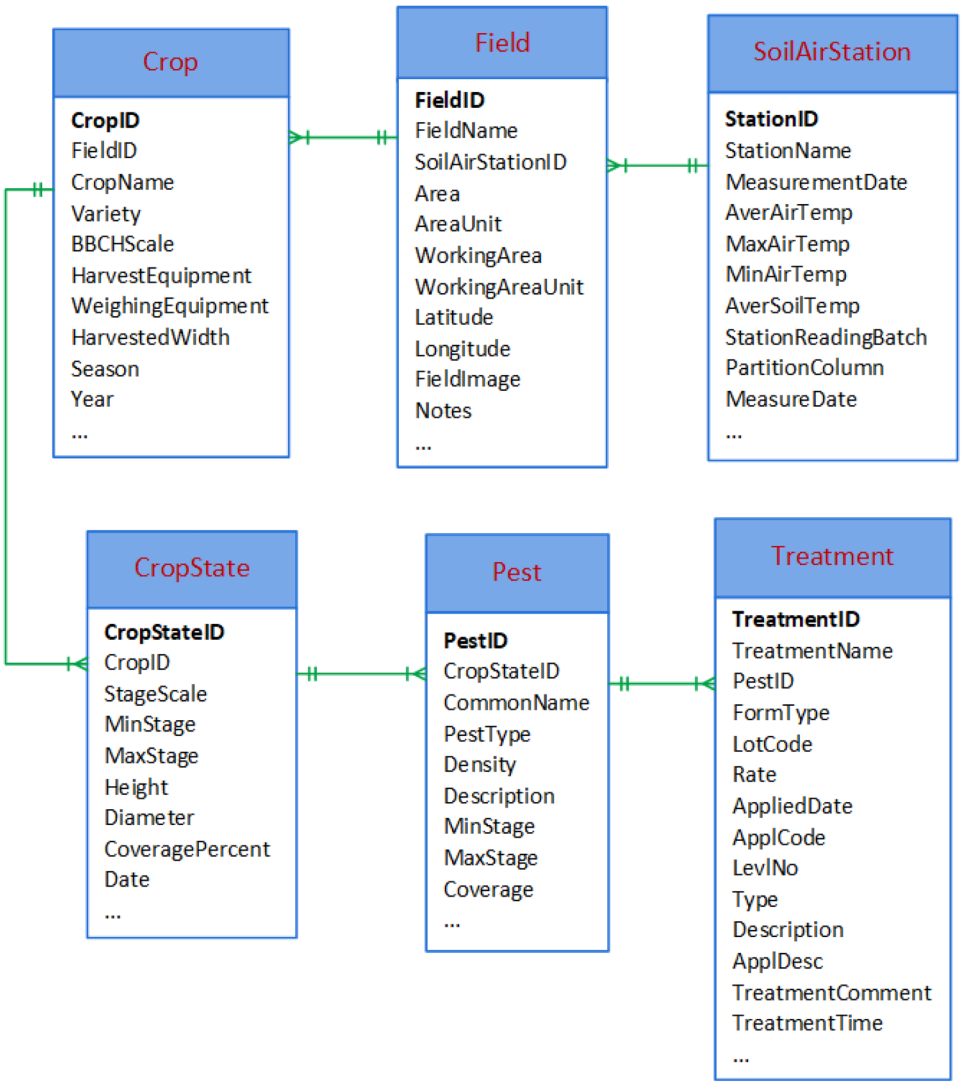


Fig. 7 A part of the ERD of Dataset 4

The proposed EAR schema (Electronic Agricultural Record) is presented in Fig. 8 which can handle high performance and high scalable. The EAR contains 5 fact tables being `FieldFact`, `Management Action`, `Order`, `Sale` and `Testing`. Among them, the `FieldFact` fact table describes data about fields, soils, fertilisers, nutrients, treatments, weather and pests. The `Management Action` fact table presents management operations on nutrients, fertilisers, inspection, treatments and spray. While, the `Order` and `Sale` fact tables include information about business operations. Finally, the `Testing` fact table includes testing operations on crops, soils and nutrients on zones.

In EAR, there also are 22 dimension tables, such as `CropState`, `Fertiliser`, `Field`, `Inspection`, `Soil`, `Treatment`, `Weather Station`, etc. Every dimension table includes information in detail about instances which are related framing operations. Some representative attributes of the dimension tables are described in Table 1. To exploit information, the HQL (Hive Query Language) or SQL (Structured Query Language) queries need to combine fact tables and dimension tables.

For example: Listing the information of fertiliser and treatment for each crop. The crops were harvested in spring or summer 2018, attached by 'black twitch' or whitefly pests, and have yield > 8 tons/ha. Besides, the soil in field has pH > 5.5, potassium >= 100 mg/l and magnesium (Mg) <= 80 mg/l. To answer this requirement, the HQL/SQL query needs to use the `FieldFact` fact table and the 6 dimension tables, being `Crop`, `Fertiliser`, `OperationTime`, `Pest`, `Soil` and `Treatment`, as folows.

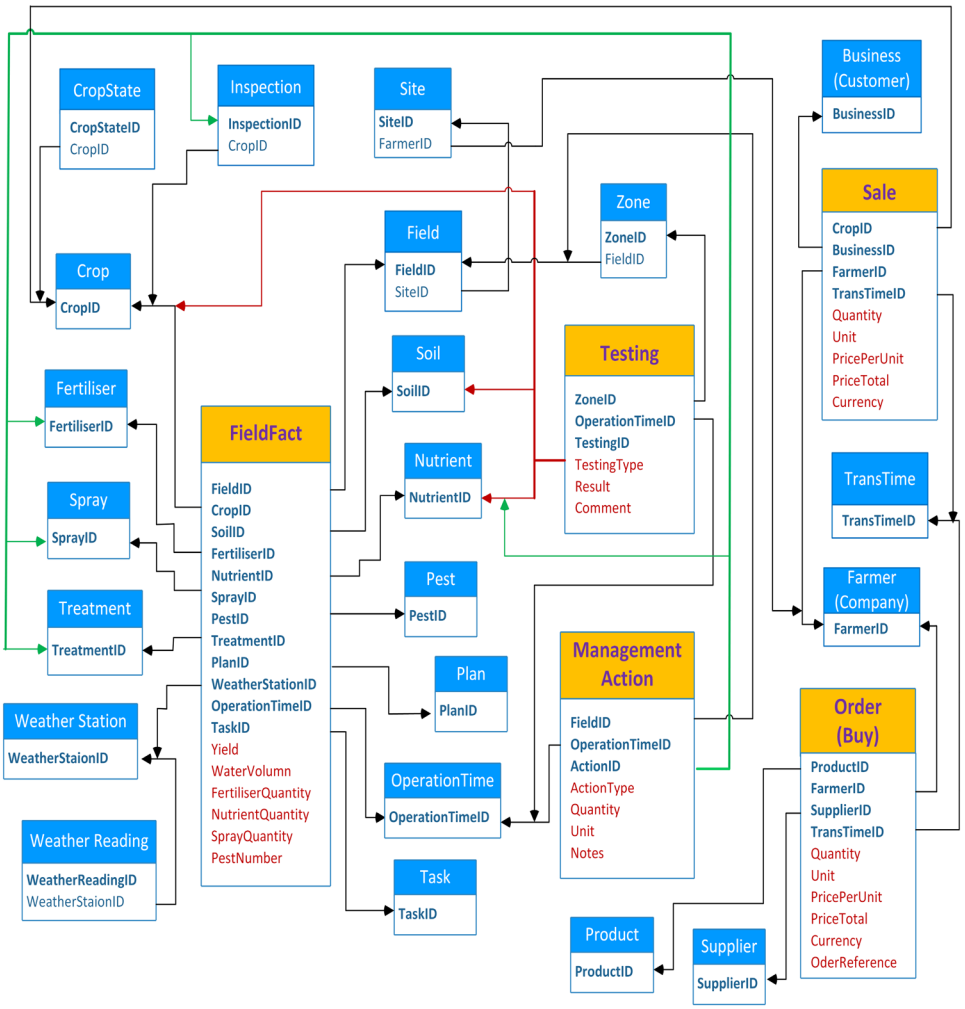


Fig. 8 A part of our EAR for smart farming

```
Select CR.CropName, FE.FertiliserName, FF.FertiliserQuantity,
       FE.ComponentNames, FE.ComponentPercentages,
       TR.TreatmentName, TR.TreatmentComment, TR.Rate
From Crop CR, Fertiliser FE, FieldFact FF, OperationTime OT,
     Pest PE, Soil SO, Treatment TR
Where FF.CropID = CR.CropID and
      FF.FertiliserID = FE.FertiliserID and
      FF.OperationTimeID = OT.OperationTimeID and
      FF.PestID = PE.PestID and
      FF.SoildID = SO.SoilID and
      FF.TreatmentID = TR.TreatmentID and
      Year(OT.StartDate) = '2018' and
      (OT.Season = 'Spring' or OT.Season = 'Summer') and
      (PE.CommonName = 'Black twitch' or PE.CommonName = 'Whitefly') and
      FF.Yield > 8 and SO.PH > 5.5 and
      SO.Potassium >= 100 and SO.Magnesium <= 80
```

Table 1 The 22 dimension tables and their representative attributes

No.	Tables	Representative attributes
1	Business	Name, Phone, Mobile, Address, Email, ContactPerson
2	Crop	CropName, EstimatedYield, BbchScale, HarvestEquipment, SeasonStart, SeasonEnd
3	Crop State	Height, Diameter, CoveragePercent, StageScale, MinStage, MaxStage
4	Farmer	Name, Phone, Mobile, Address, Email
5	Fertiliser	FertiliserName, ComponentNames, ComponentPrecentages, Status, Description, CompanyName, ManufactureDate
6	Field	Name, Reference, Block, Latitude, Longitude, Area, GeoPoints, Notes
7	Inspection	Date, ProblemType, Severity, AreaValue, Order, GrowthStage, Notes
8	Nutrient	NutrientName, Quantity, Unit, RecordedDate
9	Operation Time	Year, Season, StartDate, EndDate
10	Pest	Name, CommonName, Description, PestType, Coverage, Density, MinStage, MaxStage
11	Plan	PlanName, RegisNo, ProductName, ProductRate, Date, WaterVolume
12	Product	Name, Company, DateOfPurchase, GroupName
13	Spray	SprayName, ProductRate, ConfirmedHumidity, ConfirmedTemperature, ConfirmedWindSpeed, WaterVolume, SprayArea
14	Site	Name, Area, Address, GPS, OwnedBy
15	Soil	Sand, Silt, Clay, Nitrogen, Potassium, Phosphorus, Magnesium, Calcium, Unit, PH, TestDate, SoilType, SoilTexture, OrganicMatter, SupSoil, TopSoil
16	Supplier	Name, Phone, Mobile, Address, ContactName
17	Task	Description, Status, AppCode, DateStart, DateEnd, TaskInterval, Note
18	TransTime	OrderDate, Note, DeliveryDate, ReceivedDate
19	Treatment	Name, Description, Type, Rate, ProductCode, ApplicationCode, LevelNo, ApplicationDescription, Comment
20	Weather Reading	Date, Time, SPLite, Rainfall, AirTemperature, SoilTemperature, Humidity, LeafWetness, WindDirection, WindSpeed
21	Weather Station	Name, Region, Latitude, Longitude
22	Zone	Name, ZoneArea, ZoneType, GeometricPoints, Latitude, Longitude, SatelliteImage, YieldMap

4 Big data system implementation and design

4.1 Hive and elasticsearch

A data warehouse is a unified repository system for various heterogeneous data sources that a big company can collect from its business systems, research results and external inputs. The DW should adapt all the criteria of agricultural Big Data being volume, variety, velocity and veracity. Redshift[1], Cassandra[2], MongoDB[3], and Hive[4] are popular databases supporting efficiently the DW. Hence, we analyse them on data management, DW and technical features, and see Hive as be the best suited for our data problem. Hive is a data warehouse system built on Hadoop Distributed File System (HDFS)[5] for processing, writing and storing large datasets and running distributed applications [3, 20]. Hive supports many main features: (1) Online analytical processing (OLAP); (2) Storage capacity; (3) Data extract - transform - load (ETL); (4) Governance and data lifecycle management (via Hadoop); (5) Data science; (6) Security and monitoring; (7) Workload management; (8) Hive query language, similar to SQL; and (9) Replication-recovery.

However, Hive was not built for: (1) Real-time queries; (2) Data variety adaptation; (3) Online transaction processing (OLTP); (4) Iterative execution; and (5) Row-level update. So, it needs to be combined with Elasticsearch to overcome its disadvantages.

Elasticsearch [35] is an open-source, distributed search engine server built on top of Apache Lucene [14]. So, it is high scalable and high performance. Besides, Elasticsearch can support agricultural information and documents, such as JSON, text, images, figures, geo-spatial, multi-media. It uses the JSON over HTTP API and gets back a JSON reply for indexing and searching data. It is built on the Java programming language and hence it can run on different platforms. Finally, Elasticsearch supports functions to visualise, analyse and search easily.

4.2 System architecture

Our system architecture for agricultural Big Data is illustrated in Fig. 9 which contains three modules, namely

[1] http://aws.amazon.com/redshift

[2] http://cassandra.apache.org

[3] http://www.mongodb.com

[4] http://hive.apache.org

[5] http://hadoop.apache.org

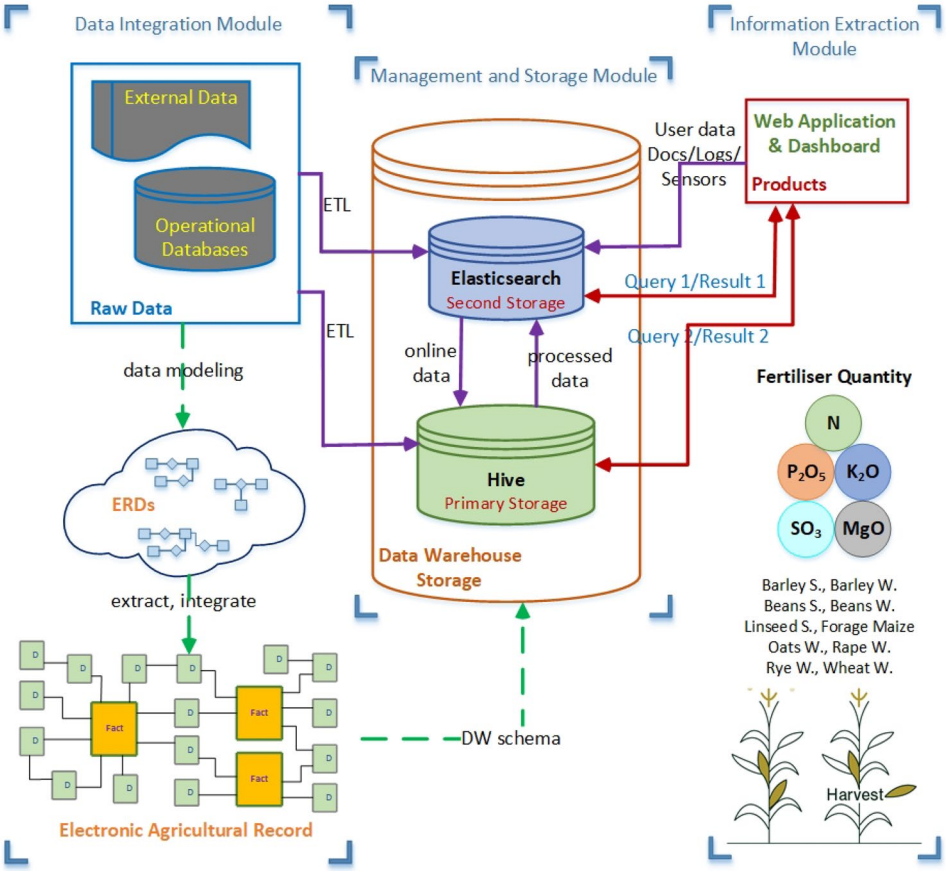


Fig. 9 Agricultural data warehouse

`Data Integration`, `Management and Storage`, and `Information Extraction`. The `Data Data Integration Module` has components being Raw Data (External and Operational Databases), ERDs and Electronic Agricultural Record (see more in Sect. 3). The `Management and Storage Module` is DW Storage including the Elasticsearch component and the Hive component (see more in 4.1 and below). The `Information Extraction Module` presents Web Application and Dashboard products which apply data mining algorithms to extract statistical information about fertiliser components corresponding to each crop (see more in Sect. 5).

In DW Storage, Elasticsearch and Hive receive data in Operational Databases and External Data from Raw Data module through the ETL tool. Products module also sends information to Elasticsearch and Hive. In that, Elasticsearch receives real-time data in dashboard and web application. Otherwise, Elasticsearch sends analysed answers which need to be retrieved in real-time to Products module. With queries having multiplex calculation, the Hive component will receive from and process for the Products module directly. Hive also stores the online data from Products module through Elasticsearch and sends processed data to store in Elasticsearch.

4.3 Our primary storage performance evaluation

The reading performance of our primary DW storage (i.e. Hive) needs to be evaluated because a DW is used primarily for reporting and analysing data, not for writing data. In addition, the secondary DW storage (i.e. Elasticsearch) is near real-time in indexing and searching [11]. So it does not need to be evaluated the performance. We use Hadoop 2.6.5, Hive 2.3.3, JDK 1.8.0_171 and MySQL 5.7.22 for evaluation. The software are installed on Ubuntu Bash 16.04.2 on Windows 10 and a Dell laptop having 16 GB memory and Intel Core i7 CPU (2.40 GHz).

Our database in Hive is copied to MySQL to evaluate and compare run-time performance. The popular HQL/SQL commands, namely Where, Group by, Having, Right (left) Join, Order by and Union, are used to create 10 query groups for testing. Each query group uses a few of commands and includes five queries (see Table 2). In addition, the queries also applies operations, e.g. Sum, Count, Or, And, Like, Min and $\leq$, to the commands to express complex queries. Each query is evaluated the runtime in three times and taken its average runtime.

The mean executive times of the 10 query groups on MySQL and our primary storage (Hive) are shown in

Table 2 The query groups with combined commands

Group	Commands	Group	Commands
G_1	Where	G_6	Where, Right (left) Join and Order by
G_2	Where and Group by	G_7	Where, Group by and Having
G_3	Where and Right (left) Join	G_8	Where, Group by, Having and Order by
G_4	Where and Union	G_9	Where, Group by, Having, Right (left) Join and Order by
G_5	Where and Order by	G_{10}	Where, Group by, Having, Union and Order by

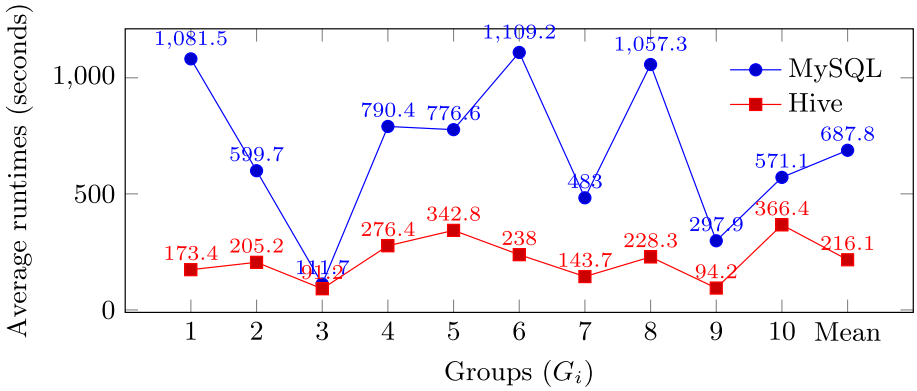


Fig. 10 Mean executive times of MySQL and Hive in every query group

Fig. 10. The average runtimes of a query on the primary storage and MySQL are 216 seconds and 688 seconds, respectively. So, MySQL is lower 3.2 times than the primary storage. If our DW is deployed on a distributed

Table 3 Mean yield (ton/ha) in each yield group of every crop type

Crop Name	Gr.	Mean Yield	%	Crop Name	Gr.	Mean Yield	%
Barley S.	1	8.16	+91.5	Maize F.	1	47.00	+99.0
	2	7.32	+71.8		2	44.67	+89.1
	3	6.52	+53.1		3	40.27	+70.5
	4	5.81	+36.4		4	32.63	+38.1
	5	4.26	0		5	23.62	0
Barley W.	1	10.92	+111.6	Oats W.	1	8.06	+42.9
	2	8.29	+60.7		2	7.50	+33.0
	3	7.30	+41.5		3	7.00	+24.1
	4	6.40	+24.0		4	6.93	+22.9
	5	5.16	0		5	5.64	0
Beans S.D.	1	5.21	+382.4	Rape W.	1	4.59	+94.5
	2	4.32	+300.0		2	4.00	+69.5
	3	3.79	+250.9		3	3.59	+52.1
	4	1.92	+77.8		4	3.15	+33.5
	5	1.08	0		5	2.36	0
Beans W.D.	1	6.15	+80.9	Rye W.	1	39.90	+124.5
	2	5.51	+62.1		2	32.39	+82.3
	3	4.97	+46.2		3	28.23	+58.9
	4	4.52	+32.9		4	23.19	+30.5
	5	3.40	0		5	17.77	0
Linseed S.	1	2.28	+430.2	Wheat W.	1	11.74	+71.9
	2	1.57	+265.1		2	10.22	+49.6
	3	1.30	+202.3		3	9.32	+36.5
	4	0.84	+95.3		4	8.55	+25.2
	5	0.43	0		5	6.83	0

Table 4 The mean quantities of general fertiliser and NPK group (kg/ha)

Crop Name	Gr.	Total (kg/ha)	NPK (kg/ha)	NPK in total (%)	Crop Name	Gr.	Total (kg/ha)	NPK (kg/ha)	NPK in total (%)
Barley S.	1	882	415	47	Maize F.	1	812	460	57
	2	712	422	59		2	1,172	574	49
	3	881	368	42		3	789	257	33
	4	816	342	42		4	1,698	723	43
	5	1069	376	35		5	485	186	38
Barley W.	1	1240	486	39	Oats W.	1	724	344	48
	2	805	423	53		2	397	289	73
	3	1316	539	41		3	461	297	64
	4	498	274	55		4	616	367	60
	5	652	255	39		5	291	193	66
Beans S.D.	1	436	315	72	Rape W.	1	920	330	36
	2	391	294	75		2	730	374	51
	3	391	270	69		3	998	351	35
	4	291	215	74		4	1,190	470	40
	5	268	212	79		5	942	343	36
Beans W.D.	1	300	224	75	Rye W.	1	1,036	421	41
	2	304	209	69		2	1,129	495	44
	3	257	128	50		3	1,403	501	36
	4	657	201	31		4	1,083	479	44
	5	306	205	67		5	799	478	60
Linseed S.	1	580	230	40	Wheat W.	1	1,403	578	41
	2	403	249	62		2	1,424	580	41
	3	377	223	59		3	1,179	516	44
	4	315	217	69		4	1,009	491	49
	5	367	164	45		5	1,134	532	47

system or a cloud storage, we believe that its runtime performance will be faster than MySQL many times.

5 Case study: fertiliser knowledge extraction

5.1 Classification based on yield

From the EAR dataset, we analyse information related to fertiliser and crop yield in every field. This includes crop name, yield, field identification, year, season, the quantities of total fertiliser and main elements of fertiliser, being *N*, *P* and *K*. We classify each EAR record into one of the five yield groups of every crop type that is based on crop type and yield of each record. Each yield group includes 20% of the amount of records of each crop type. Among them, based on yield, group 1 is the highest 20% group, group 3 is medium 20% group and group 5 is the lowest 20% group. After that, in each group, the mean values of yield, total fertiliser, NPK group, and fertilisers *N*, *P*, and *K* are calculated.

Table 3 describes the top ten most popular crops in EU, which are Barley S. (Spring Barley), Barley W. (Winter Barley), Beans S.D. (Spring Dried Beans), Beans W.D. (Winter Dried Beans), Linseed S. (Spring Linseed), Maize F. (Forage Maize), Oats W. (Winter Oats), Rape W. (Winter Rape), Rye W. (Winter Rye), and Wheat W. (Winter Wheat). The mean yield of each yield group of each crop type is shown in this table. In addition, in each crop yield, the different percentages between yield group 5 (the lowest yield group) and other yield groups are also presented clearly. Specifically, in Barley S., group 5 has mean yield of 4.26 ton/ha. While, group 1 and 2 have mean yield of 8.16 ton/ha and 7.32 ton/ha, and are higher than group 5 about 91.5% and 71.8%, respectively. Besides, group 3 and 4 have mean yield of 6.52 ton/ha and 5.81 ton/ha, and are higher than group 5 about 53.1% and 36.4%, respectively. Specially, in Linseed S. and Beans S.D., group 1 is higher than group 5 about 430.2% and 382.4%.

Fertilisers have been used since the start of agriculture to supply one or more essential nutrients to the growth of crops. Today, farmers often use fertilisers being either mined or manufactured. However, fertilisers are very expensive and can harm the environment. Besides, excess fertilisers will badly impact crop quality and yield. So, the right fertiliser quantities for every crop should be used. The fertilisers are composed of many major, secondary and trace elements.

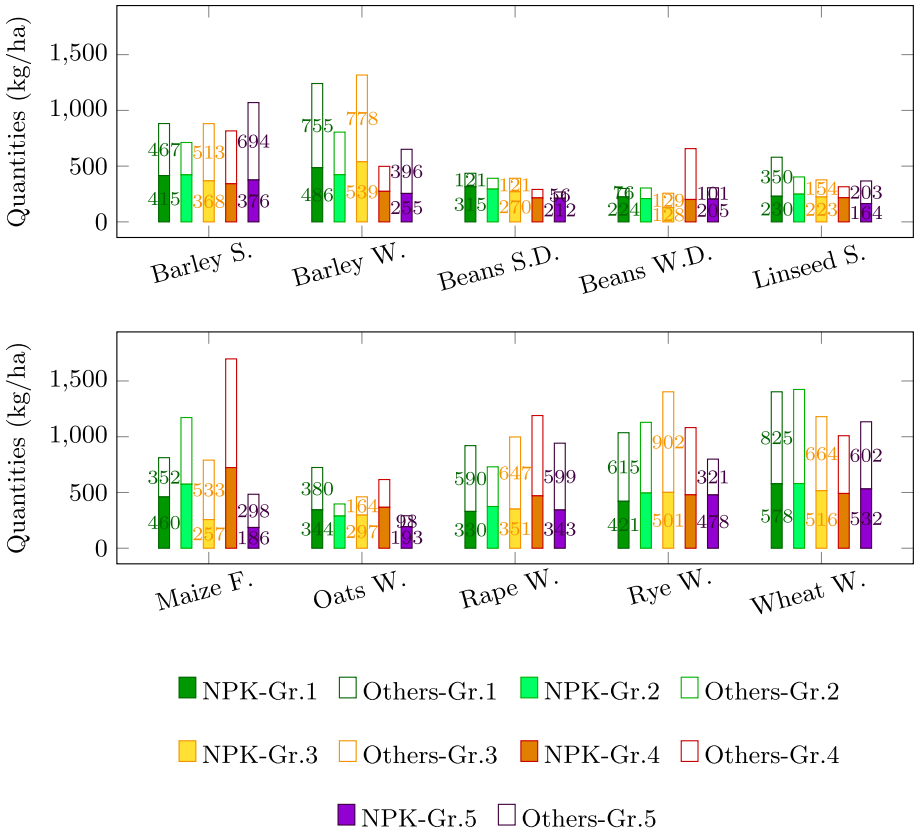


Fig. 11 Mean fertiliser quantities of the NPK groups and the other groups

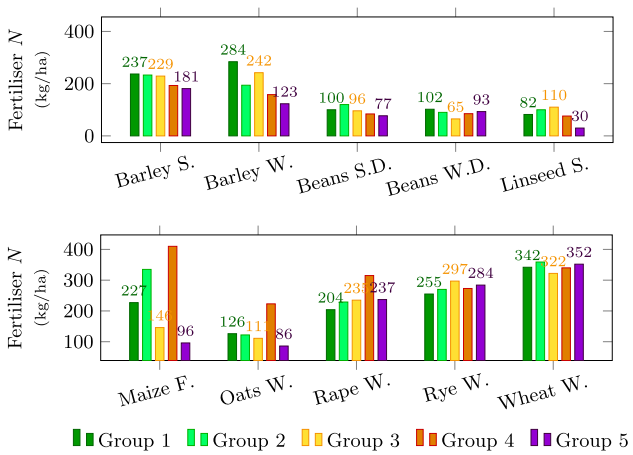


Fig. 12 Mean nitrogen fertiliser quantities

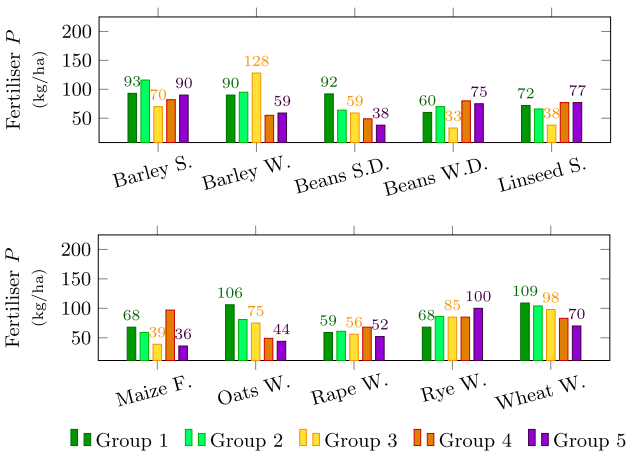


Fig. 13 Mean phosphorus fertiliser quantities

Among them, the trace elements are ions of Chlorine (*Cl*), Iron (*Fe*), Manganese (*Mn*), Zinc (*Zn*) and Copper (*Cu*). The secondary elements are Calcium (*Ca*), Magnesium (*Mg*) and Sulfur (*S*). While, the primary elements, being Nitrogen (*N*), Phosphorus (*P*) and Potassium (*K*), are used in large quantities by plants and play a key role in plant nutrition [32]. So, the "Big 3" nutrients will be detected and analysed more careful.

5.2 Crop and NPK group correlation

With classifying through the five yield groups of ten crops, we extract information about the quantities of fertiliser in Table 4 and Fig. 11. They show the mean quantities of total fertiliser, NPK group and other group of each crop in each

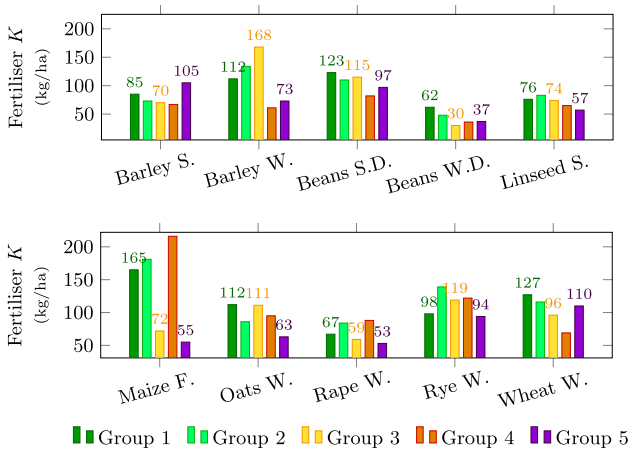


Fig. 14 Mean potassium fertiliser quantities

yield group. In addition, they also present the percentage of NPK group in total fertiliser.

From Table 4 and Fig. 11, there are not significant differences among yield groups in Barley W., Beans S.D., Rape W. and Rye W. While, in the remain 6 crops being Barley S., Beans W.D., Linseed S., Maize F., Oats W. and Wheat W., the percentages of NPK groups between high-yield groups and low-yield groups are clearly separate. Hence, we recommend the suitable percentages of NPK groups are about 47% for Barley S., 75% for Bean W.D., 40% for Linseed S., 57% for Maize F., 48% for Oats W. and 41% for Wheat W. Besides, the suitable total fertilisers are about 882 kg/ha for Barley S., 300 kg/ha for Bean W.D., 580 kg/ha for Linseed S., 812 kg/ha for Maize F., 724 kg/ha for Oats W. and 1,403 kg/ha for Wheat W. Moreover, the ratio of *N*, *P* and *K* in the NPK group is also important for developing crops. So we continue to analyse these ratios in next section.

5.3 Crop and N–P–K ratio correlation

Nitrogen is very important because it is a major component of chlorophyll, amino acids (being the building blocks of proteins) and nucleic acids (e.g. DNA). Without nitrogen, plants wither and die. Moreover, nitrogen deficiency will limit plant growth, make yellow leaves and be easily attacked by diseases and insects. On other hand, nitrogen redundancy can cause excessive growth of aquatic plants and algae which use up dissolved oxygen and clog water intake to affect growth of crops. Besides, nitrogen can pervade in drinking water, environmental damage, and be harmful to human or livestock [6]. So, in each crop, we need to determine the suitable quantity of nitrogen fertiliser to make the highest yield. Figure 12 presents mean quantities of fertiliser *N* used in yield groups of crops that are extracted from the data warehouse.

Phosphorus is a vital component of DNA and RNA. Especially, it captures and converts the sun's energy into useful plant compounds. *P* deficiency makes a stunting of the plant in the early growth, and affects both seed development and normal crop maturity in the late growth [30]. While, too much *P* can be toxic. Because, waste *P* can easily flow into water and cause algal blooms and excessive vegetative growth. Besides, it also impedes the uptake of *Fe* and *Zn*. The mean quantities of fertiliser *P* in yield groups for crops are presented in Fig. 13.

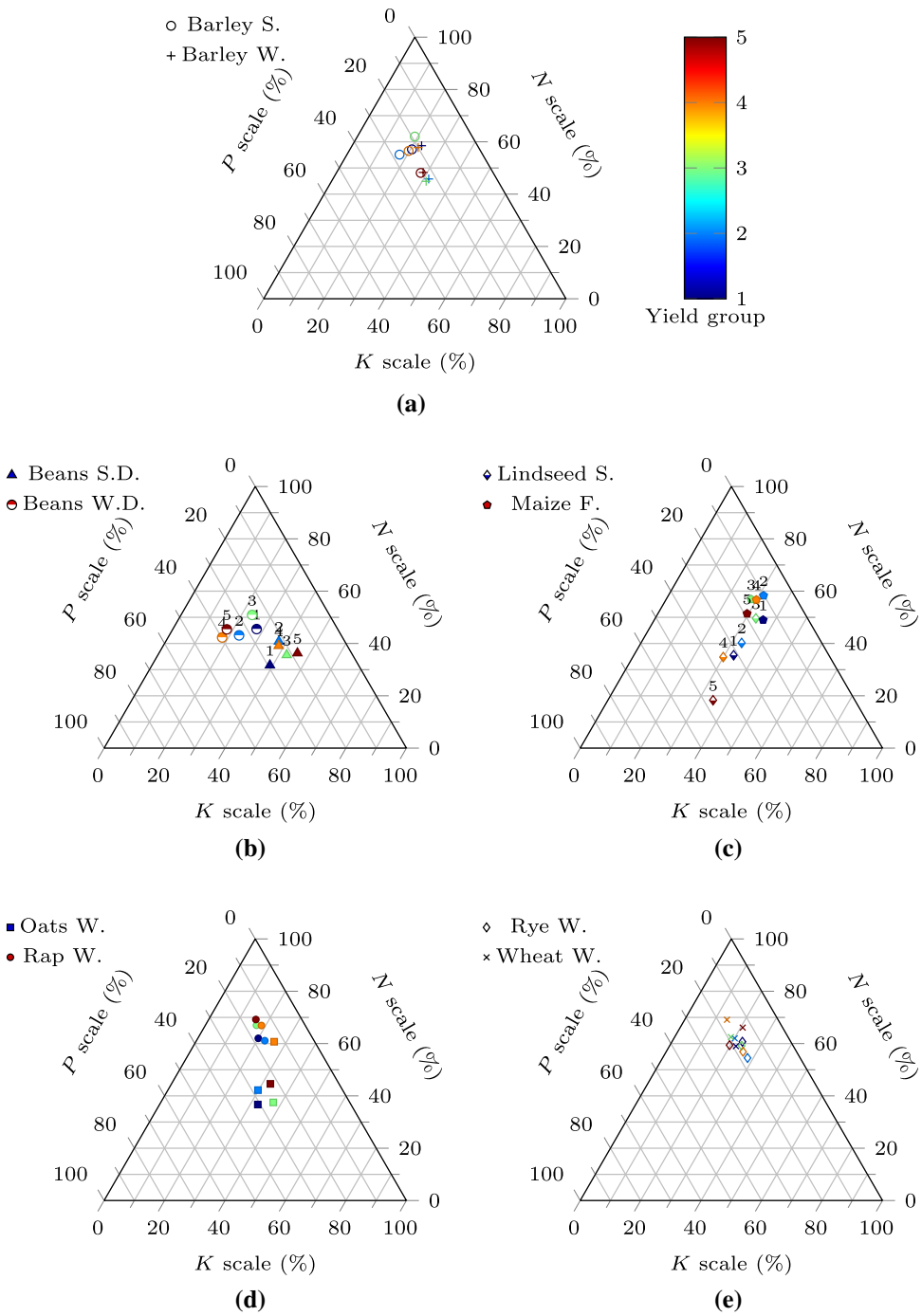
Potassium is essential in photosynthesis, enzyme activation and protein increment to sustain growth and reproduction of plants. *K* deficient plants are less resistant to drought, excess water, disease, insect attack, frost and cold [21]. Specially, *K* deficiency makes yellow firing leaf and poor root development. While, excess *K* can affect uptake other nutrients, such as *N* and *Mg*. We extract the mean quantities of fertiliser *K* in yield groups of crops from the data warehouse and present in Fig. 14.

The N–P–K ratio in fertiliser is important for developing crops. So we use ternary graphs to analyse correlation between the ratio of *N*, *P* and *K* and crop yield. In Fig. 15a, c, e, we don't see the separation among yield groups of Barley S., Barley W., Lindseed S. and Rye W. However, there are significant differences between high-yield groups and low-yield groups of Beans S.D., Beans W.D., Maize F., Oats W., Rape W. and Wheat W. in Fig. 15b–e. So we can propose the suitable ratio of *N*, *P* and *K* in the NPK group for the 6 crops based on information of their group 1. Combining with information extracted in Sect. 5.2 and information in Figs. 12, 13, 14 and 15, we propose the suitable ratio of *N*, *P* and *K* in the NPK group, and the suitable percentage of NPK group in fertiliser total for crops in Table 5. The suitable quantities of *N*, *P* and *K* for Beans S.D., Beans W.D., Maize F., Oats W., Rape W. and Wheat W. are extracted and recommended. While, we can only propose the suitable quantity of group NPK for Barley S. Besides, we do not have enough information to recommend the fertiliser quantities for Barley W., Linseed S. and Rye W.

6 Conclusion and future work

In this paper, we analysed and integrated many original agricultural datasets to determine useful dimensional and fact tables, and their attributes and relationships for proposing an EAR. Based on the EAR, also being a fact constellation schema, various separate datasets are extracted, transferred and loaded into a unified crop dataset. The EAR is adjustable and scalable to new datasets and variety standards of Big Data analytics in agriculture. Besides, we also designed and implemented an agricultural DW based on Hive and Elasticsearch which adapted criteria about DW and Big Data, such

Fig. 15 N–P–K ratio



as security, high performance, high storage, variety data and data science support. Specially, from the unified EAR dataset, we presented a data analysis method based on crop yield classification with fertiliser components. The studied results showed that in some crops, the more fertilisers used, the more yield increased. However, in many other crops, they are suitable to medium fertiliser quantities and their yield decreased as using more fertilisers. We proposed the suitable quantities of the NPK group, *N*, *P* and *K* in various season and farms on the top ten famous crops in continental Europe, Ireland, United Kingdom and Brazil.

With the scope of the paper, we exploited information about fertiliser as a case study. However, the crop yield improvement is affected not only the fertiliser components, but also available soil properties, soil texture and nutrient translocation. So, in the future, we will apply our deep learning [10] and machine learning [23] algorithms to discovery relation of fertiliser components, soil properties, adjuvants and water requirements on increasing crop yield. Besides,

Table 5 The proposed quantities of the NPK group and its elements in each crop are extracted from Big Data

Crop Name	Total (kg/ha)	NPK (kg/ha)	NPK in total (%)	N (kg/ha)	P (kg/ha)	K (kg/ha)	N in NPK (%)	P in NPK (%)	K in NPK (%)
Barley S.	882	415	47	n/a	n/a	n/a	n/a	n/a	n/a
Barley W.	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Beans S.D.	436	315	72	100	92	123	32	29	39
Beans W.D.	300	224	75	102	60	62	46	27	27
Linseed S.	580	230	40	n/a	n/a	n/a	n/a	n/a	n/a
Maize F.	812	460	57	227	68	165	49	15	36
Oats W.	724	344	48	126	106	112	37	31	32
Rape W.	920	330	36	204	59	67	62	18	20
Rye W.	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Wheat W.	1403	578	41	342	109	127	59	19	22

time series and the weather factors, such as temperature of air and soil, sunshine, rain fall, humidity and wind speech, are powerfully affect to crop yield and also will be studied.

Acknowledgements This research is an extended work of [26]. Besides, the research also uses a part of dataset of [24]. The two conference papers [26] and [24] were partly funded under the Science Foundation Ireland, Strategic Partnerships Programme (16/SPP/3296).

Funding Open Access funding provided by the IReL Consortium.

Declarations

Competing interest The authors declare that they have no known competing personal relationships or financial interests that could have appeared to influence the work reported in the paper.

References

1. Admass W (2022) Developing knowledge-based system for the diagnosis and treatment of mango pests using data mining techniques. Int J Inf Technol 14:1495–1504
2. Anupama G, Jain R, Falk T et al (2020) Data warehousing for open data sharing and decision support in agriculture: a case study of the vdsa knowledge bank and its development process. Int J Inf Technol 12:923–931
3. Apache Hive: Apache Hive A Complete Guide - 2021 Edition. The Art of Service - Apache Hive Publishing (2021)
4. Barrett CE et al (2018) Optimization of irrigation and n-fertilizer strategies for cabbage plasticulture system. Sci Hortic 234(14):323–334
5. Cambouris AN (2016) Corn yield components response to nitrogen fertilizer as a function of soil texture. Can J Soil Sci 96(4):386–399
6. Carswell A et al (2022) Combining targeted grass traits with red clover improves grassland performance and reduces need for nitrogen fertilisation. Eur J Agron 133:126433
7. Cordero E et al (2018) Fertilisation strategy and ground sensor measurements to optimise rice yield. Eur J Agron 99:177–185
8. Cui Z et al (2018) Pursuing sustainable productivity with millions of smallholder farmers. Nature 555:363–366
9. Dong Y et al (2020) Precision fertilization method of field crops based on the wavelet-bp neural network in china. J Clean Prod 246:2
10. Duong LT, Le NH, Tran TB, Ngo VM, Nguyen PT (2021) Detection of tuberculosis from chest x-ray images: Boosting the performance with vision transformer and transfer learning. Expert Syst Appl 184:115519
11. Elasticsearch team: Near real-time search (2022) https://www.elastic.co/guide/en/elasticsearch/reference/master/near-real-time.html, accessed January 01, 2022
12. FAO-FSIN: Global Report on Food Crises 2019 (2019) Food Security Information Network, FAO
13. FAO-Land-Water report: Promoting coherence and coordination on land and water. Land and Water Division, FAO (2020) http://www.fao.org/land-water/overview/en/. Accessed 01 January 2022
14. Gheorghe R et al (2016) Elasticsearch in action. Manning Publications Co, New York
15. Huang J et al (2017) Nitrogen and phosphorus losses and eutrophication potential associated with fertilizer application to cropland in china. J Clean Prod 159:171–179
16. Islam S, Akter M, Uddin M (2021) Design and implementation of an internet of things based low-cost smart weather prediction system. Int J Inf Technol 13:2001–2010
17. Jain R, Kingsly I, Chand R et al (2019) Methodology for region level optimum crop plan. Int J Inf Technol 11:2019
18. Jiang YM et al (2019) Big data analysis applied in agricultural planting layout optimization. Appl Eng Agric 35(2):147–162
19. Kaizzi KC et al (2017) Fertilizer use optimization: Principles and approach. In: Wortmann CS, Sones K (eds) Fertilizer use optimization in sub-Saharan Africa, pp. 9–19. CABI
20. Kunigk J, Buss I, Wilkinson P, George L (2019) Architecting modern data platforms: a guide to enterprise hadoop at scale. O Reilly
21. Liu J, Hu T, Feng P, Yao D, Gao F, Hong X (2021) Effect of potassium fertilization during fruit development on tomato quality,

potassium uptake, water and potassium use efficiency under deficit irrigation regime. Agric Water Manag 250:106831
22. Market & Market: Agriculture Analytics Market (2019) https://www.marketsandmarkets.com/Market-Reports/agriculture-analytics-market-255757945.html, accessed January 01, 2022
23. Ngo VM, Duong TVT, Nguyen TBT, Nguyen PT, Conlan O (2021) An efficient classification algorithm for traditional textile patterns from different cultures based on structures. J Comput Cult Herit 14(4):1–22
24. Ngo VM, Kechadi MT (2020) Crop knowledge discovery based on agricultural big data integration. In: Proceedings of the 4th International Conference on Machine Learning and Soft Computing (ICMLSC). pp. 46–50. ACM
25. Ngo VM, Le-Khac NA, Kechadi MT (2018) An efficient data warehouse for crop yield prediction. In: Proceedings of the 14th International Conference on Precision Agriculture (ICPA-2018). pp. 3:1–3:12
26. Ngo VM, Le-Khac NA, Kechadi MT (2019) Designing and implementing data warehouse for agricultural big data. In: Proceedings of the 8th International Congress on BigData (BigData-2019). LNCS, vol. 11514, pp. 1–17. Springer
27. Ngo VM, Le-Khac NA, Kechadi MT (2020) Data warehouse and decision support on integrated crop big data. Int J Bus Process Integr Manag (IJBPIM) 10(1):17–28
28. Noel, A. M.: Data becomes cash crop for big agriculture (2019), https://www.bloomberg.com/news/articles/2019-03-13/data-becomes-cash-crop-for-big-agriculture, accessed January 01, 2022
29. Origin team: Perform, sustain, grow (2019) In: Annual report and accounts
30. Peçanha DA et al (2021) Phosphorus fertilization affects growth, essential oil yield and quality of true lavender in brazil. Ind Crops Prod 170:113803
31. Rogovska N et al (2019) Development of field mobile soil nitrate sensor technology to facilitate precision fertilizer management. Precision Agric 20(1):40–55
32. Shtull-Trauring E, Cohen A, Ben-Hur M, Israeli M, Bernstein N (2022) Npk in treated wastewater irrigation: regional scale indices to minimize environmental pollution and optimize crop nutritional supply. Sci Total Environ 806:150387
33. Silva JV (2021) Agronomic analysis of nitrogen performance indicators in intensive arable cropping systems: an appraisal of big data from commercial farms. Field Crop Res 269:108176
34. Smith CJ et al (2019) Using fertiliser to maintain soil inorganic nitrogen can increase dryland wheat yield with little environmental cost. Agr Ecosyst Environ 286:1–15
35. Srivastava A (2020) Learning elasticsearch 7.x: index, analyze, search and aggregate your data using elasticsearch. BPB Publications, Berlin
36. Todorovic M et al (2018) Impact of different water and nitrogen inputs on the eco-efficiency of durum wheat cultivation in mediterranean environments. J Clean Prod 183:1276–1288
37. Udiasa A et al (2018) A decision support group to enhance agricultural growth in the mekrou river basin (west africa). Comput Electron Agric 154:467–481
38. United Nations: 17 Goals to Transform Our World, Sustainable Development Goals (2021), release on 21 June 2021
39. Vermeulen SJ et al (2012) Annual review of environment and resources. Clim Change Food Syst 37:195–222
40. Wang S et al (2016) Effect of split application of nitrogen on nitrous oxide emissions from plastic mulching maize in the semiarid loess plateau. Agricult Ecosyst Environ 220:21–27
41. World Ometers: Current World Population (2022) https://www.worldometers.info/world-population/, accessed January 01, 2022
42. Xu J et al (2020) Effects of irrigation and nitrogen fertilization management on crop yields and long-term dynamic characteristics of water and nitrogen transport at deep soil depths. Soil Tillage Res 198:2
43. Zhang X et al (2020) Optimizing fertilization under ridge-furrow rainfall harvesting system to improve foxtail millet yield and water use in a semiarid region, china. Agric Water Manag 227:1–12