Research on the Importable of Machine Learning Technology in Japanese Vegetable Distribution Field

Research on the Importable of Machine Learning Technology in Japanese Vegetable Distribution Field

（日本の野菜流通現場への機械学習技術の導入に関する研究）

2018

The United Graduate School of Agricultural Sciences, Gifu University

Science of Biological Production

（Gifu University）

Li Ning

Research on the Importable of Machine Learning Technology in Japanese Vegetable Distribution Field

（日本の野菜流通現場への機械学習技術の導入に関する研究）

Li Ning

Contents

Contents 2

Contents 3

1. General Introduction

1.1. Practice of Big Data Analysis in Agriculture

In the 21st century, human beings have achieved unprecedented development in network information technology and become the most dynamic and influential high-tech industry in the world. The intensive integration of various communication technologies and computer technologies has changed all aspects of human society, and the technological progress and impact brought by them are unparalleled.

In recent years, new concepts such as mobile Internet, intelligent sensor network, Internet of Things and cloud computing have emerged in an endless stream. New technologies are developing with each passing day, data information and knowledge are exploding, and the era of big data is quietly coming.

In the history of human development, the agricultural industry based on the development of land resources provides a guarantee for human survival; it also developed industry based on the mineral resources of the earth, significantly extending the physical ability to transform the world, improving the survival ability and expanding the living space. With the increasing dependence of social development on data and information resources, the creative industry is known as the "new mineral source" is emerging. This is an emerging industry that takes advantage of information and mental resources outside the physical world. By making full use of these resources, it will greatly expand human wisdom and greatly improve human's ability to spread knowledge, cognition and transform the world. The integration of information technology and network communication technology has greatly promoted the rapid rise of the Internet, mobile Internet, Internet of things and cloud computing, as well as the rapid popularisation and wide application of various intelligent mobile terminals. The growth of data and information obtained by human society has reached an unprecedented speed. The Internet of things, cloud computing, and other technologies are continuously popularised and extended, becoming an important source of data and information. The application of intelligent systems in various fields will surely lead to the explosive growth of data: data is growing faster and faster, with larger and larger capacity, more and more complex data structure, more and more redundant data, and more and more difficult data processing and application. It is with this rapid development trend that a new concept comes into being -- "big data" which is hailed as another wave of information technology by the industry and is bringing a new direction to technological progress and social development. Some media and experts even say that "big data" is another disruptive technological change in IT industry after cloud computing and Internet of things. IT will have a huge impact on social management, development prediction, decision-making of enterprises and departments, and even all aspects of society. On March 22, 2012, the Obama administration announced the investment of 200 million us dollars to promote the development of big data-related industries, upgraded the "big data strategy" to a national strategy, and called big data the "new oil" of the future. From the spring tide of big data and keen competition, people increasingly feel that the information technology has been from the digital and network era, gradually forward to the era of big data and intelligent, the new wave of IT industrialisation is pregnant with a new round of major technological breakthroughs in information technology.

1.1.1. Big data overview

1.1.1.1. The primary connotation of big data

The concept of "big data" gradually became popular in the Internet information technology industry and attracted the attention of the public and media around 2009. The emergence of the concept of big data is indeed accompanied by the explosive growth of information in recent years, and the understanding and understanding of it are also deepening. In particular, with the emergence of new generation technologies such as "mobile Internet", "Internet of Things" and "cloud computing" in recent years, the understanding of "big data" has been constantly improved.

In the 1990s, the concept of "big data" was a catchy term for collections of data that could not be captured, managed, or processed by traditional methods for a given period.

In September 2008, Nature further proposed the concept of big data. McKinsey, a world-renowned consulting company, released a detailed report on big data in June 2011 -- "big data: a guide to future innovation, competition and productivity" which conducted a detailed analysis on the impact of big data, key technologies and application fields. In March 2013, IBM released the white paper "analysis: the application of big data in the real world" in Beijing, which further analyzed and defined the latest "4V" theory of big data, namely the four basic characteristics: Volume, Variety, Velocity and Veracity. There is another very similar 4V theory that follows the same 3V definition as IBM, only turning item 4 into value. If the two are combined, the so-called "5V" theory of big data can also be formed. In addition, some well-known domestic media put forward big data presents the characteristics of "4 V + 1 c", in addition to the "volume", "variety" and "velocity" gives some new definitions, and puts forward 4 V - "vitality", namely data dynamically continuous regularly updated and expanded, fast response speed, strong timeliness, has the vigour of widely used, etc.; The description of complexity of 1C mainly means that in view of the complexity of the data structure, new technical methods are needed to meet the requirements of unified access and real-time data processing of heterogeneous data.

To sum up, there is no unified definition of the concept of big data at present, and the expression methods are not the same, but the expression of core and essential content is the same. Big data, in particular, refers to the kind of data sets that are extremely difficult to store, process and mine because of their massive data capacity and primarily because of their diverse forms and visible unstructured characteristics. Some data show that more than 85% of the data in practical application belongs to unstructured data. For many years, people are familiar with the analysis and application based on structured data; the relational database has dominated the IT application. Typical semi-structured data includes E-mail, word-processing files, and much news published on the web, based on content. Unstructured data widely exists in social networks (SNS), Internet, Internet of things, e-commerce, such as office documents, text, XML, HTML, various reports, pictures, images and audio/video information. The acquisition, storage, mining and processing of the above unstructured or semi-structured data is a significant challenge faced by big data.

1.1.1.2. Big data brings a shift in thinking

With the emergence of new technologies such as SNS, mobile computing and sensor networks, people's way of thinking will change significantly in the era of big data. Research and analysis of a phenomenon will rely on the use of all data rather than sampling data; and don't have to be precise about the data, but it is valuable to adapt to the diversity and richness of the data and to accept even the wrong data. The key to the great charm of applying big data technology is to make it possible for people to extract insights from the fragmented and seemingly redundant and disorganized and irrelevant massive data slag (garbage) and generate great wisdom. With the development of big data technology, people's way of thinking and working is bound to change significantly. Through the analysis of big data, people are more likely to obtain or pay more attention to the result (what is) produced by the event rather than the cause (why). Correlation analysis of data is better than the exploration of causality. Of course, sufficient understanding of correlation will inevitably promote the cognition of causality. Big data inherits some characteristics of statistical science, but it is different from traditional logical reasoning research. Instead, it conducts a statistical search, comparison, clustering and classification analysis and induction on a large number of data. Perhaps it is because of the "inadequate" statistical methods for big data processing that data mining and big data processing technology have been applied more widely in the commercial field. Generally speaking, enterprises do not need to follow the research idea of "from data to information to knowledge and wisdom" when collecting and processing big data, but take the shortcut of "from data to value directly". Big data applications the most typical case is the forecast of "flu" of Google company, based on their every day from all over the world more than 30 one hundred million search instructions set up a system, the system before the 2009 swine flu outbreak began in various areas of the United States successfully for forecast of "flu" and "Google flu trends" service. Alibaba also has a large amount of data. It applies big data technology and integrates the "cornerstone" of all e-commerce models under Alibaba to form a big data platform and become a model of big data business application.

This feature of big data is definitely more suitable for the application in agriculture: The agricultural production cycle is long, and the impact factors are involved, so it may be challenging to understand the causal relationship, but through the big data technology, relevant information can be obtained to guide how to implement according to the needs, so as to ensure the average production and development of agriculture. It is expected that agricultural big data will play a more significant role in the future. Based on local meteorological information, crop and soil information, management information, market circulation and consumption information over the years, through data statistics, case comparison and pattern discrimination, more intelligent agricultural services will be provided.

1.1.2. The magnitude analysis of big data

The network information age, data is surging at an alarming rate, in the end how fast the growth rate, from some data analysis, can be a glimpse. From the dawn of civilization until 2003, a total of 5 exabytes of data were created. However, now people can create the same amount of data in just two days (note: 1 024KB = 1MB, and so on: GB, TB, PB, EB, ZB, YB, DB, NB...). At present, the accepted standard for the scale of big data is that the size of a single data set is at least between dozens of terabytes and several petabytes. According to research by IDC, the volume of data produced worldwide in 2011 was as high as 1. 82ZB, equivalent to more than 200GB of data per person per year. IBM's research shows that 90 per cent of all the data available to the entire human civilisation was generated in the last two years. With the advance of time and the development of science and technology, as well as the rise of the Internet of things, mobile Internet and SNS, the amount of data generated every year grows exponentially. By 2020, the data generated worldwide will reach 44 times the size today.

The explosive growth of data volume has also brought about a revolution in the way of data storage. In 2000, the digital storage of information accounted for only a quarter of the global data volume, but in 2007, only 7% of all data information was stored in newspapers, books, pictures and other media, and the rest was all digital data. The cost of storage today is about 1% of what it was ten years ago. Low-cost and high-performance storage devices provide a basic guarantee for big data storage.

On the one hand, the main reasons for the emergence of big data are the substantial improvement of human's ability to obtain information, especially that of automatic data acquisition. On the other hand, the scope of human cognitive activities and exploration is expanding. A significant feature of the information age is automation, including the automation of data generation and processing, which frees humans from simple, tedious and repetitive tasks to solve problems requiring innovation. In today's all kinds of agricultural monitoring system, for example, as long as the installation of various kinds of sensors at the scene (including images and video), can be automatically, high frequency, and 24-hour non-stop collect plant growth environment of the information such as temperature, humidity and plant diseases and insect pests, the precision of monitoring and management for the growth of plants, data volume growth with astonishing speed.

The above is just the tip of the iceberg of big data. In addition to the improvement of automation technology, big data also covers the needs of many fields and industries: (1) data generated by scientific research, including astronomy, biology, high energy physics and many other disciplines; (2) Smart management: smart government, smart city, smart transportation and smart agriculture; (3) Mobile communications and Internet operators or based on the provision of a variety of services; (4) Internet enterprises.

1.1.3. The key technology of big data

In order to obtain useful information from data, discover knowledge and make use of it, make predictions and guide decisions, it is necessary to conduct in-depth analysis of data rather than generate simple reports. These complex analyses must rely on complex analysis models and tools. After nearly 40 years of development, relational database technology has become a mature and still evolving mainstream data management and analysis technology. Mainstream technologies for relational data management include OLTP and OLAP applications, as well as data warehousing. The OLTP is on - line transaction processing, OLAP- on-line analytical processing. SQL language as a language to access relational database system has been standardised, after continuous expansion, its function and expression ability have been enhanced. However, in the era of big data, in the face of massive semi-structured and unstructured data, relational data management technology loses its advantages, the main reason is that the scalability of relational data management system (parallel database) has encountered unprecedented obstacles and bottlenecks, and it is not up to the requirements of big data analysis, and it is difficult to rely on simple technology for analysis and knowledge extraction. The objects of relational data management are mainly standardised structured data, that is, the raw data that can be logically expressed and implemented with two-dimensional table structure and stored in the database. At present, most mature databases are structured designs. However, as shown above, in practical application, unstructured data gradually rise to the dominant position, including all formats of office documents, texts, pictures, XML, HTML, various reports, images and audio/video information, etc., the challenges are inevitable.

The core technology of big data is store-based computing. In essence, big data mainly solves the problems of mass data collection, storage, computing, mining, presentation and application. It can also be summarized merely into three levels: cloud storage of big data (virtualisation of computing resources), significant data processing (cloud computing model) and big data mining (construction of various algorithm libraries and model libraries). From the perspective of service for users, should also provide more application functions: (1) visual interactive analysis engine, to provide heuristic, human-computer interaction, visual data mining new technology, mass data mining highly interactive functions; (2) the establishment of workflow engine, for the user to create mass data processing, analysis process to provide graphical process design tools, automatically execute the user created data processing analysis process, to provide resource scheduling and optimization services; (3) provide OPEN API function, provide data mining platform and third-party application system extension interface.

At present, there are various models of big data solutions in the world. In the application field and profit model, Google, Amazon, Microsoft and VMware have launched their own big data solutions at different times. Hadoop is an open source implementation of Google big data platform, which is very typical in significant data processing and application. Hadoop is a distributed system platform that makes it easy to build an efficient and high-quality distribution system, and it also has many other related sub-items, which are a great expansion of its functions, such as Zookeeper, Hive, HBase and so on. Because of its open-source nature, more and more applications are adopting Hadoop to their needs. Hadoop distributed platform mainly includes two points; one is pointing cloth type documents system (HDFS - distributed file system), run on commodity hardware, two its is distributed computing model - graphs, both be short of one cannot. Also, HBase (Hadoop database) can be applied to establish a high-performance, highly reliable and scalable column-oriented distributed storage system on Hadoop, and make it possible to set up large-scale modular structure storage set group on cheap PC Server. MapReduce is one of the core components of Hadoop. It was first proposed by Google in 2004. Its parallel processing model for big data has the advantages of good scalability and high robustness. If parallel databases belonging to relational databases have the characteristics of high query efficiency and are the crystallisation of database development, then MapReduce solves the problem of big data analysis platform's scalability and fault tolerance at the system level and is a typical representative of non-relational databases. It is suitable for unstructured data processing, large-scale parallel processing, easy to use and other prominent advantages, in the Internet information search and other fields of big data analysis have made significant progress, has become the mainstream technology of big data analysis. Therefore, more and more researchers improve MapReduce from the aspects of performance and ease of use and indeed solve various technical problems of big data effectively through the fusion of MapReduce technology and parallel database.

Based on the above key technologies, a series of application or business service models can be generated. For the field, have to adapt to their own needs of the application. In this sense, big data technology is only meaningful for the solution of specific fields and problems, and there can be no unified solution. However, in general, the core technologies involved, such as data transmission, storage, computing, mining, presentation and development of application platforms, have universal consistency. The advent of big data era brings new challenges to data storage, processing and analysis, but the general trend is to solve the "bottleneck" problem through distributed computing. Distributed computing means that in order to improve the overall performance of the system, it cannot rely on the vertical expansion to improve the performance of a single node, but can achieve the purpose of calculation and analysis by increasing the number of nodes in the system through the horizontal expansion, which means that cheap equipment can participate in a variety of computing tasks of big data. The core of the solution is to distribute the tasks of storage, processing and analysis to each node in the system in a distributed way, to accelerate the speed of data storage, processing and analysis.

1.1.4. Agricultural big data

1.1.4.1. The connotation of agricultural big data

Agriculture is the endless source of big data and the broad field of big data application. Agricultural data cover a wide range of data sources complex. As the name suggests, big data in agriculture refers to the use of big data concepts, technologies and methods to solve a series of problems in the collection, storage, calculation and application of agricultural or agriculture-related data. It is the application and practice of big data theory and technology in agriculture. Agricultural big data is the professional application of big data theory and technology. In addition to the public attribute of big data, agricultural big data must have its characteristics. Generally speaking, agriculture should cover rural areas, agriculture and farmers. It has the characteristics of covering a wide range of regions, covering a wide range of fields and contents, influencing factors, sophisticated data collection and difficult decision-making and management. Unique agricultural production means crop-plantation, including the production of food crops, cash crops and forage crops and green manure crops production activities, not only involves the cultivated land, planting, fertilizer, pesticide, harvest, storage, and breeding crops every link of the whole process of production, but also involves the cross-trade and cross major, across business data analysis and mining, as well as the result of demonstration and application, and the whole industry chain of resource, environment, process, safety monitoring and management decision-making, etc. In a broad sense, agricultural production refers to five modern forms including planting industry, forestry, animal husbandry, fishery and sideline, all of which should be included in the category of agricultural significant data research.

With the rapid development of precision agriculture, smart agriculture, Internet of things and cloud computing, agricultural data also presents an explosive increase, and data from storage to mining applications are facing enormous challenges. The penetration of the Internet of things in various fields of agriculture has become an inevitable trend in the development of agricultural information technology and the most critical data source of agricultural big data. A large number of agricultural workers and managers are both users and producers of big data. Due to the complexity and particularity of agriculture, agricultural data will be transformed from relational data type based on the structure to semi-structured data type and unstructured data type. Compared with the relational data structure logically expressed by two-dimensional tables, there are more unstructured data in the field of agriculture, such as a large number of hypermedia elements in the form of text, charts, pictures, animation, voice/video, as well as expert experience and knowledge, agricultural models and so on. A large number of facts have proved that unstructured data show a momentum of rapid growth, the amount of which has dramatically exceeded structured data. In particular, the main body of the agricultural production process is biology, which is susceptible to external environment and human management and other factors. There are diversity and variability, individual and group differences, which all determine the difficulty of data collection, mining and analysis application. How to excavate data value, improve data analysis and application ability, reduce data redundancy and data garbage is an important subject facing agricultural big data.

1.1.4.2. Significant applications of big data in agriculture

Based on the analysis of the main application fields of agricultural information technology and the primary sources of big data, the main application fields of big data include the following aspects:

a) Production process management data: facility planting industry, facility breeding industry (livestock, poultry and aquatic products, etc.), precision agriculture, etc. It is an urgent task of agricultural informatisation to improve the precision monitoring, intelligent decision-making, scientific management and regulation of the whole production process.
b) Agricultural resource management data: land resources, water resources, agricultural biological resources, means of production, etc. China's agricultural resources are in short supply, and the ecological environment and biological diversity are degenerate. By understanding the family background, we should further optimise the allocation and rational development to achieve sustainable development of high yield, high quality, energy saving and high efficiency in agriculture.
c) Agricultural ecological environment management data: soil, atmosphere, water quality, meteorology, pollution, disasters, etc. Comprehensive monitoring and accurate management are needed.
d) Big data of agricultural products and food safety management: origin environment, industrial chain management, pre-production, in-production and post-production, storage and processing, market circulation field, logistics, supply chain and traceability system, etc.
e) Agricultural equipment and facilities monitoring big data: equipment and

implementation condition monitoring, remote diagnosis, service scheduling, etc.

f) Big data generated by various scientific research activities, such as a large number of remote sensing data, including space and ground data; A large number of experimental biological data, such as gene mapping, large-scale sequencing, agricultural genome data, macromolecules and drug design.

1.1.5. The main task of agricultural big data

Based on the theory and technique of the extensive data, to push the innovation of agricultural big data technology and application practice, combined with the national agricultural modernization and agricultural informatization development strategy, break through some key technologies of agriculture big data, plotting and concise and application of a batch of sizeable agricultural data demonstration projects, to enhance data and the status of the Internet of things and cloud computing are equally important, employed the leading technology of the whole big data, a new era of information, promote agricultural development wisdom.

Under the condition of the market economy, decentralised operation and production mode of agriculture make it more critical than ever to rely on information in participating in market competition: the lag of information and services often has a substantial negative impact on the whole industry chain. Due to the characteristics of the market economy, it is difficult to form a unified plan for agricultural production nationwide, so agricultural production is significantly affected by market fluctuations. Moreover, agricultural production relies on feelings and experience in many aspects and lacks quantitative data support. In the era of big data, we can not only establish a comprehensive data platform to regulate and control agricultural production but also record and analyse the dynamic changes in agricultural planting and breeding process and circulation process. By analysing data, we can formulate a series of regulation and management measures to make agriculture develop efficiently and orderly.

As for the definition of smart agriculture, there is no accepted definition and unified standard for it because of its complex contents. However, based on comprehensive data and analysis, the core content of "smart agriculture" can be summarised as the following aspects:

As for the definition of smart agriculture, there is no accepted definition and unified standard for it because of its complex contents. However, based on comprehensive data and analysis, the core content of "smart agriculture" can be summarized as the following aspects: (1) scope of coverage: the whole agriculture is complete rather than partial, and the whole process rather than stage; (2) technical means: it is mainly based on various advanced information and network technologies, including perception, transmission and intelligent processing technologies, to achieve the main production model of automated production, optimal control, digital and network services, intelligent decision-making management; (3) overall goal: to achieve the sustainable development of high efficiency, high quality, energy saving and environmental protection. "Smart agriculture" is a vast system, or a top-level virtual concept. Its specific implementation is composed of subsystems in various fields, such as intelligent facility agriculture, original agricultural conditions, intelligent plant protection, intelligent irrigation, intelligent market management and other subsystems. With the deepening of smart agriculture, advanced technologies such as the Internet of things, cloud computing, big data and the semantic network will be applied more widely. The key to smart agriculture lies in fully integrating knowledge base and model base of various fields, applying reasoning, analysis and other mechanisms to make predictions and provide intelligent control and decision management. Current agricultural research and application of the wisdom, the true wisdom is still far, but every journey begins with one step, in facility agriculture, for example, through automation network monitoring system, real-time acquisition greenhouses in various environmental elements parameters (such as air temperature and humidity, soil temperature and humidity, CO 2, illumination, dew point temperature and other environmental parameters), automatic open or close the specified device (such as irrigation, shading, ventilation, heating and cooling equipment etc.); According to user demand, it can provide not only automatic monitoring means for comprehensive information on facility agriculture but also provide scientific basis for automatic control and intelligent management of environment. By monitoring the field crops and environmental information, various disaster information, we can better judge the appropriate time for irrigation and fertilization, timely release early warning information, and take effective disaster prevention and reduction measures. Although these are the primary applications of smart agriculture, they have more or less influenced the change of agricultural production mode.

1.2. The practice of Machine Learning in Agriculture

1.2.1. Technical background and application significance

In recent years, agricultural IOT has attracted considerable attention worldwide. In the report of the governments of many countries, it is pointed out that we should vigorously promote the "Internet +" modern agriculture, and apply the Internet of things, cloud computing, big data, mobile Internet and other modern information technologies to promote the transformation and upgrading of the whole agricultural industry chain.

The development of the agricultural Internet of things involves many agricultural science and technology fields and policies of relevant units and departments. Led by the government, it can accelerate the development of agricultural science and technology, accelerate the upgrading of agricultural industry infrastructure and decision-making management platform, build a comprehensive logistics management platform to ensure the intelligent scheduling of agricultural market, and build an entire agricultural industry chain. The development that affects relevant industry by agriculture adjusts industrial structure more reasonable.

Some developed countries, represented by Europe and America, have made relatively early studies on agricultural IoT and environmental monitoring, mainly establishing environmental monitoring systems, meteorological and pest monitoring and early warning systems, and performance control systems. Crops can be appropriately used for medicine, irrigation and fertilisation to improve resource utilisation rate and reduce environmental pollution. Concerning performance control, it is reflected in the rational and coordinated allocation of agricultural resources, the constraint of energy consumption and yield improvement.

Machine learning has many advantages over traditional solutions, such as low cost, long-term monitoring, scalability, accuracy, and ease of deployment. Therefore, in the agricultural Internet of things, a large number of sensor nodes of various types can be arranged near crops to form a systematic and comprehensive monitoring network, to help farmers find problems in time and propose corresponding solutions. Thus, an integrated production mode based on intelligent decision-making of agricultural IOT system is formed.

The environmental data collected based on machine learning techniques are non-linear and non-stationary, which can be regarded as a complex of different scales and time, which is conducive to the establishment of soil-vegetation-atmosphere model and physical model, as well as the improvement of the expert system design in precision agriculture. Among the traditional data processing methods, the more popular processing methods include Fourier transform and wavelet transform, etc., but they have strict requirements on the essential characteristics of data and cannot process the data with nonlinear and non-stationary characteristics. However, the data of the characteristics of the growth environment of farmland crops collected based on WSN nodes are non-linear and non-stationary data, so the traditional data processing and analysis methods are no longer applicable to the data in this paper. At this time, a new method should be sought for its analysis.

With the development of the agricultural Internet of things, the decision-making system still needs to establish a more comprehensive data analysis method to explore the profound value of data and elevate it to the level of intelligent decision-making. Machine learning algorithms are used to monitor surface temperature, soil moisture, ambient light and temperature to help analyse current plant nurseries in the country. Therefore, the interrelation of various sensor data collected in the agricultural Internet of things decision-making system, especially the time-space correlation analysis based on different scales, cannot only enrich the data processing and analysis methods in the decision-making system but also has great significance, mainly reflected in the following aspects:

a) Firstly, it quantitatively reflects the relationship between the growing environment of crops. In the growth cycle of crops, it is necessary to collect various types of environmental information (temperature, soil moisture, sunshine, precipitation, etc.) to reflect the growth of crops, and all elements are of great significance. Soil moisture content can directly affect the growth, maturity time and yield of vegetation. Temperature and sunshine have direct or indirect effects on soil moisture to a certain extent, resulting in the decline of soil moisture. However, the environmental elements based on time series are all non-linear and non-stationary data, and they are often in a fuzzy influence relationship. After the feature extraction of each environmental element, the component is stable and has practical physical significance. Therefore, the quantitative response of the time-space correlation performance of each element to its correlation is analysed based on different scales.
b) Secondly, it is helpful for more reasonable performance control of agricultural IOT. In general, the collection frequency and total working time are determined based on the coverage of crops, project requirements and work experience. In the time-space correlation of various environmental factors, the redundancy between the collected data information is reflected. Thus the node layout and acquisition mode are changed, the performance control of agricultural Internet of things is optimized, and the cost of human and material resources is reduced to achieve the reasonable allocation of agricultural resources.
c) Finally, it is conducive to the establishment of a more accurate agricultural IOT decision-making system. Through correct judgment and analysis of the data, relevant information is extracted from the data based on various knowledge, and then a perfect decision-making system is established. That is to say; the decision-making is based on the data integrity analysis report. In the agricultural IOT decision-making system, the judgment of the decision-making system can be improved, and more accurate and useful decisions can be made

for crops by analysing the time-space correlation of the collected environmental variables of crops and combining with the changes of meteorology in the future.

1.2.2. Research history and current situation at home and abroad

B.L. Barnhart et al. used the EMD method to decompose the fluctuating total solar radiation, global temperature, sunspot number and carbon dioxide concentration into the components of different periods, analyzed and compared the components of each period from the time domain, and empirically estimated the relative radiation effect in the different components of solar radiation and carbon dioxide concentration. Patrick Flandrin et al. conducted multiple experiments by randomly adding part of Gaussian noise to the data and obtained that EMD has the characteristics of filtering signal noise. Abdel-ouahab Boundraa et al. decomposed the signal containing noise and reconstructed the relevant components again, to achieve the purpose of filtering noise. In practical application, HHT has the effect of abnormal detection. Dejie Yu applied HHT to the vibration signal of the roller bearing to obtain the energy spectrum, to analyse and diagnose the fault mode of the roller bearing.

On the analysis of meteorological data of time-space correlation, Asim Biswas et al., respectively, using discrete wavelet transform (discrete 'wavelet transform) and the EMD of soil water content, soil water storage) of geographical space is analyzed from different scales, and analyze the contribution of each component of rainfall on the soil, with the actual environment variable time correlation analysis, to find the essential impact factors of the actual environment. Ouyang et al. studied the northwest Ohio PAR (photosynthetically active radiation) and the relationship of the soil surface temperature, showed that because both have shared an annual cycle of day and night and present a high degree of correlation, but there are few integrity studies of PAR in the same region, soil surface temperature, the time-spatial correlation between soil moisture were analyzed.

In the model prediction of intelligent irrigation decision subsystem, the precipitation prediction provides an essential basis for crop production management, engineering construction, flood management, disaster prediction and other processes. Due to the diversity, variability and complexity of the influencing factors of precipitation, it is often difficult to establish a model and forecast based on multiple influencing factors.

Machine learning algorithm can process complex data through different models so that its prediction or classification can achieve high accuracy. W.O.Ochola developed a Muscoff model based on daily precipitation data from two weather stations in Kenya's kano plain for extreme weather prediction. C.L. Wu pre-modeled and compared the classification after data preprocessing. The model comparison methods were an artificial neural network, k-nearest-neighbours algorithm, linear regression and a hybrid model with three methods. The results showed that the effect of singular spectrum analysis as data preprocessing was better, and the effect was significantly improved in predicting daily precipitation. Tiesong Hu et al. used principal component analysis (PCA) to analyze and process the precipitation, and then built a neural network model based on the processed data to predict the runoff of Da-Rong river.

According to the characteristics of the data, the prediction efficiency of the machine learning algorithm will be significantly improved by appropriate conversion. Manning Hu et al. collected the wind speed of three farms in northwest China and used the mixed method of ARIMA model, support vector machine, empirical mode decomposition and support vector machine to predict the wind speed, and the results showed that the mixed model method had the best prediction effect. Shengzhi Huang et al. combined the empirical mode decomposition with the support vector machine in the monthly runoff prediction of slide county, Lintong and Xianyang, and the results showed that the combined model had a more stable and more accurate prediction effect.

Kamilaris and Kartakoullis et al. (2017) and Kamilaris and Francesc et al. (2018) performed reviews of agricultural big data and analytical techniques1), and confirm that although in contrast to the rather profuse bibliography on big data, the agricultural subset appears to be relatively recent and limited, but the number of studies has a tendency to increase rapidly. They indicated that the reason for the increase is when researchers experiment with agricultural big data analysis, applying it to resolve various agricultural problems involving classification or prediction, the complex, multivariate and unpredictable agricultural ecosystems can be better understood.

Furthermore, the results of their study confirm that 79% of papers solve problems like land degradation and water contamination, climate change from a technical perspective. Additionally, the rest of the papers solve problems like governmental policies, market fluctuations and sociocultural development (e.g. dietary preference of meat protein), they speculated that a critical reason of the lack of studies in these areas is that compared with climate and geographic data, social behaviour data are less open and complete (especially business data), which makes the research progress lag behind.

To address these governmental policies, market fluctuations issues studies, researchers in Japan have found an opportunity.

In the construction of Japan's agricultural product distribution system for many years, Japanese farmers set up group, companies or join Japan Agricultural Cooperatives (JA) for mass production. Moreover, products are traded in wholesale markets which administered by the Wholesale Market Act. After years of standardisation, these links of the distribution system have kept a large number of complete transaction data. Combined with the macro data collected by Ministry of Agriculture, Forestry and Fisheries (MAFF), can provide high-quality data sets for agriculture big data to solve problems such as government policies and market fluctuations in the field of domestic agricultural products distribution.

1.3. The basic theory of machine learning

In 1997, Tom m. Mitchell defined Machine Learning in his book Machine Learning as "the behaviour of computers to improve the performance of their systems by using experience". Machine learning can be roughly divided into two categories: supervised learning and unsupervised learning. Supervised learning refers to a model that can recognise and judge the future from labelled training data, and each sample contains an input object and an output object. It is mainly used for regression and classification. The commonly used models include a support vector machine, artificial neural network and decision tree. Data for unsupervised learning and training are not labelled, but are learned according to the characteristics of data to achieve specific purposes. They are mainly used for clustering, mainly including hierarchical clustering, k-means clustering and other models.

1.3.1. Linear regression

Linear regression is a relatively simple statistical learning method, which assumes a linear relationship between dependent variables and independent variables and can reflect the influence weight of each variable from the relationship, to identify important characteristic variables. The mathematical expression of multiple linear regression is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Going to simplify it as a vector

$$y = X\beta + \varepsilon$$

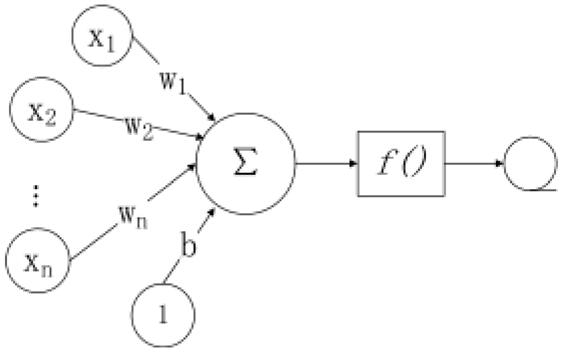
Where Y is the dependent variable, X is the independent variable, β is the partial regression coefficient, namely the weight coefficient of the independent variable, ε is the random error except the influence of n variables on Y.

After establishing the relationship expression, linear regression can not only know the influence weight of each characteristic variable on the target value but also predict the new data through the expression. It belongs to the simple model in machine learning, that is, establish the model after learning the data, and then predict the target variable according to the new data characteristics.

1.3.2. Artificial neural network theory

Artificial Neural Network (ANN) abstracts the human brain's response to sensory input information, connects a large number of nodes (neurons), and models the relationship between input information and output information to solve the problem by simulating the brain. Artificial neural network mainly includes artificial neuron, activation function, network topology and training algorithm. The activation function defines the method of information transmission, the network topology defines the overall framework of the network, and the training algorithm provides the learning method of the neural network, which is introduced below.

1.3.2.1. Artificial Neuron



Single neural network model is similar to biological models, from the input unit receives multiple signals, and the weight of those signals are multiplied by the corresponding additive after (w), and then according to the unified $f()$ function get output value y(x) processing, including w weight to the importance of the signal, said $f()$ function as the activation function, the mathematical expression for

$$y(x) = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

The activation function is a process of signal processing and filtering, which converts multiple input signals of neurons into a single output and transmits the signals in the whole network in this way. It simulates the processing mode of biological neurons. When the signal reaches a specific condition after processing, the signal is transmitted to the next layer. Otherwise, it prevents the signal from being transmitted downward, which is called the threshold activation function. There are many forms of activation functions, which are briefly described below:

a) The unit jump activation function

The unit jump activation function is a relatively simple function, which can be used for a simple component model. When the sum of the input signals is greater than or equal to zero, the output is 1; otherwise, the output is 0. The mathematical expression is:

$$f(x) = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases}$$

b) S type activation function

The shape of the s-type activation function is similar to that of S, and the range of its output value is (0,1), and the expression is:

$$f(x) = \frac{1}{1 + e^{-x}}$$

From the expression, we can know that this function is differentiable, no matter the input is any real number, the output is always in a fixed interval, which is the most commonly used activation function.

c) Linear activation function

The model generated by the linear activation function is very simple, and the expression is:

$$f(x) = x$$

In addition to the above-mentioned activation functions, there are also linear activation function, full linear activation function, hyperbolic tangential activation function, Gaussian activation function and other functions, among which the radial basis network model generated by the Gaussian activation function is the most famous.

1.3.3. Network topology

The reason why the neural network can solve the complex nonlinear problems lies in the diversity of its network topology structure, that is, the productive structure of neurons interconnection, which mainly includes three essential parts:

a) Number of hidden layers

Generally, the structure of the artificial neural network model is divided into three layers. The first layer is the input layer, which is mainly used for the input of data information. The second layer is the hidden layer, which can process and filter the data. The third layer is the output layer, which outputs the final result. Among them, there is a fully connected structure between each layer, and there is no connection between nodes of the same layer, and the number of hidden layers is variable. It is precise because the number of hidden layers can be adjusted that the structure model of different layers can be constructed to deal with the relationship of more complex data.

b) Number of nodes per layer

The number of input layer nodes is determined by the number of essential characteristics of input data, while the number of output nodes is determined by the result of modelling requirements, and the number of hidden layer nodes needs to be set. The number of neurons in the hidden layer is also directly related to the complexity of the model, so the number of nodes is also a direct reason for the effect of the model.

c) The direction of information transmission

From signal input to signal output, signals are transmitted from one layer to another in a fixed order and direction. All signals are transmitted in the order from front to back. Such a network structure is called feed-forward network. A loop structure is added into the feedforward network structure, which allows forward transmission of partial hierarchical signals, thus making the signals recursive in the network, which is called the recursive network.

1.3.4. Training methods

At present, after the classic method of training artificial neural network is to spread (backpropagation) algorithm, which includes two phases: first, back to the stage to the stage, through the two-phase loop iteration for many times, and each iteration is called a new era (epoch), finally to achieve the set a stop criterion, for example, a new era to the expected number, the prediction error is less than a specific value criterion, etc.

1.3.4.1. The forward propagation

The forward propagation, merely speaking, refers to the feed-forward network, the process of data from the input layer to the output layer, in which each neuron is weighted and activated and propagates layer by layer to finally produce the output.

1.3.4.2. Backpropagation

The output signal in the forward propagation is compared with the actual one, to calculate the error value, and then the error is back propagated, and the weight between neurons is adjusted in each layer. Finally the total error of the model is minimised. The method of weight adjustment is a gradient descent algorithm to find the place where the error declines the fastest. The error back-propagation algorithm is shown from the perspective of data as follows:

$$f(x) = f(\sum_{j=1}^{n} w_j f(x_j) - b)$$

$$E = \frac{1}{2}(f(x) - y)^2$$

$$\nabla w_j = \frac{\partial E}{\partial w_j} = \frac{\partial E}{\partial f(x)} \frac{\partial f(x)}{\partial w_j} = (f(x) - y) f'(x) f(x_j)$$

$$w_j = w_j - \eta \nabla w_j$$

$$\nabla w_{i,j} = \frac{\partial E}{\partial w_{i,j}} = \frac{\partial E}{\partial f(x)} \frac{\partial f(x)}{\partial f(x_j)} \frac{\partial f(x_j)}{\partial w_{i,j}} = (f(x) - y) f'(x) w_j f'(x_j) d_i$$

$$w_{i,j} = w_{i,j} - \eta \nabla w_{i,j}$$

There are two modes of weight updating. The first mode is to update the weight after each instance is processed, but the error calculated by this method is not necessarily reduced in real time, but related to the characteristics of the instance. Another mode is to update the weight after the instance processing of specific data, which is called batch learning. At the same time, the size of instance batch is also related to the error of final optimisation.

In the process of training, due to the slow convergence speed of the error back-propagation algorithm, it is easy to fall into local extreme values and other problems. Some improvements can be made by the original algorithm. According to certain specific indicators to stop training, this data stop method; In some methods, some penalty factors are added into the error function to adjust the updating of weights, such as L1 regularisation and L2 regularisation.

This chapter introduces the basic theory of machine learning. Linear regression is the most common, simple and understandable model in life. Then it introduces the most important essential elements of the artificial neural network, including artificial neurons, network topology and training methods. Then another critical support vector machine in machine learning and support vector regression machine theory are introduced, including its basic definition, formula reasoning process and its introduction of relaxation variables, kernel function and other core ideas to improve performance algorithm.

1.4. Japanese domestic vegetable supply situation

1.4.1. Japanese vegetable import trends

In recent years, Japan's vegetable imports have gradually increased, accounting for about 3 million tons in 2015, accounting for about 21% of domestic demand, about 15 million tons. Regarding the types of imported vegetables, fresh vegetables increased by more than 1 million tons, followed by 800,000 tons of frozen vegetables, totalling less than 1 million tons, among which processed tomato products and other frozen vegetables also increased.

This is the result of consumers' low-price intention, simple diet and other comprehensive factors after the Plaza agreement, under the keynote of strong yen and the deflation trend after the collapse of the bubble economy. From the perspective of import proportion, the proportion of crude grain import is 15%, the demand for household consumption accounts for 2%, and processing and business use account for 26%. The simplification of diet and externalization increase the proportion of import of fast food and processing enterprises.

Japan's domestic vegetable self-sufficiency has been hovering around 25% of the total agricultural output since the 1990s. It is the same major sector as rice farming and livestock production. In 2014, when the rice price dropped, it reached 24.6%, surpassing rice (22.8%). However, the output of vegetables in Japan was influenced by the increase in imported vegetables and the decrease in domestic consumption.

1.4.1.1. Trends of vegetable imports

In 2015, the top import varieties of fresh vegetables and their import volume were 358,000 tons of Onions (32.1%), 122,000 tons of pumpkins (11.0%), 101,000 tons of carrots (9.1%), and 71,000 tons of Onions (6.4%), which accounted for 58.6%. A total of 43 species of imports (species classification by the agricultural and livestock industry revitalisation agency). The second largest importer of pumpkin was New Zealand (69%).

It is particularly pointed out that, the Japanese winter cabbage import rate is at a low level (in 2015, it was less than 5%) (MAFF 2015);

1.4.2. The supply of winter cabbage

Winter cabbage was used as the analysis object, because winter cabbage is one of the designated vegetables in Japan, and its shipment volume ranks the first among leaf ad stem vegetables. The national production of winter cabbage in Japan increased from

1,211,000 tons in 2009 to 1,316,000 tons in 2014, with an obvious tendency to increase.

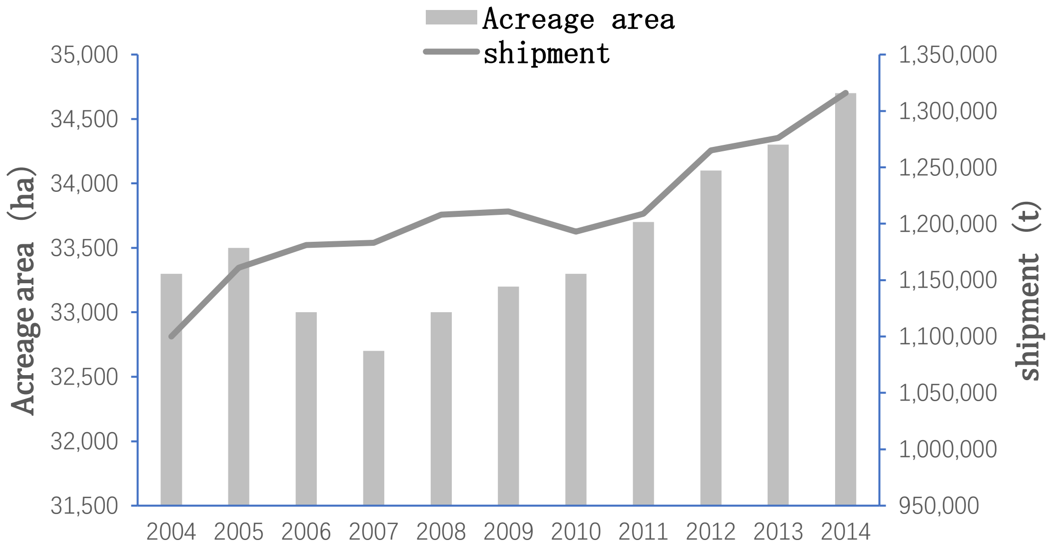


Figure: Planting area and shipment volume of Japanese cabbage annually

Source: Created by author from "Ministry of Agriculture, Forestry and Fisheries" Vegetable production shipment statistics

"

In addition, the annual import volume of winter cabbage in Japan is less than 5%. Meanwhile, the designated origin of domestic winter cabbage is concentrated in Aichi and Gunma prefectures in Japan. Therefore, as a survey variety of domestic vegetables in Japan, the judgment with fewer interference factors can be made.

Table: Comparison of the top three suppliers of fresh vegetables in Japan

Item	Daikon	Cabbage	Onion
Quantity of main production area (production amount> 10%)	3	2	3
Percentage of processing/ business for total distribution volume	60%	50%	59%
Percentage of processing and business for market distribution	20%	7%	15%
Percentage of processing/ business dose brokers	40%	43%	44%
Import ratio	＜5%	＜5%	42%

Source: Ministry of Agriculture, Forestry and Fisheries "Crop survey - vegetable production shipment statistics" H22 fiscal year

Agriculture, Forestry and Fisheries Policy Research Institute "Proportioning ratio for each eye" H22 fiscal year

Ministry of Agriculture, Forestry and Fisheries 'Fruit and vegetable wholesale market survey report' H22 fiscal year

Ministry of Agriculture, Forestry and Fisheries "Transition of import volume (fresh) by current products over processing/business vegetables" H25

Ministry of Agriculture, Forestry and Fisheries "Summary of Imports of Agriculture, Forestry and Fisheries" H.26 Fiscal Year

Created by the author

2. A Case Study on Producer wholesale market

2.1. Introduction

Over the past decade, at the international level, the term "Green Data Revolution" was born from the optimism that big data can and will deliver benefits to the agricultural industries and global society (FAO,2009). By recognizing the potential of revolutionizing Japanese agriculture, the Japanese government and several researchers had invested in this field. Key priorities like precision agriculture, plan breeding and yield prediction incorporate machine learning technology on the basis of big data.

Matsui et al. (2014) developed a factor analysis and a prediction model of abandoned, cultivated land in regions of Japan. Their research showed that the Random Forest (RF) algorithm of machine learning technology can detect major factors in specific regional situations, it is one of the best algorithms to predict cultivation abandonment in Japan. Yamaya and Sonbe et al. (2017) used satellite data and crop classifications to establish mapping methods, in place of existing ground surveys, and assessed the accuracy of classifications performed by the RF algorithm. The results of their study confirm that lower-cost crop classification is possible. RF applied in these studies is an important algorithm of machine learning technology. The main advantage of RF algorithm is that it can investigate nonlinear and hierarchical relationships between the predictors and the response by using an ensemble learning approach. There are many cases wherein the RF has outperformed traditional linear regression and linear discriminant analysis approaches, since using multiple efforts to predict a response can increase the robustness and accuracy of predictions compared to using a single data set or model (Everingham and Inman-Bamber et al.,2015).

Agricultural products distribution is another area that relies on machine learning technology. In a case in Japan, the regional agricultural promotion, which has received widespread attention and has been explored extensively in recent years, has become an important topic related to agricultural products distribution. Many researchers are aware that sustainable improvement in regional agricultural economy requires reforming the distribution system structure. However, the reform is subject to several constraints. In many less adaptable agricultural areas, traditional, decentralized, and complex distribution structures still dominate. In these areas, Producer wholesale markets are established at a distance from each other, to such an extent that each of the markets tend to follow a different business model. Therefore, there is insufficient research on this topic.

Given the advantages offered by the RF algorithm to make early and accurate predictions by using big data, the objective of this study is to fill the gaps in the application of data mining technology in the agricultural distribution of Producer wholesale market, explore a business scheme, and determine how accurately the RF algorithm predicts trading price.

2.2. Materials and methods

2.2.1. Research Subject and Area

After investigation, we decided to choose the winter cabbage as our research subject for domestic vegetable distribution in Japan. The reasoning behind this decision is that the winter cabbage has several important attributes, which are as follows: Firstly, the Japanese winter cabbage import rate is at a low level (in 2015, it was less than 5%) (MAFF 2015); Secondly, the winter cabbage's production is concentrated in only one prefecture (Aichi); Thirdly, there is evidence that a high ratio of winter cabbage (over 50% in 2015) are used in the foodservice industry (MAFF, 2011, 2015).

The cultivation of the winter cabbage first began in the Aichi prefecture, Japan, and the Atsumi area soon became the most important area for the growth of Aichi's agriculture. Through investigating the distribution system of the Atsumi area, the following main conditions, which aided in the growth of the winter cabbage, were observed:

2.2.1.1. Production conditions

Atsumi area is a peninsula in the southern part of the Aichi prefecture, central Honshu, Japan. Due to the influence of the warm current (Kuroshio) flowing offshore, it has a mild climate even in winter, has abundant sunshine, and it is considered to have one of the best climatic and geographical conditions in the nation for agriculture. Especially after the completion of the Toyokawa water project, agriculture developed greatly, and it became one of Japan's representative agricultural zone of vegetables.

2.2.1.2. Distribution conditions

In the process of forming large winter cabbage production areas, Producer wholesale market merchants, who have a horizontal competitive relationship with JA (Japan Agricultural cooperatives) Group in terms of purchasing the agricultural products, had a significant effect on local distribution networks. They show a unique presence in vegetable distribution, commensurate with their agricultural productivity. Currently, 26 Producer wholesale market merchants have set up six Producer wholesale markets as limited companies. Half of the agricultural output in the Atsumi area is traded through these markets.

2.2.1.3. Market conditions

This research focuses on market A, which is the largest of the six Producer wholesale markets and trades more winter cabbages than any other Producer wholesale market in the Atsumi area.

2.2.2. Data Acquisition and Preprocessing

2.2.2.1. Acquisition

After investigating the trade system of Producer wholesale market A, we found that all vegetable products are traded by auction and separate settlement1). In contrast, JA Group's markets use consignment sale and pooling settlement2)3). Consequently, the producers who want to ship their products easily tend to choose JA Group's market but producers who want to sell their products at a higher price tend to choose a Producer wholesale market.

In this study, out of consideration for commercial confidentiality agreement, we were allowed to collect auction data between January 2011 and April 2016 in market A.

2.2.2.2. Preprocessing

Outlier: We use Z-scores to identify outliers. $\{x=Z\_score \mid x \leqslant -3, x \geqslant 3\}$ is considered an outlier. Identified outliers will be listed separately and analyzed in subsequent studies.

Average price: We defined average price in three steps. First, we used sales date, size, and variety data to classify winter cabbage. Second, we calculated the average prices for each category. Third, we subtracted the average price from auction price for each item, and the results were defined as "price deviation".

$$\textbf{Price Deviation} = \textbf{Auction Price} - \textbf{Average Price}$$

In this study, price deviation prediction has following advantages: Firstly, it can reduce the impact of year price fluctuation4); Secondly, it reduces the database dimension and hence, there is no need to separate database by year index; Lastly, it improves the readability of the predicted results and facilitates development of trading scheme.

2.2.2.3. Scheme and RF Classifiers

Based on the price deviation results, we learned that in daily auction of prices in market A, the instances of price deviation over 20% are almost 0. At this point, we carried out a new business model concept for market A in which price deviation would be predicted and classified before auctions using the RF classification model, and then according to the results of the prediction, the products will be divided into 2 parts; the first part will be traded by auction and separate settlement, and the second part will be traded by centralized auction and pool settlement. Thus, it shortens the total time of the auction. Hence, the target of RF classification model is

$\{y = Price\ Deviation | -15 \leq y \leq 15\}$(unit: yen).

After classification, the cabbages that meet classifier target conditions will be traded by centralized auction and pool settlement. The rest will still use auction and separate settlement. The new scheme is depicted in Figure 1.

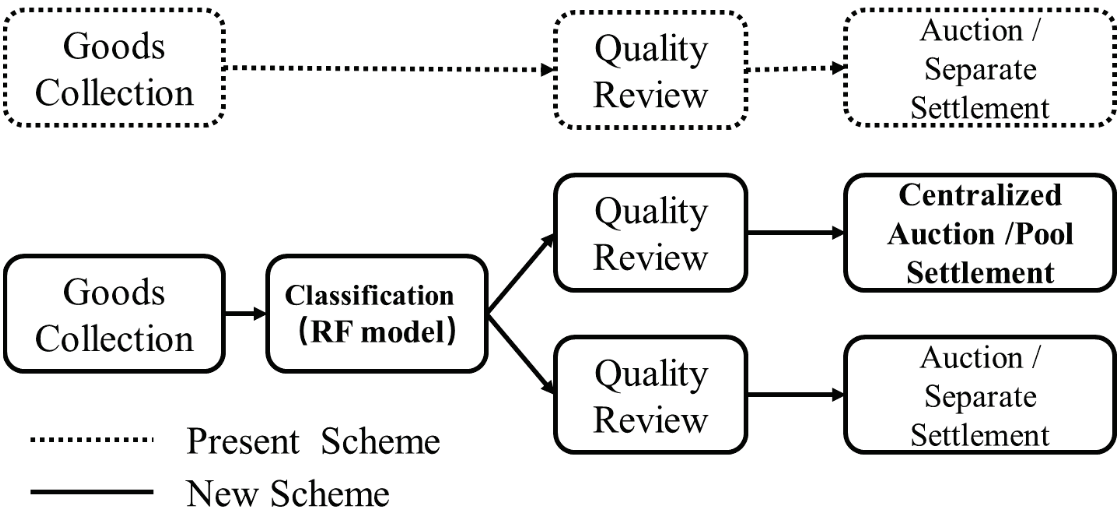


Figure 1 The Flow Chart of Present /New Scheme

Note: **[Centralized Auction/Pool Settlement] --** Products will be auctioned irrespective of producer and settled at same price.

The RF algorithm also can rank the relative importance of each predictor variable. Using the RF algorithm, the mean decrease in accuracy (MDA) for a variable allows it to assess the importance of each variable based on the out-of-bag (OOB) regression prediction error. Higher values of MDA indicate that the variables are more important to the classification (Culter and Edwards et al., 2007; Breiman., 2001; Liaw and Wiener., 2002).

A range of predictor variables that could be related to predictions of price deviation was entered in RF classification model. Details about all the variables can be found in Table 1

Table 1 Variable Details and MDA Results of RF Model

Name	Unit	Measure	MDA	Details
Size	-	Nominal	7.6	The size of winter cabbage: L, M, S,2L,3L,4L,5L,2S,3S
Quantity/Auction	10kg	Scale	9.2	The quantity of winter cabbage (number of cases) in an auction
Week	-	Ordinal	10.1	The week (number) in the year that production is to be auctioned
Scale(Producer)	ton/year	Ratio	10.2	Calculated based on the delivery volume of each producer in per year
Daily Quantity	ton/day	Ratio	12.5	Daily shipments. (Includes production for all producers)

Source: The author's survey, 2016.

Note: Calculated by SPSS Modeler 18

2.3. Results and Discussion

2.3.1. Importance of Classification Variables

MDA results of each variable are calculated (Table 1). Discussions regarding each variable's MDA results are as follows:

2.3.1.1. Daily quantity

Daily total shipment has become a highly important variable. This result is in line with our survey's result. The Producer wholesale market merchants believe that price fluctuations of winter cabbage are greatly influenced by supply and demand. Especially due to weather or other factors affecting supply shipments.

2.3.1.2. Scale

The second item is the scale of producers' shipments. Combined with actual survey results, if producers supply winter cabbage with long-term and stable shipment agreements, they always tend to have better trust relationships, and their winter cabbage's auction prices tend to be higher. This is also considered as an important reason why producers can build good contractual relationships with Producer wholesale market.

2.3.1.3. Week

We use one week as the time unit, because it effectively incorporates climate change on winter cabbage output and price. It also provides better prediction accuracy than unit of one day. It is necessary to mention that, for the construction of RF model, we still use the day index as the weekly and daily average prices are two independent variables which have no effect on each other.

2.3.1.4. Quantity/auction

Our results also show that the quantity of winter cabbage (number of cases) in an auction has high MDA value. Combined with actual survey results, the Producer wholesale market merchants tend to downstream orders as reference to participate in auctions, and if coupled with the increased demand for processing use of winter cabbage in recent years, we can infer that the quantity per order from downstream has been increasing. Accordingly, for Producer wholesale market merchants, winter cabbages with the same quality are more appealing. The raw data also shows us that, even with the same quality, winter cabbages sold in larger quantities in an auction (more than 200 cases) always fetch a higher price than those sold in smaller quantities (less than 50 cases).

2.3.1.5. Size

The size of a cabbage is also an important variable. In recent years, most winter cabbages are purchased in supermarkets, where size L is the most basic and popular size with retail customers. On the other hand, winter cabbage processing plants are inclined to purchase larger sizes due to the limited number of processing machines, in which sizes greater than 3L or smaller than M often cannot be machined. This also affects auction prices.

2.3.2. Model Parameters and Performance

2.3.2.1. Model parameters

We set the number of decision trees to 1,500, and the random splitting variable on the leaf node to greater than 3. The details of dataset can be found in Table 2. For this model, we set 50% of data to be the OOB data. The target(y) set as price deviation between ±15 yen.

Table 2 Details of RF Model for New Scheme

Dataset		**Model**	
Item	Outlier	Variable	Target
93,410	1,509	5	$-15 \le y \le 15$

2.3.2.2. Model performance statistics

The performance comparison of RF model can be found in Table 3. As seen in Table 3, in the RF model testing with OOB data, correct rate up to 86.82%, and when compared with the traditional artificial neural network (ANN) method, the RF algorithm had better accuracy and stability. This experiment determined that if this model is applied in market A’s new scheme, we can classify winter cabbages by predict price deviation before the auction to assist in the auction later on. To statistically analyze the RF model, we assumed that for the new scheme, each auction would last 5 minutes, and each centralized auction would last 8 minutes. According to this assumption, the result showed that our new auction scheme could reduce 31% of the auction time (Table 4).

Table 3 Model Performance Statistics of the Random Forest and ANN Classifiers

Variables	Correct Classification Rate (%)	
	RF	ANN
Selected predictor variables 1	70.53	70.18
Selected predictor variables 2	76.65	72.44
All variables	86.82	81.53

Source: The author's survey, 2016

Note: * Calculated by SPSS Modeler 18. All variables are randomly combined. Table 3 shows the combination schemes with top three accuracy rate.

**Selected predictor variables 1: Quantity/per, Week, Daily Quantity

*** Selected predictor variables 2: Size, Quantity/per, Week, Scale (Producer)

**** All variables: Size, Quantity/per, Week, Scale (Producer), Daily Quantity

Table 4 The Expected Effect of the RF Algorithm Classifier (Estimated Value)

	Total Number (auction)	Time (per auction)	Time (total)	Ratio
Separate	93,410	5	467,050	1
Separate & Centralized	52,905*＋20,253**	5(8***)	318,533	0.68

Source: The author's survey, 2016`

Note: a) * Auction and separate settlement

** Centralized auction and pool settlement

*** Centralized auction

b) Time data in Table 4 estimate data based on my own survey.

c) **Assuming that after changes to a centralized auction, the number of auctions will be reduced by two-thirds.

2.4. Summary

Faced with challenges and problems encountered by the application of machine learning technology in regional agricultural promotion and given the opportunities created by RF algorithm, we have developed a new trade scheme for vegetable trading which can be utilized by Producer wholesale markets. This trade scheme uses RF model, which has already proven to be efficient in many fields like precision agriculture, plan breeding, and yield prediction. This study ascertained that machine learning technology can offer a new insight in to regional agricultural promotion and provide benefits for Producer wholesale markets.

In the process of building the RF model, five variables that were related to auction price fluctuations were considered and verified. Results of this study confirmed that these variables can accurately predict price deviation and by using this RF model scheme, Producer wholesale markets can improve their auctions' efficiency. Relevant personnel of Producer wholesale markets can reference this study to develop more effective and targeted selling strategies and logistically sound planning.

It is necessary to state that the RF model built in this study is not perfect. Due to experimental conditions, this RF model has not been able to add more data (more than five years) or variables (such as weather and vacation) and since the learning data was only collected from Producer wholesale market A, it is not appropriate to compare our results with other case studies. Clearly, more research is required to improve the prediction accuracy and application range of the RF algorithm. Despite this fact, the result of this study still contributes a partial solution to help reform Producer wholesale markets.

In conclusion, the key findings of this study support an efficiency trade scheme for Producer wholesale markets and give a meaningful exploration for the ideas on how to incorporate machine learning technology on the basis of big data in order to revolutionize Japanese agriculture.

2.5. Notes

1) Auction and separate settlement: The source area wholesale market's trade system ensures that each producer can get a price corresponding to their own product quality.

2) Consignment sale and pooling settlement: Products provided by all JA Group members will be settled at the same price.

3) JA Group members: Producers that have a contractual obligation relationship with the JA Group.

4) The price of winter cabbage fluctuated greatly from 2011 to 2016. Since the price deviation is calculated as auction price minus average price, in case of a fluctuation, the deviation remains unchanged if the auction price and average price changed in equal measure.

3. A Case Study on Consumption-Area Wholesale Market

3.1. Introduction

According to the report of the FAO of 2009, it is estimated that the global population would increase by more than 30% until 2050, which means that it is necessary to increase food production by 70%. To satisfy this increasing demand, several initiatives have been launched like use emerging digital technologies and biotechnologies such as cloud computing, remote sensing, and Internet of Things to support agricultural practices.

To address the challenge of these initiatives, many studies related to monitoring, measuring and analyzing continuously various physical aspects and phenomena of agricultural practices have been carried out. The aforementioned emerging digital technologies import to these new studies producing large quantities of data in an unprecedented pace, and implies the need for storage, preprocessing, analysis and modeling of huge amounts of data coming from various heterogeneous sources, so that researchers realize the implementation of these technologies is based on big data analysis, hence leading to the new notion of "agricultural big data".

Kamilaris and Kartakoullis et al. (2017) and Kamilaris and Francesc et al. (2018) performed reviews of agricultural big data and analytical techniques1), and confirm that although in contrast to the rather profuse bibliography on big data, the agricultural subset appears to be relatively recent and limited, but the number of studies has a tendency to increase rapidly. They indicated that the reason for the increase is when researchers experiment with agricultural big data analysis, applying it to resolve various agricultural problems involving classification or prediction, the complex, multivariate and unpredictable agricultural ecosystems can be better understood.

Furthermore, the results of their study confirm that 79% of papers solve problems like land degradation and water contamination, climate change from a technical perspective. Additionally, the rest of the papers solve problems like governmental policies, market fluctuations and sociocultural development (e.g. dietary preference of meat protein), they speculated that one important reason of the lack of studies in these areas is that compared with climate and geographic data, social behavior data are less open and complete (especially business data), which makes the research progress lag behind.

To address these governmental policies, market fluctuations issues studies, researchers in Japan have found an opportunity.

In the construction of Japan's agricultural product distribution system for many years, Japanese farmers set up group, companies or join Japan Agricultural Cooperatives (JA) for mass production. And products are traded in wholesale markets which administered by the Wholesale Market Act. After years of standardization, these links of the distribution system have kept a large number of complete transaction data. Combined with the macro data collected by Ministry of Agriculture, Forestry and Fisheries (MAFF), can provide high-quality data sets for agriculture big data to solve problems such as government policies and market fluctuations in the field of domestic agricultural products distribution.

Based on Japan's agricultural distribution system has the advantage of providing high-quality data sets, Li N. and Maezawa S. (2018) developed a new trade scheme for vegetable trading that can be utilized by Producer wholesale markets. Their study provides agriculture big data with machine learning Technology (ML) could undertake better decision making and informed actions in real-world scenarios without (or with minimal) human intervention by using agriculture big data, also provides meaningful exploration for the ideas on how to incorporate ML based on big data in order to solve market fluctuations problem.

Considering the characteristics of the distribution system of agricultural products in Japan, the Producer wholesale market in Li’s study belongs to the wholesale market element in the production area, leaving the wholesale market for consumption area that is yet unstudied.

The motivation for this study stems from the fact that big data analysis in agriculture is a modern technique with growing popularity, while recent advancements and applications of big data in other domains indicate its large potential. And given the advantages by the ML as the most extensive technology of big data analysis. The objective of this study is to explore a scheme with ML for wholesale market in consumption area so that could solve governmental policies and market fluctuations problems. This study will verify the possibility of ML playing an active role in the whole agricultural distribution system in Japan on the basis of filling the blank in previous research.

3.2. Materials and methods

3.2.1. Target of Scheme

After determining that the wholesale market in consumption area is the target of this study, we chose Tokyo, the most populous economic circle in Japan, as the research area. Specific findings are as follows:

3.2.1.1. Target Market

In order to eliminate the problems of Arrow distribution and low efficiency of goods turnover in the wholesale market, since the 1950s, a long period of research has been carried out. One of the most important results of this research is the reconstruction of the old wholesale market in the south of Tokyo, so as to resolve the problem of overcrowding of shops in the Kamida wholesale market in the middle of Tokyo. Then, in order to ensure the smooth distribution of fresh produce in Tokyo in the future, a policy of establishing a comprehensive wholesale market of vegetables, aquatic products and flowers were formulated. Thus, the Ota wholesale market was established.

With its excellent location and policy-advantaged traffic conditions, Ota market has now become the largest vegetable wholesale market in Japan. The average daily trade of fruits and vegetables has reached more than 3,600 tons.

3.2.1.2. Target Data Source

We have learned from the previous survey of market situation, given that it had become the largest wholesale market for consumption in Japan, the daily vegetable trading price in the Ota market had also become the indicator price for vegetable trading throughout the country. Therefore, if ML can be used to improve the system of price formation in the Ota market, it will have a beneficial impact on vegetable distribution system and it will be a useful scheme to solve market fluctuations problems.

But in subsequent investigations, we found that the transaction prices for vegetables in the Ota market were formed between a number of wholesalers and downstream distributors. Furthermore, the transaction result data involved too many commercial secrets, such that we were told that we could not obtain plenary transaction price data. Therefore, under the condition that the transaction result data cannot be obtained, we analyzed the trading varieties and quantity of vegetables for all items in the Ota market and ranked it according the trading varieties2). The results are shown in Table 5. It is clear that the most traded item during the year are cabbage from table 5.

Table 5 Ranking of Vegetables Traded in Ota Market in 2016

	Item	Quantity (t)	Amount (1000yen)	Price (yen/kg)
1	Cabbage	90,213	9,144,410	101
2	Onion	61,335	6,612,918	108
3	Hakusai	60,297	5,448,344	90
4	Daikon	55,871	6,111,111	109
5	Lettuce	54,982	12,123,703	221

Source: MAFF, Survey Report on Fruit and Vegetable of Wholesale Market in 2016, Published in 2018.01.18

Then, in order to analyses the relevant data of transaction prices, while ensuring that the data source does not lack the characteristics of the consumption sector, we finally determined and analyzed the data contained in the Tokyo food price index, published by the Tokyo municipal government and compared it with the price index changes for several categories related to vegetables. The results are shown in Figure 2.

3.2.1.3. Target Agricultural Product

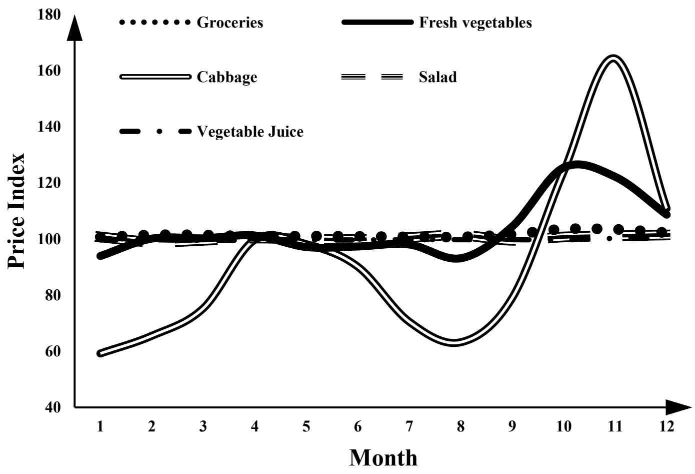


Figure 2 Tokyo Food Price index change chart 2016

Source: Statistics Japan, 2017.

Figure 2 allows us to understand that the price index of cabbage fluctuates in a wide range and there is a highly positive correlation trend between it and the data of other fresh vegetables. In this regard, we used cabbage as the key word to capture and analyses the contents of Twitter in the same period of time for residents in Tokyo. The results are shown in Table 6 and Figure 3.

Table 6 Details of Tweets Collected

Period	2016.01.01-2016.12.31
Area	Tokyo
Key word	Cabbage
Sampling	Random
Method	(Average monthly)
Total	184,177 (Twitters)

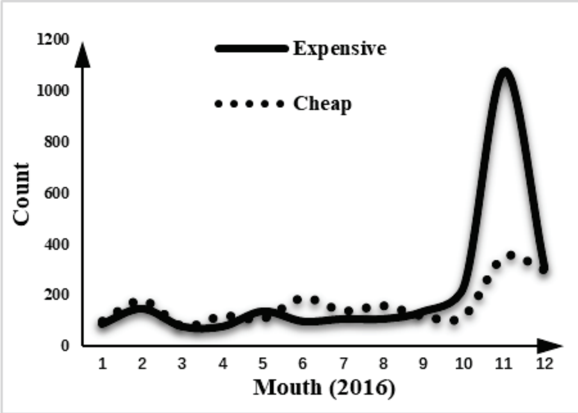


Figure 3 Text Mining Results of Twitters

Source: The author's survey, 2018

It can be seen from Figure 3 that, as the price of cabbage fluctuates, the degree of concern for cabbage among residents in Tokyo also changes synchronously. Therefore, for the following reasons we chose to use cabbage as the base vegetable: Firstly, in the wholesale market in consumption sector, cabbage transaction volume is the highest among all vegetables; Secondly as a kind of fresh vegetable, the transaction price of cabbage in the consumption area is related to the production cycle and has obvious seasonal fluctuations; Thirdly when prices change dramatically, consumers respond quickly and sensitively. Based on those reasons, we set cabbage as the target vegetable of this study.

3.2.2. Logic of Schemes

Through interviews with market managers, we learned that the price of cabbage in the market was generated by both an auction price and a contract price. Due to the changes in the distribution mode of vegetables in recent years, large-scale chain supermarkets and restaurants have become the main downstream distributors. These large-scale distributors usually sign annual purchase contracts with wholesalers, since future vegetable prices are difficult to predict when making the contract, so these contracts are basically done in a way that guarantees a certain quantity supplied at current prices. Therefore, although the demand for produce auctions has gradually disappeared, wholesalers need to judge the change of the market's supply and demand relationship by the auction price and use it to adjust the current contract price. Accordingly, Ota market participants hold regular meetings at the beginning of each month, and the auction volume of that month is set and becomes mandatory[3]. But this mandatory amount set aside for auction makes it impossible for them to understand the total supply of the Ota market on any given day, in turn making it difficult to bid on the basis of market-clearing prices. Therefore, in order to provide vegetable traders with the information pertaining to the amount supplied for their reference during trading, market managers investigate and forecast the next day's supply in the afternoon of each market business day and the forecast data are published on the Tokyo Metropolitan Central wholesale Market's website. We collected and analyzed these forecast data and the Paired samples T test results are shown in Table 7.

Table 7 Paired samples T test results

	Mean	N (df)	Range	Minimum	Maximum	Std. Deviation	Std. Error Mean	Difference Lower	Difference Upper	t	Sig. (2-tailed)
Actual	**309.1**	**265.0**	**405.1**	**151.0**	**556.2**	**60.1**	**3.7**				
Forecast	**257.8**	**265.0**	**495.0**	**104.7**	**599.6**	**64.1**	**4.0**				
Actual-Forecast	**52.1**	**265.0**	**267.6**	**-130.4**	**137.2**	**1.9**	**2.0**	**48.2**	**56.0**	**26.5**	**0.0**

Source: The author's survey, 2018.

Note: Calculated by SPSS Statistics

From Table 7, we can see that the forecast data released by the managers has a large error term with respect to the actual transaction data. Such error also affects all participants' bidding activities, thus affecting the formation of a general market price. Therefore, we can use ML technology to predict the daily supply amount by learning the existing auction data and other variables' data. In particular, it needs to be emphasized that in this process, the role of our model is not to directly predict the supply amount, but to act as an auxiliary tool to assist the market manager to correct his forecasts.

3.2.3. Algorithm of Scheme

Kamilaris and Francesc et al. (2018) pointed out that the most popular techniques used for agriculture big data analyzing in these studies include machine learning (ML), linear polarizations, wavelet-based filtering, vegetation indices (NDVI) and regression analysis, and 89% of these studies are chosen using ML, since ML as an emerging technology has become even more attractive especially when dealing with more realistic and complex situations.

One of the main advantages of ML technologies is their ability to automatically solve large nonlinear problems using data sets from multiple (possibly interconnected) data sources. They first learn to identify structures and patterns in complex data sets and then use the acquired models (similar to human experience) to predict future events. ML provides a powerful and flexible framework for not only data-driven decision making, but also integrating expert knowledge into the system itself.

In recent years different algorithm of ML have been implemented to achieve accurate prediction. The most successful algorithm has been artificial neural networks (ANNs), random forest (RF), K-means, support vector machines (SVM) and support vector regression (SVR). K-means, SVM and SVR can be collectively referred to as linear regression and polynomial regression, their advantage is that they can be modeled quickly, especially when the relationship to be modeled is not particularly complex and the data set is small, however, the model design is difficult because it must have accurate information about the relationship between characteristic variables and data structure. On the contrary, ANNs and RF can have multiple nonlinear layers (and parameters), so the model can has a good fit for modeling more complex nonlinear relationships and usually have high performance than polynomial regression (Chlingaryan, A. and Sukkarien, S. et al. 2018). Considering that RF is usually having the same performance as ANNs, we decided to use both algorithms for modeling in this scheme and to compare the performance of two models.

3.3. Results and Discussion

3.3.1. Features and Data pre-processing

There are many features that affect the supply of cabbage in Ota market, some of the features are fundamental and, thus, may not be omitted, e.g. the previous year's supply and forecast error. While some additional features can either be included or not, like climate data or the transaction prices for cabbage production, in order to improve prediction accuracy, we decided to collect and perform accurate classification of all relevant features. After randomly grouping these features, the resulting groups will be able to compare and filter the following model training process.

The accuracy of the predictions produced by the ML algorithms strongly depends on data quality. Erroneous data, data with a high level of noise and incomplete datasets may significantly reduce the predictive power of the models. Therefore, under the premise of ensuring the integrity of each feature's dataset, we conducted some pre-processing of outlier detection and also performed automated dataset cross-validation.

To ensure the best fit and to avoid the randomization and overfitting effects, we divided the datasets between training and testing and verification data, using a ratio of 80–20 and 90–10, respectively, and then different datasets were used to test the scheme.

3.3.2. Model Parameters and Performance

We use four parallel control groups that use ML algorithms to study with all randomly combined feature groups under conditions wherein all the other parameters are the same. Then, the model with smallest mean relative error (i.e. the mean error between predicted and observed values, stated as a percentage) was elected to be the optimal model within the group. Finally, the optimal models of four parallel control groups were compared (Table 9).

The comparison results in Table 9 indicate that the prediction accuracy of the ML model is much higher than artificial prediction yields. And using different ratios for training or testing of datasets has little effect on the prediction accuracy of the models. A ratio of 90–10 is better than 80–20, however, the possibility of overfitting cannot be precluded, yet at this stage, one cannot prove it with data.

Finally, through comprehensive consideration, the model using ANNs and 90–10 data sets was determined to be the scheme's model of choice. On the principle of satisfying the precision requirements, the number of hidden layers was as small as possible. We optimized the scheme model. The final model utilized multilayer perceptron architecture, with two hidden layers of eight, two nodes and one single output variable.

Table 8 Summary of Features

Feature		Source
Actual Amount	Numeric	Ota market, 2016
Forecast Amount	Target	
Error Amont	Numeric	
Weather 1	Categorical	Tokyo Area, 2016
Temperature 1	Numeric	
Price index	Numeric	Tokyo Area, 2015
Living Expenditure	Numeric	
Weather 2	Categorical	Atsumi Area[4)], 2016
Temperature 2	Numeric	
Supply amount	Numeric	

Source: The author's survey, 2018.

Table 9 Comparing results of models

	Mean	Range	t	df	Sig. (2-tailed)
Actual-Forecast	52.1	267.6	26.5	264	0
Actual-ANNs(90-10)	12.6	28.2	3.52	264	0.696
Actual-ANNs(80-20)	16.8	31.8	4.3	264	0.425
Actual-RF(90-10)	23.5	79.6	8.6	264	0.493
Actual-RF(80-20)	21.2	46.3	11.3	264	0.436

Source: The author's survey, 2018.

3.4. Summary

Given the advantages offered by ML technology to make better decision making and informed actions in real-world scenarios without (or with minimal) human intervention by using agricultural big data, we developed a scheme with an auxiliary forecasting function for cabbage trading that can be utilized by Japanese wholesale markets dedicated to the consumption area. This scheme uses ML technology that has had significant developments in recent years, along with successful application in many areas like precision agriculture, plant breeding and yield prediction. This scheme uses ML technology that has had significant developments in recent years, along with successful application in solve issues like land degradation and water contamination, climate change from a technical perspective. This study ascertained that ML technology can optimize the wholesale market trading system in consumption area and offer new insights into solve agriculture market fluctuations issues in Japan.

In the process of defining the optimal model by comparing different ML algorithms, we found that ML models provide much higher predictive accuracy than traditional artificial ones. The results of this study confirmed that the most suitable model for the scheme designed in this study is an ANN model. Through the auxiliary function of the model, specific suggestions can be provided to market managers to effectively improve the problem of forecasting error. Relevant personnel of the vegetable consumption sector can reference this study to develop more effective and targeted trading strategies and logistical planning, too.

It is necessary to mention that—due to the fact that each scheme involved different datasets, pre-processing techniques, metrics, models and parameters—it was difficult to generalize and perform comparisons between schemes. In addition, in the domain of agriculture, because many publicly available datasets for researchers to work with are not available, researchers need to develop their own datasets, and this could require much time and effort. Clearly, more research and contributions are required to improve the application range of ML technology.

In conclusion, the key findings of this study is support a scheme with an auxiliary forecasting function that can be utilized by managers in vegetable wholesale markets in consumption area, and give meaningful exploration regarding the ideas on how to incorporate ML technology by analyzing agricultural big data, the result of this study can provided a valuable reference materials for solve market fluctuations issues and the agricultural revitalization of Japan.

4. General discussion

4.1. Full case summary

Based on machine learning technology of Internet of things of agriculture decision support system in the system as its core, data from the depth of mining in the process of the vegetable trading characteristics, convert it to the height of the scientific research, reflect its actual value, and formulate the corresponding decision measures, from macro to micro precision agriculture to agriculture. This paper studies the theory and application of the data analysis subsystem of the agricultural Internet of things decision system.

First of all, the whole theoretical system of agricultural big data is deeply studied from the theoretical level. The international and domestic research and development situation are analyzed and compared, and the advantages of neural network algorithm of machine learning and other algorithms in data processing are analysed.

Secondly, the machine learning theories such as linear regression, ANN and SVR are explained from the theoretical level. The autocorrelation function and partial autocorrelation function of model data relevance are sorted out to grasp the construction process of machine learning as a whole system.

Thirdly, in the data analysis system, the analysis and modelling are carried out based on the data samples of the market of origin and the market of consumption. The results show that the new improved scheme can predict the required data efficiently and accurately. This not only achieved the purpose of improving the transaction process but also made a more reasonable layout plan and saved the labour and material costs. Based on Japan's agricultural distribution system has the advantage of providing high-quality data sets, the first part of this paper developed a new trade scheme for vegetable trading that can be utilised by source-area wholesale markets. We have developed a new trade scheme for vegetable trading which can be utilised by source-area wholesale markets. This trading scheme uses a random forest model, which has already proven to be efficient in many fields like precision agriculture, plan breeding, and yield prediction.

This part of study ascertained that machine learning technology could offer new insight into regional agricultural promotion and provide benefits for source-area wholesale markets. In the process of building the random forest model, five variables that were related to auction price fluctuations were considered and verified. Results of this part of the study confirmed that these variables could accurately predict price deviation and by using this random forest model scheme, source-area wholesale markets can improve their auctions' efficiency. Relevant personnel of source-area wholesale markets can reference this study to develop more effective and targeted selling strategies and logistically sound planning. This part of study provides ML could undertake better decision making and informed actions in real-world scenarios without (or with minimal) human intervention by using agriculture big data, also provides some meaningful exploration for the ideas on how to incorporate ML based on big data in order to solve market fluctuations problem.

Considering the characteristics of the distribution system of agricultural products in Japan, the source-area wholesale market in the first part of this study belongs to the wholesale market element in the production area, leaving the wholesale market for consumption area that is yet unstudied.

The second part of this paper with the motivation that stems from the fact that big data analysis in agriculture is a modern technique with growing popularity, while recent advancements and applications of big data in other domains indicate its tremendous potential. Moreover, given the advantages by the ML as the most extensive technology of big data analysis, we developed a scheme with an auxiliary forecasting function for cabbage trading that can be utilised by Japanese wholesale markets dedicated to the consumption area. This scheme uses ML technology that has had significant developments in recent years, along with successful application in many areas like precision agriculture, plant breeding and yield prediction. This part of the study ascertained that ML technology could offer new insights into regional agricultural promotion in Japan and optimise the wholesale market trading system used in the consumption sector.

In the process of defining the optimal model by comparing different ML algorithms, we found that ML models provide much higher predictive accuracy than traditional artificial ones. The results of this study confirmed that the most suitable model for the scheme designed in this study is an ANN model. Through the auxiliary function of the model, specific suggestions can be provided to market managers to improve the problem of forecasting error effectively. Relevant personnel of the vegetable consumption sector can reference this study to develop more effective and targeted trading strategies and logistical planning, too.

In conclusion, the key findings of these two parts of our study support schemes with auxiliary forecasting functions that can be utilized by managers in vegetable wholesale markets and give some meaningful exploration regarding the ideas on how to incorporate ML technology through analyzing agricultural big data that might, in turn, be used to revolutionize Japanese agriculture.

4.2. Future work outlook

4.2.1. In this study

This study support schemes with auxiliary forecasting functions that can be utilized by managers in vegetable wholesale markets and give some meaningful exploration regarding the ideas on how to incorporate ML technology through analyzing agricultural big data that might, in turn, be used to revolutionize Japanese agriculture. Due to the limitation of time and energy, the research work of this paper needs to be further improved. It is hoped that the following aspects can be improved in the future:

a) In the market of the place of production, the prediction is mainly based on the historical data of market A, and the prediction effect is relatively accurate, but there is still room for improvement. According to the characteristics of the current production market, the machine learning model can be improved correspondingly. At the same time, the formation of vegetable price is affected by a variety of complex factors. The accuracy of the model can be further improved by adding factors with large influence factors into the current model.
b) In terms of data analysis, there are many data analysis tools, many of which

require computer knowledge. In the whole agricultural Internet of things decision-making system, more knowledge bases should be added, and different data analysis tools should be used to analyze the acquired data samples.

c) In the second part of this paper, the data analysis and decision system mainly analyzes the 2016 data, and USES the artificial neural network algorithm, which has achieved very good results. However, there are many uncertain factors in production and circulation, which have different influences on experimental data. In order to further explore the time-space correlation between data, more data of years or data samples of larger areas should be collected to improve the value analysis of sample data. At the same time, more agriculture-related data should be collected to deeply explore the value of data from multiple perspectives, so as to form a more comprehensive analysis and decision-making system.

d) With the continuous development and application of deep learning in the data based on time series, some scholars have applied it in the data analysis of meteorology. In the future model research, it can be studied as an important model and compared with the current model for analysis.

e) It is necessary to mention that—due to the fact that each scheme involved different datasets, pre-processing techniques, metrics, models, and parameters—it was difficult to generalize and perform comparisons between schemes. In addition, in the domain of agriculture, because many publicly available datasets for researchers to work with are not available, researchers need to develop their own datasets, and this could require much time and effort. Clearly, more research and contributions are required to improve the application range of ML technology.

f) In addition, there are some problems and development directions mentioned in the relevant research, which are summarized and described next.

4.2.2. Related research

4.2.2.1. Open problems of big data in agriculture

The application of big data analysis in agriculture has not been beneficial in all cases, as it has created (or is expected to create) some problems too. We list below the problems, as identified and mentioned in the revised research:

a) From a sociopolitical perspective, the creation of large monopolies in the agri-food industry and dependence of the farmers on large corporations about their farming operations becomes possible (Sykuta, 2016). Big data concentrated in the hands of big agribusinesses limits the potential of this technology, only reinforcing the capacities and business advantages of a few corporations (Carbonell, 2016).

b) Privacy issues are raised, in respect to who owns the data and who can monetise it (Nandyala and Kim, 2016). Farmers are concerned about the potential misuse of information related to their farming activities (Shin and Choi, 2015), by seed companies or competitor farms (Carolan, 2016). (Schuster, 2017) warns that hedge funds might use real-time data at harvest time from a large number of sources (e.g. weather data, yields predictions, remote sensing, data from machinery such as combines, etc.) to speculate in commodity markets.

c) The practice of big data collection and analytics has raised questions over its security, accuracy and access, as discussed in (Nandyala and Kim, 2016) and (Sykuta, 2016).

d) Moreover, the use of big data differs in developed vs developing countries, according to Kshetri (2014) and Rodriguez et al. (2017). RIICE Partnership (2014) and Syngenta (2010) believe a digital divide exists between developed and developing economies, due to unbalanced access to technology (i.e. computing power, internet bandwidth and sophisticated software), and lack of skilled analysts in the developing world. Especially in respect to volume and variety, big data in the developing world is smaller-scale and less diverse (Rodriguez et al., 2017), as the surveyed papers suggest

(Tesfaye et al., 2016; Frelat et al., 2016; Sawant et al., 2016; GSMA, 2014; Akinboro, 2016). Big data collection efforts mainly benefit big, well-educated farmers who have the means and the expertise to collect it successfully and accurately (Kshetri, 2014; Oluoch-Kosura, 2010).

e) From a technical perspective, product developers have only limited access to ground truth information (Atzberger, 2013), an issue that has been observed in many of the revised papers (Armstrong et al., 2007; Waldhoff et al., 2012; Sakamoto et al., 2005; Jóźwiaka et al., 2016; Frelat et al., 2016). Ground truth information is necessary for evaluating products and services in various settings and physical or weather conditions (Capalbo et al., 2016). Also, visualisation of large data volumes is still tricky (Schnase et al., 2014; Karmas et al., 2016).

4.2.2.2. Barriers for broader adoption of big data analysis

Related work indicated various barriers hindering the wider use of big data analysis, such as lack of human resources and expertise (Sawant et al., 2016) and limited availability of reliable infrastructures to collect and analyse big data (Akinboro, 2016; Syngenta Foundation for Sustainable Agriculture, 2016). Frelat et al. (2016) note that accurate and actionable data requires considerable technical skills to handle data mining and analysis methods, while infrastructures are needed for efficient data storage, management and processing of multi-modal and high-dimensional datasets, including provisioning for real-time processing in many critical geospatial applications (Karmas et al., 2016).

Further, there is generally a lack of structure and governance related to agricultural big data, as pointed out by Nandyala and Kim (2016) and Nativi et al. (2015), as well as identified and addressed in some of the revised papers (Schnase et al., 2014; Marcot et al., 2001; Becker-Reshef et al., 2010). Kempenaar et al. (2016) suggest that business models are needed that are attractive enough for solution providers, enabling at the same time a fair share between the different stake-holders. In addition, Lokers et al. (2016) consider that the general absence of well-defined semantics complicates big data understanding and reuse by other researchers and organizations.

As observed in this study (see Table 1), much of the attention on big data has focused mainly on large volumes (i.e. applications in weather and climate change, land identification, farmers' decision-making, insurance and finance, remote sensing). This has led to a skewed and narrow perspective of the value of big data to organizations and the society since aspects of data velocity, variety, veracity and valorization are equally important, as pointed out by Shin and Choi (2015) and Capalbo et al. (2016).

Moreover, technical challenges of remote sensing systems for farm management still exist (Zhang and Kovacs, 2012), such as the collection and delivery of images in a timely manner (Galford et al., 2008), sampling errors and the lack of high spatial resolution data (Nativi et al., 2015), image interpretation and data extraction issues (Karmas et al., 2014), the influence of weather conditions (Barrett et al., 2014), etc. Finally, common barriers involve the absence of data itself (or part of it) and its limited reliability, variety or time relevance as observed and discussed in some of the revised papers (Fuchs and Wolff, 2011; Schnase et al., 2014; Schuster et al., 2011; Armstrong et al., 2007; Frelat et al., 2016; RIICE Partnership, 2014; Marcot et al., 2001).

4.2.2.3. Addressing open problems and overcoming barriers

From a sociopolitical view, many farmers from around the world started to mobilize and organize themselves (e.g. in cooperatives, on-line communities), increasing their power in terms of sharing of know-how and experiences, and big data understanding (Farm Hack, 2010). Shin and Choi (2015) believe that the data-driven economy has the potential to create suitable knowledge for the users of data ecosystems, such as the one of agriculture.

Moreover, formation of a policy framework on data ownership is required Sykuta (2016)), which will protect owners' copyrights and control user access (Shin and Choi, 2015). Policies for data manage- ment and security are needed (Kshetri, 2014), towards the democratization of big data, broadening its potential impact and value through the adoption and accessibility of appropriate support tools (World Economic Forum, 2012).

From a technical aspect, investments in cloud infrastructures are essential for large-scale storage, analysis and visualization of agricultural data (Hashem et al., 2015), supporting business analytics in high scale and speed (Kempenaar et al., 2016). The infrastructures should be easily accessible by non-technical personnel and not be expensive (Hashem et al., 2015). Techniques such as data aggregation, data reduction and proper analysis can contribute towards more userfriendly platforms (Karmas et al., 2014). Various ways of technically managing high volumes of widely varied data, addressing the “V”s dimensions of big data effectively are discussed in Karmas et al. (2016) and Nativi et al. (2015).

Also, well-defined and commonly accepted technologies for data semantics (e.g. RDF, OWL, SPARQL and Linked Data) and ontologies (e.g. AGROVOC, Agricultural Ontology Service and AgOnt) can be used as common terminologies towards data interoperability, as proposed in Lokers et al. (2016), Cooper et al. (2013), Kamilaris et al. (2016). Karmas et al. (2016) suggest that open standards need to be adopted and agreed for data integration, such as OGC.

Furthermore, open-source software tools and libraries would be useful, such as crop type maps and calendars (Sakamoto et al., 2005), biophysical measures and vegetation indices (Wardlow et al., 2007), yield models (RIICE Partnership, 2014), crop area estimates (Becker-Reshef et al., 2010) and seasonal weather forecasts (Tripathi et al., 2006). These tools should be easily mergeable with other platforms to support large-scale, highly-varied data analysis, a strategic goal of the GEOSS platform (Nativi et al., 2015) and MERRA Services (Schnase et al., 2014).

More big datasets should become publicly available (Carbonell, 2016), and there is a growing trend towards this direction already (OADA, 2014; GODAN, 2015; AgGateway, 2005). Besides, numerous A. Kamilaris et al. Computers and Electronics in Agriculture 143 (2017) 23–37 organizations on the web have started to provide various large-scale datasets covering a wide spectrum of agricultural areas.

4.2.2.4. Potential areas of application of big data analysis

This subsection lists potential areas of applying big data analysis for addressing various agriculture-related problems in the future. These areas have not been covered adequately (or not covered at all) by the existing research and papers under study (according to the authors' opinion), and include the following possibilities:

a) Platforms enabling supply chain actors to have access to high-quality products and processes (RIICE Partnership, 2014; Sawant et al., 2016), enabling crops to be integrated to the international supply chain, according to the global needs (Cropster, 2007; Syngenta Foundation for Sustainable Agriculture, 2016).

b) As farmers are sometimes not able to sell harvests due to over-supply or not getting the planned harvest (Frelat et al., 2016; Syngenta Foundation for Sustainable Agriculture, 2016), tools for better yield and demand predictions must be developed (Tesfaye et al., 2016; Kempenaar et al., 2016; Becker-Reshef et al., 2010).

c) Providing advice and guidance to farmers based on their crops' responsiveness to fertilizers is likely to lead to a more appropriate management of fertilizer use (Giller et al., 2011). This could apply as well to better use of herbicides and pesticides (i.e. (Gutiérrez et al., 2008; Sawant et al., 2016; aWhere Inc., 2015)).

d) Scanning equipment in plants, shipment tracking and retail monitoring of consumers' purchases creates the potential to enhance products' traceability through the supply chain (Armbruster and MacDonell, 2014), increasing food safety (Jóźwiaka et al., 2016; Lucas and Chhajed, 2004). Prevention of foodborne illnesses is an issue that requires international collaboration and investment by local/global organizations and governments (Grace and McDermot, 2015), to ensure safer food (Chedad et al., 2001; RIICE Partnership, 2014). Moreover, since the agricultural production is prone to deterioration after harvesting (Wari and Zhu, 2016), optimization procedures are essential to minimize losses and maximize quality (Pierna et al., 2004). Promising optimization techniques already being applied to food processing involve (meta-) heuristics and genetic algorithms (Wari and Zhu, 2016), as well as neural networking

(Erenturk and Erenturk, 2007).

e) Remote sensing for large-scale land/crop mapping will be critical for monitoring the impacts of various countries and areas in respect to measuring and achieving their productivity and environmental sustainability targets (Barrett et al., 2014; Waldhoff et al., 2012; Schuster et al., 2011; Becker-Reshef et al., 2010).

f) More advanced and complete scientific models and simulations for environmental phenomena could provide a basis for establishing platforms for policy-makers, assisting in decision-making towards sustainability of physical ecosystems (Schnase et al., 2014; Nativi et al., 2015; Song et al., 2016)

g) High-throughput screening methods that can offer quantitative analysis of the interaction between plants and their environment, with high precision and accuracy (Furbank and Teste, 2011; Karmas et al., 2014).

h) Self-operating agricultural robots could revolutionize agriculture and its overall productivity, as they may automatically identify and remove weeds (Gutiérrez et al., 2008; Blue River Technology, 2011), identify and fight pests (PEAT UG, 2016), harvest crops (Waldhoff et al., 2012), etc.

i) Fully automatized and data-intensive closed production systems (i.e. greenhouses and other indoor led-illuminated aeroponics) (Love et al., 2014; Anon., 2016), would be on the rise within the framework of the circular economy (e.g. less use of pesticides, water and nutrient recycling, proximity to the consumer, etc.).

j) Precise genetic engineering, known as “genome editing”, would make it possible to change a crop or animal’s genome down to the level of a single genetic “letter” (Hartung and Schiemann, 2014). As (González-Recio et al., 2015) comment, this could be more acceptable to consumers, because it simply imitates the process of mutation on which crop breeding has always depended, and it does not imply the generation of transgenic plants or animals. This technology would supplement existing research in epigenetics (McQueen et al., 1995; Tesfaye et al., 2016).

The majority of the aforementioned potential applications would produce large amounts of (big) data, which could be used by future policy-makers to balance offer and demand (applications #1 and #2), reduce the negative impact of agriculture on the environment (applications #3, #5 and #9), raise food safety (applications #2, #4 and #6) and security (applications #2, #7, #8 and #10), increase productivity (applications #8, #9 and #10). The potential open access of this data to the public could create tremendous opportunities for research and development towards smarter and more sustainable farming.

5. Conclusion

This paper performed research on the importable of ML Technology in the agricultural field, based on this research, the reader can be informed about which types of agricultural applications currently use big data analysis, which characteristics of big data are being used in these different scenarios, as well as which are the conventional sources of big data and the general methods and techniques being employed for big data analysis. Open problems have been identified, together with barriers for broader adoption of the big data practice. Various approaches to addressing these problems and mitigating barriers have been discussed.

As we saw in this research, the availability and openness of hardware and software, techniques, tools and methods for big data analysis, as well as the increasing availability of significant data sources and datasets, shall encourage more initiatives, projects and startups in the agricultural sector, either addressing some of the problems or focusing on some of the emerging future application areas we have identified, or even creating radical-new services and products applied in new agricultural areas.

This increasing availability of big data and big data analysis techniques, well described through common semantics and ontologies, together with adoption of open standards, have the potential to boost even more research and development towards smarter farming, addressing the big challenge of producing higher-quality food in a larger scale and in a more sustainable way, protecting the physical ecosystems and preserving the natural resources.

6. Acknowledge

Years are like a shuttle, three years of time is so quiet and quietly flowing, the doctoral student's study career is about to be celebrated, and I have learned a lot in the past three years. Upon the completion of this thesis, I sincerely thank my family and friends who have been accompanying and supporting me all these years.

First of all, I would like to thank my doctoral supervisor Maezawa Shigenori who has been cultivating this field for three years. Professor Maezawa is unique in scientific research and academic research because of his rigorous and serious attitude, rigorous thinking logic, problem discovery and information mining. He discusses problems and papers from multiple perspectives. In the past three years, there have been considerable progress in both social conduct and scientific research and learning, which cannot be separated from the careful instruction of the professor. On the occasion of graduation, I would like to thank professor Maezawa for his tireless teachings over the past few years, which have laid a solid foundation for my future development.

I also want to thank the deputy supervisor Arahata Professor and Shibagaki Professor, whose valuable feedback, suggestions and comments increased significantly the overall quality of this survey.

Secondly, I would like to express my gratitude to all of you in the food circulation science laboratory who have been pursuing scientific research, for their support which has created an excellent environment for scientific research and learning. Here we are fighting shoulder to shoulder and exploring into unknown fields.

I would also like to thank the managers of wholesale market who provided me with data and market operation in the experimental investigation. It is their help that makes me have the courage to continue in the difficult research again and again.

Thanks to the evaluation teacher and all the people I know and do not know who have helped me.

7. Reference

[1] Breiman, L., “Random forests”, Kluwer Academic Publishers Press, 2001 pp.5-32.

[2] Cutler, D. R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., “Random forests for classification in ecology”, Ecology 2007, 88, 2007, pp.2783-2792.

[3] Everingham, Y., Inman-Bamber, G., Sexton, J. and Stokes, C., “A Dual Ensemble Agroclimate Modelling Procedure to Assess Climate Change Impacts on Sugarcane Production in Australia”, Agricultural Sciences, Vol.6 No.8, 2015, pp.870-888.

[4] Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., Müller, J., “Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes”. Agricultural Water Management Vol.116, 1 January 2013, pp.142-150

[5] FAO., “How to feed the world in 2050”, 2009, http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf(Reference 2017.8.3)

[6] Lebourgeois, V., Dupuy, S., Vintrou, É., Ameline, M., Butler, S., Bégué, A., “A Combined Random Forest and OBIA Classification Scheme for Mapping Smallholder Agriculture at Different Nomenclature Levels Using Multisource Data (Simulated Sentinel-2 Time Series, VHRS and DEM)”, Remote Sens, 9, no. 3,259, 2017, pp.20

[7] Liaw, A, Wiener, M, “Classification and regression by Random Forest”., RNews 2/3,2002, pp.18–22. ftp://131.252.97.79/Transfer/Treg/WFRE_Articles/Liaw_02_Classification%20and%20regression%20by%20randomForest.pdf

[8] MAFF., “Forestry and Fisheries crop survey (vegetables) H22 Vegetable Production and Shipment Statistics”, 2011,

http://www.maff.go.jp/j/tokei/kouhyou/sakumotu/sakkyou_yasai/(Reference 2017.8.3)

[9] MAFF., “Changes in the import volume (fresh) by current products over processing and business vegetables” 2015, http://www.maff.go.jp/j/seisan/kakou/yasai_kazitu/pdf/kg-yasai.pdf (Reference 2017.8.3)

[10] Matsui, T., Ugata, Machimura, T., “A development of factor analyzing and predicting model of abandoned agricultural land with machine learning algorithms”. Journal of Japan Society of Civil Engineers, Ser. G. Vol70, No.6, 42,2014, pp.131-139.

[11] Yamaya, Y., Sonbe, R., Tan, H., Wang, X., Kobayashi, N., Mochizuki, K., “Crop Classification by Random Forest Using TerraSAR-X Data”, Memoirs of the Research Faculty of Agriculture, Hokkaido University,34,2017, pp.1-11

[12] Chlingaryan, A. and Sukkarien, S. et al., Machine Learning Approaches for Crop Yield Prediction and Nitrogen Status Estimation in Precision Agriculture: A Review, Computers and Electronics in Agriculture 151,2018, pp.61-69

[13] Kamilaris, A. and Kartakoullis, A. et al., A Review on the Practice of Big Data Analysis in Agriculture, Computer and Electronics in Agriculture 143,2017, pp. 23-37.

[14] Kamilaris, A. and Francesc X. et al., Deep Learning in Agriculture: A Survey, Computer and Electronics in Agriculture 147,2018, pp.70-90.

[15] Li, N. and Maezawa, S., Data Mining Efficiency in Auctions on the basis of Random Forest Algorithm: A Case Study on Producer wholesale market in Atsumi Area, Aichi Prefecture, Japanese Journal of Rural Economics (forthcoming).

[16] Matsui, T. and Ugata et al., A development of factor analyzing and predicting model of abandoned agricultural land with machine learning algorithms, Journal of Japan Society of Civil Engineers 170,2014, pp.131-139.

[17] Yamaya, Y. and Sonbe, R. et al., Crop Classification by Random Forest Using TerraSAR-X Data, Memoirs of the Research Faculty of Agriculture, Hokkaido University 34,2017, pp.1-11.

[18] Ganasi. and Sakurai, S., “Research on Specialized Farmers’ Cooperatives in the Region of Inner Mongolia and its Members: Based on the Cases of Y and S Specialized Farmers’ Cooperatives”, Agricultural Marketing Journal of Japan, 24-1, 2015, pp. 1-11 (in Japanese)

[19] Liu, L. and Yoshinaga, K., “Evaluation by Farmers on the Chinese Farmer’s Professional Association: The case study in Changsha & Huarong, Hunan Province”, Journal of Rural Planning Association, 30, 2011, pp. 237-242 (in Japanese)

[20] Ma, J., Kobayashi, H., Taniguchi, K., and Sato, T., “Study of Development of Farmers' Cooperative Organization and Farm Management for Rice Production in Northeast of China”, Journal of Rural Issues, 44-2, 2013, pp. 53-63 (in Japanese)

[21] Narita, T. and Sui, S., “The Present Condition and the Problem for Farmers’ Professional Cooperative Based on a Group of Merchants: Case of Cherry Production Area of China”, Journal of Rural Society and Economics, 31-1, 2013, pp. 18-24 (in Japanese)

[22] Tuya, “The Present Status of Development and Countermeasure Suggestion of Specialized Herders’ Cooperative in Grassland Area”, Northern Economy, 15, 2011, pp. 45-47 (in Chinese)

[23] Wang, F., “The Status, Problems and Countermeasures of Pasturing Area in Baotou City”, Journal of Anhui Agricultural Sciences, 20, 2013, pp. 8752-8753, 8755 (in Chinese)