IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 1, 2021, accepted December 7, 2021, date of publication December 10, 2021,
date of current version December 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3134488

Automated Zone Identification for Variable-Rate Services in Precision Agriculture

JIWEI XU[1], NESTOR VELASCO BERMEO[1], MENGYA ZHENG[1], DAVID LANGTON[2], MICHAEL O'GRADY[1], (Senior Member, IEEE), AND GREGORY M. P. O'HARE[3], (Member, IEEE)

[1]School of Computer Science, University College Dublin, Dublin 4, D04 V1W8 Ireland
[2]Origin Enterprises PLC, Dublin 24, D24 DCW0 Ireland
[3]School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, D02 PN40 Ireland

Corresponding author: Jiwei Xu (jiwei.xu@ucd.ie)

This work was supported in part by the SFI Strategic Partnerships Programme under Grant 16/SPP/3296, and in part by Origin Enterprises PLC.

ABSTRACT Varying the rate of application of agronomic inputs generates many positive economic and environmental impacts. Increasingly, technologies that enable variable rate application are becoming a distinctive feature of precision agriculture. Nonetheless, a prerequisite, and crucial challenge, remains the optimal and operational designation of distinct application zones for differing agronomic operations. Core to this challenge is the conflation and fusion of diverse data sources ranging from satellite imagery to real-time in-situ data from farms. At present, zones for variable rate application are often defined manually by agronomists and farmers. This paper proposes a novel methodology for the automatic definition of zones for variable rate application. This approach comprises multi-dimensional spatio-temporal data integration methods, clustering-based data classification and a zone creation and representation procedure. In this way, the harmonization of heterogeneous data sources, augmented with different clustering algorithms, enable the delineation of management zones and subsequent construction of maps for potential variable rate applications. Experimental results confirm the effectiveness and efficiency of the proposed approach.

INDEX TERMS Precision agriculture, variable rate application, clustering, data fusion.

I. INTRODUCTION

Precision agriculture is being continuously invigorated through the adoption of advanced ICT technologies, including the Internet of Things (IoT) [1], Cloud computing [2], 5G [3], data analytics, amongst others. Increasingly, intelligent agricultural machinery can interact with Cloud services in near real-time [4], interpreting contextual information acquired from real-time mobile sensors and the Global Navigation Satellite System (GNSS). Such interactions enable *on-demand* intelligent services such as Variable Rate Applications (VRA). Examples of such applications include fertiliser, crop protection and growth regulator application. Applying such products in a fine-grained manner to zones within fields on an *as-needed* basis contributes to yield improvement and cost-effectiveness while enabling a more environmentally sustainable model of agriculture. Fundamental to VRA is the need for a plan. Such a plan will align with farmer objectives and enable the automated application of inputs according to the characteristics of the point of application. For example, variable-rate seeding can be used to help optimise crop canopy and yield potential across the field. However, calculating how the seeding is applied demands detailed consideration of a wide range of factors by farmers or agronomists.

Effective VAR planning demands the resolution of two issues in the first instance. Firstly, different spatio-temporal traits from diverse data sources must be harmonised. On the spatial scale, examples include point data, for example, historical yields, or raster data, for example satellite imagery. On the temporal scale, the data collection intervals could range from hourly, for example, weather data, to daily, monthly, or even yearly, for example, yield data in the latter instance. A uniform spatio-temporal dimension for all data is required. Secondly, application zones must be internally coherent yet externally different. Thus the objectives for this work are:

The associate editor coordinating the review of this manuscript and approving it for publication was Kashif Sharif.

1) How to integrate multi-dimensional spatio-temporal data for consistent feature extraction?
2) How to classify the data for coherent zone creation according to the different data features available?
3) How to define zones in an interoperable standard such that they can be interpreted consistently for use by diverse agricultural machinery?

The contribution of this paper is as follows.

1) A set of grid-based methods is defined by which to integrate different kinds of data. These methods include *dense point data integration*, *sparse point data integration*, and *raster data integration* in the spatial dimension. In the temporal dimension, the method is a *weighted mean value integration*.
2) k-means clustering methods and Expectation Maximization (EM) clustering methods are harnessed to group data; both clustering methods are compared through experimentation.
3) An n-looking forward graph traverse method is defined to generate the application zone. Zones are then cast as standard geospatial polygon formats - GeoJSON and Shapefile, thereby ensuring interoperability.

The remainder of this paper is organised as follows: Section II introduces some preliminary terms and definitions. Section III proposes the grid-based integration methods for different kinds of data. Section IV details the clustering methods for data grouping. Section V presents the graph traverse method for zone creation. Section VI presents a set of experiments to test the effectiveness and efficiency of our approach. Finally, Section VII considers some future research directions and concludes the paper.

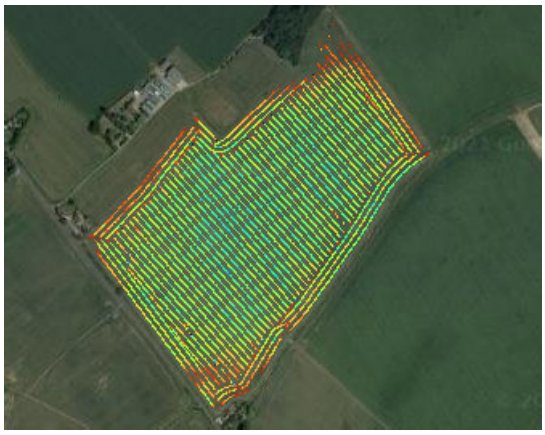


FIGURE 1. Dense point data.

II. TERMINOLOGY

This section briefly considers three different kinds of data synonymous with smart agriculture practice. A simple initial categorisation for data is that of point data and raster data. Point data may be regarded as a tuple of a geographic coordinate (GPS) and corresponding data values. Depending on the density of the points, point data can be further categorised into dense point data or sparse point data. Raster data is typified by data collected by wide-range scanning devices such as satellite, radar or drones. In contrast to point data, raster data consists of data values of a geographic area. However, the quality and usefulness of such data depend upon its resolution.

FIGURE 2. Sparse point data.

A. DENSE POINT DATA

Data with a distance interval of acquisition of less than N is defined as dense point data. There is a minimum of two data points located in any $N \times N$ grids. A typical example of dense point data is yield map data. Yield data is recorded by a yield monitor on a harvester when crops are being harvested. Each point reflects the yield output of a particular area of the field. Many benefits accrue from the correct use of yield maps. Optimizing sustainable production has significant economic, environmental and social impacts. Globally, food security is increased. At the farm level, problems resulting from geomorphological conditions, for example, poor soil nutrients and uneven water availability, may be quickly identified and remedial strategies identified. Thus, a yield map may be used as a key data feature for planning variable-rate seeding and variable-rate fertiliser application. Figure 1 illustrates a sample yield map. As can be seen, the yield map consists of a set of densely distributed data points. The route of the harvester is also visible.

B. SPARSE POINT DATA

In contrast to the dense point data, sparse point data refers to data where the distance interval of acquisition is larger than N. In this case, if $N \times N$ grids covered a field, some of the grids will have zero data points covered. Soil's electrical conductivity (EC) is an example of such data. EC data measures the electrical conductivity of the soil which is a key indicator of soil health that can reflect the soil moisture, the soil texture, and the content and utilisation of soil nutrients. Thus, it can serve as a feature for guiding the variable-rate fertiliser application and variable-rate seeding, among others. Figure 2 illustrates a map of the EC data of a field. As can be

seen, an EC data map consists of a set of sparsely distributed data points.

C. RASTER DATA

Raster maps convey values for geographic areas. For example, Normalized Difference Vegetation Index (NDVI) and Green Chlorophyll Vegetation Index (GCVI) maps are two commonly used images in smart agriculture. The NDVI is calculated by $NDVI = (NIR - Red)/(NIR + Red)$, while NIR represents the band of near-infrared light and the Red represents the band of red light. The GCVI is calculated by $GCVI = (NIR/Green) - 1$, while $Green$ represents the band of green light. NDVI can be used in monitoring the crop growth, while GCVI can be used to show the chlorophyll intensity. Figure 3 and Figure 4 illustrate the NDVI and GCVI of a field on 2018/05/09.

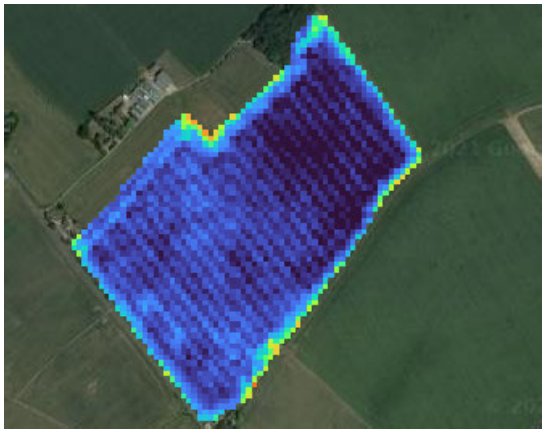


FIGURE 3. Raster data - NDVI (2018/05/09).

FIGURE 4. Raster data - GCVI (2018/05/09).

III. GRID-BASED SPATIO-TEMPORAL DATA INTEGRATION

In considering the different spatial and temporal granularity of archetypical agriculture data, a uniform grid can be adapted that ensures the same spatio-temporal granularity across data categories is now defined.

$N \times N \times T$ grids are harnessed to integrate the different data categories; here, N represents the granularity of spatial dimension, and T represents the granularity of the time dimension. The grids are then applied to a target area, for example, a field and the approximate data value for each grid is calculated. For each category of data, a different calculation is necessitated.

A. DENSE POINT DATA INTEGRATION

In a dense point data map, there are at least one or more data points located in one grid. Here, the average value of the data points located in the grid is used to approximate the value of the grid. Equation (1) calculates the value of the grid:

$$V = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{1}$$

where P_i represents the value of i^{th} point located within the grid.

B. SPARSE POINT DATA INTEGRATION

A grid may contain one or more data points in a sparse point data map, or perhaps no data point. If a grid has data point(s), equation (1) is harnessed to calculate the value of the grid. If a grid has no data point, the weighted average value of the n closest data points to the grid centre is used to approximate the value of the grid. This calculation is illustrated in (2):

$$V = \frac{\sum_{i=1}^{n} W_i P_i}{n-1} \tag{2}$$

where P_i represents the value of i^{th} closest point, and W_i represents the weight of P_i.

Here the distance is used as the factor of the weight. Its calculation is as follows (3):

$$W_i = 1 - \frac{Dis(P_i, C)}{\sum_{j=1}^{n} Dis(P_j, C)} \tag{3}$$

where $Dis(A, B)$ represents points A and B's physical distance.

The basis for this calculation is that the value of this category of data, as exemplified by EC and soil moisture, is spatially continuous. However, this calculation may not be applied to data for which there is no explicit spatial continuity. In such cases, alternate weighting criteria would be used.

C. RASTER DATA INTEGRATION

A raster may be regarded as a small grid. Thus, a grid may cover several rasters. The average value of the raster represents the value of the grid. The calculation is as (4):

$$V = \frac{\sum_{i=1}^{n} D_i * S_i}{\sum_{i=1}^{n} S_i} \tag{4}$$

where D_i represents the value of raster i and S_i represents the area of the i^{th} raster covered by the grid.

D. TEMPORAL INTEGRATION

If the interval between data collection iterations is smaller than T, there must be more than one data value in the time grid. For example, the satellite image collection period is usually of the order of 2 or 3 days. If the time grid granularity is set at 10 or 15 days, there would be typically 3-6 satellite data instances of time grid data. The average value of such data is calculated and averaged as the grid data. This integration is formalised by (5).

Assume that there are n available data within the time range $[x, x+T)$ and are represented as $\gamma_1, \ldots, \gamma_n$, the value of the grid can be calculated by:

$$V = \frac{1}{n}\sum_{i=1}^{n}\gamma_i \tag{5}$$

Strategically, data integration is the pre-processing of all the data needed to guide the variable rate application. The net result is a four-dimensional data matrix M. The first dimension is the data type, such as NDVI, EC, YieldMap, etc. The second dimension is time. The third dimension is longitude, and the fourth dimension is latitude.

IV. CLUSTERING-BASED DATA CLASSIFICATION

The application rate is informed by the value of different data metrics. Thus, the metrics data must be classified into different degrees before identifying the application rate and the application zone. To complete this classification, the following steps must be completed.

A. ATTRIBUTE SET PREPROCESSING

1) DIMENSION REDUCTION

Since all the data is integrated into a 4D matrix M, the particular data or data combination needed for agronomic operational decision-making must be a slice of M. This slice is termed the attribute set A. For calculation convenience, A is reduced into a two-dimensional matrix. To achieve this reduction, the time dimension is removed and the two spatial dimensions are serialised.

a: TIME DIMENSION REMOVAL

All the operations are related to a specified time range of the data. As the time range is the temporal grid size of M, the corresponding column data to $M[:, j, :, :]$ is fetched. Now, the attribute set becomes three-dimension.

b: SPATIAL DIMENSIONS SERIALIZATION

The spatial dimensions refer to latitude and longitude. n represents the column counts of latitude, and r and c describe the indices of the longitude and latitude dimension of M. Then $p = r * n + c$ is used to serialise the spatial grid index. Thus $M'[:, :, p] = M[:, :, r, c]$. The r and c can also be calculated reversely by (6).

$$\begin{aligned} r &= p/n \\ c &= p\%n \end{aligned} \tag{6}$$

If the selected data types are in the set S, then the attribute is represented by (7).

$$A = \bigcup_{i \in S} M'[i, j, p] \tag{7}$$

2) NORMALIZATION

The attribute set A can be treated as the subset of M. For each data type, the statistical index of M_i is used for normalisation instead of A_i. Z-score normalization is used to process each kind of data. The normalized value is then calculated by (8).

$$z = \frac{x - \mu}{\sigma} \tag{8}$$

where x is the value of a grid, and μ and σ represent the mean and standard deviation of M_i. The normalized value range is $[-1,1]$.

B. CLUSTERING

Clustering is the key step in partitioning a field into different zones based on the attribute set A. Each row of A has n attribute values $(x_1, x_2, \ldots, x_n)$. When clustering, different attributes have different weights $(w_1, w_2, \ldots, w_n)$. Three different clustering methods are now considered, k-means and two Expectation–Maximization (EM) algorithms based on different models.

1) k-MEANS CLUSTERING

Classical k-means partitions A into k clusters in which each item belongs to the cluster with the nearest mean; thus, within-cluster variances are minimised. In this case, all the attributes have the same weights. Here, different weights are set for different attributes. Then, it can be represented by (9).

$$\underset{S}{argmin} \sum_{i=1}^{k}\sum_{x \in S_i}\sum_{j=1}^{n} w_j * (x_j - \mu_{ij})^2 \tag{9}$$

The calculation is as follows: First, randomly select k items as the centroids of k groups. Then, divide all items into the k groups according to the weighted distance and select the centroid for each group. Repeat this process until all groups do not change.

2) EM CLUSTERING

Expectation-maximization (EM) algorithm alternates between two steps, the expectation (E) step and the maximisation (M) step. For clustering, EM uses probabilistic models and estimates a set of parameters iteratively to reach convergence. The probabilistic model is defined as a set of k probability distributions, and each distribution corresponds to one cluster. An item is assigned to a particular cluster based on its membership probability. EM clustering follows three steps: 1. Guess the initial parameters. 2. Iteratively refine the parameters with E and M steps until convergence is reached. 3. Assign each item to the cluster with which it has the highest membership possibility.

Here, two different probabilistic models are harnessed - Gaussian Mixture Model (GMM) and Variational Bayesian Gaussian Mixture Model (VBGMM). GMM assumes all the data points are generated from a mixture of k Gaussian distributions with unknown parameters. It uses the maximum likelihood estimation method for parameter estimation. VBGMM maximises a lower bound on model evidence instead of data likelihood.

Unlike k-means and GMM, VBGMM is able to find the optimal value for k. It can infer an approximate posterior distribution over the parameters of a Gaussian mixture distribution by implementing the Dirichlet distribution model.

V. ZONE CREATION AND REPRESENTATION

Clustering methods used in the classification phase are geography independent. The clustering results may be mapped into a two-dimensional geocoded map using Equation (6). This geocoded map, composed of $N \times N$ grids, is the minimum bounding rectangular area that covers an entire field. A 2D array G represents the rectangle with lengths equal to the rows and columns of the grids. The elements of G have different values. Two kinds of elements exist in G, within-the-field and out-of-the-field. As a result of clustering, elements within the field have already been classified into different clusters; their values range from 1 to k. Those elements out of the field are set to -1. The initialization can be formalized as (10), where $P_{r,c}$ represents the grid of row r and column c, and C_x represents the x^{th} cluster.

$$G_{i,j} = \begin{cases} -1, & P_{r,c} \notin \bigcup_{x=1}^{k} C_x \\ x, & P_{r,c} \in C_x \end{cases} \tag{10}$$

A. n-FORWARD DEPTH-FIRST SEARCH

A depth-first search merges all the adjacent grids within the same cluster into a polygon. However, there may be many isolated points in the map, resulting in excessively fragmented zones. To address this issue, we proposed an n-forward depth-first search algorithm (n-DFS). n-DFS searches eight neighbours clockwise to find grids within the same cluster. All the searched grids are set to 0. For the up, down, left, and right directions, n-DFS performs a look forward n operation which means it not only checks the proximate grids but searches n more grids in these four directions. If the m^{th} ($m \leq n$) grid has the searching value, the grids before it are also marked within the same cluster. However, the n-looking forward process will stop when meeting a 0 or -1.

B. GRID MERGE AND STANDARDIZATION

In geospatial data, vectors composed of GPS coordinates are essential to express geographical features by considering those features as geometrical shapes. Different types of geometry describe various geographical features: point, line and polygon. However, a grid is essentially a polygon with four GPS coordinates. Fields were split into small grids for data integration and grid values used as an elementary unit for clustering. However, there is much redundant information when using grids to represent the clustering result.

To simplify information presentation, all grids are merged within the same cluster into a polygon. Two benefits accrue from this. First, adjacent grids have duplicated coordinates which can be avoided in polygon representation. Second, the polygon need only store the intersecting points, while grids need to keep all vertices.

The standardised vector representations include inter alia well-known text (WKT) and its binary equivalent well-known binary (WKB), GeoJSON and Shapefile. To maintain consistency of service delivery through RESTFul APIs, GeoJSON is harnessed as the export format since it is JSON-based. The structure is defined as Figure 5. The clustering result is represented as a FeatureCollection type, and each feature specifies a cluster with both properties and multi-polygon geometry. One can add further properties that can be recognised by agricultural machinery to achieve the VRA.

```
{
    "type":"FeatureCollection",
    "features":[
        {
            "type":"Feature",
            "properties":{
                "average_yield":7.51
            },
            "geometry":{
                "type":"MultiPolygon",
                "coordinates":[
                    ...
                ]
            }
        },
        ...
    ]
}
```

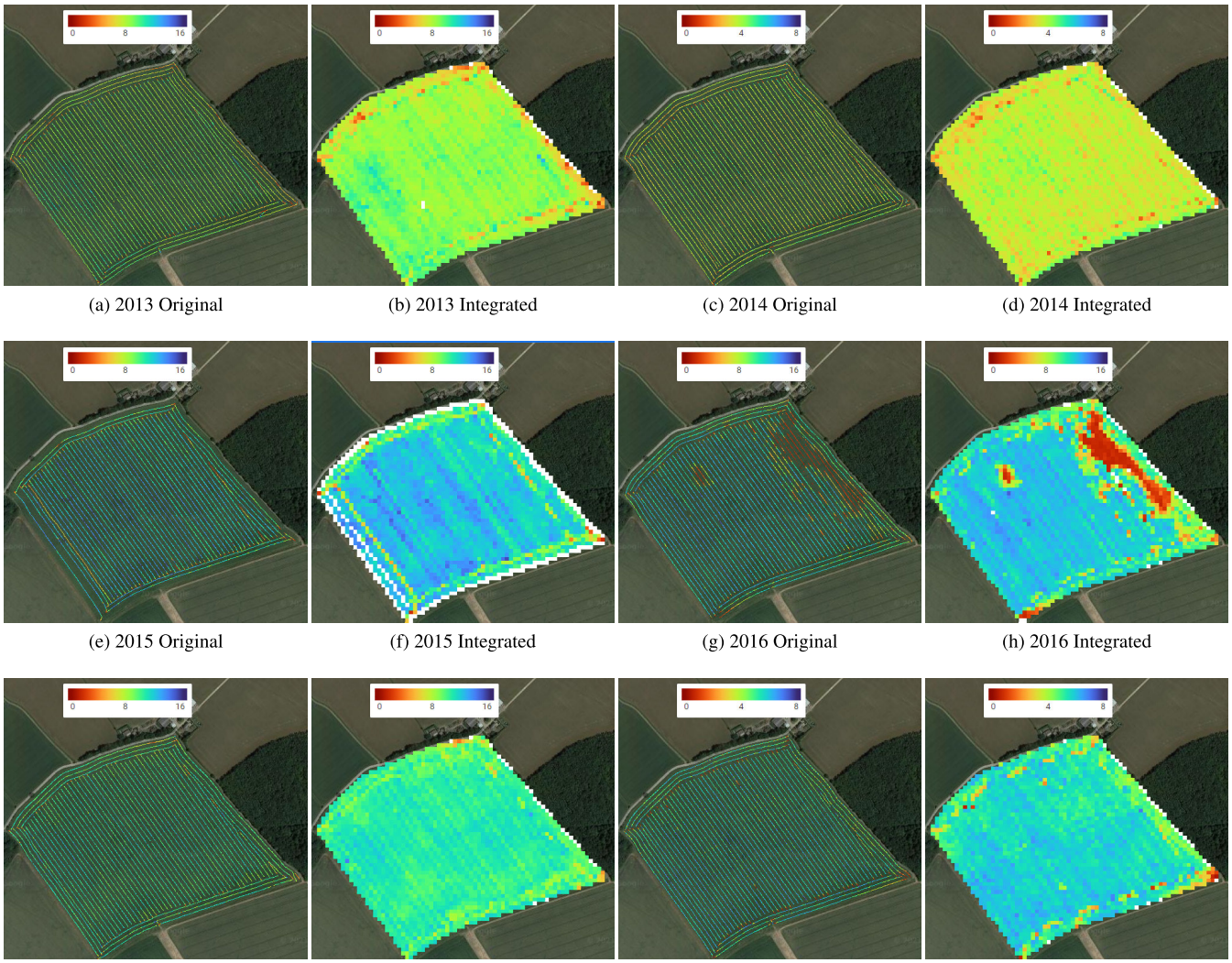
FIGURE 5. GeoJSON Structure.

VI. EXPERIMENTS

A. ENVIRONMENT SETUP

Real-world in-situ data (EC data and yield data) together with satellite imagery were used to assess the feasibility and reliability of the map generation method. EC data and yield data are common to most arable farms; in this case, it was sourced from a tillage farm in the South of England. Open and freely available satellite imagery was acquired from the Copernicus Sentinel mission. For this experiment, post-processed satellite imagery was accessed through the Sentinel Hub APIs [5].

The software stack consists of MongoDB v4.4.1 Community, Python v3.9, Node.js v14.7, Docker v20.10.5 and related libraries. MongoDB is used to address data storage, Python 3.9 is used as a vehicle for clustering and geographic data processing, Node.js is accountable for service delivery with RESTFul API, and Docker containers host all services. The services are deployed on a server with 4 Intel Xeon Processor (Skylake, IBRS) CPUs and 32GB memory.



(a) 2013 Original (b) 2013 Integrated (c) 2014 Original (d) 2014 Integrated

(e) 2015 Original (f) 2015 Integrated (g) 2016 Original (h) 2016 Integrated

(i) 2017 Original (j) 2017 Integrated (k) 2018 Original (l) 2018 Integrated

FIGURE 6. Yield Map Data (Dense Data) Integration.

TABLE 1. Yield map clustering statistical information.

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
k-means	mean	12.07597307	11.06742169	9.044257747	6.197204651	1.958473492
	var	0.127134425	0.149735022	0.460551989	1.103858465	1.240255526
	count	908	808	306	166	191
GMM	mean	11.83329336	10.61558353	8.58727914	5.481242613	1.903047132
	var	0.207315216	0.11704342	0.770160079	0.754811019	1.14954129
	count	1367	388	331	107	186
VBGMM	mean	11.66973615	9.81572856	7.491041579	2.287945989	0
	var	0.312524169	0.194410453	1.350342299	1.792100801	-
	count	1626	261	272	220	0

Spatial granularity was set at 10 meters, meaning that the spatial grid size is 10×10. The temporal granularity is set to bimonthly, meaning that there are two values for each month. The first value covers from 1^{st} to 15^{th}, while the second value covers the remainder of the month.

B. DATA INTEGRATION

Both point data and raster data must be integrated. However, the raster data is inherently grid-based; hence it will not look much different after integration. Point data integration is now considered.

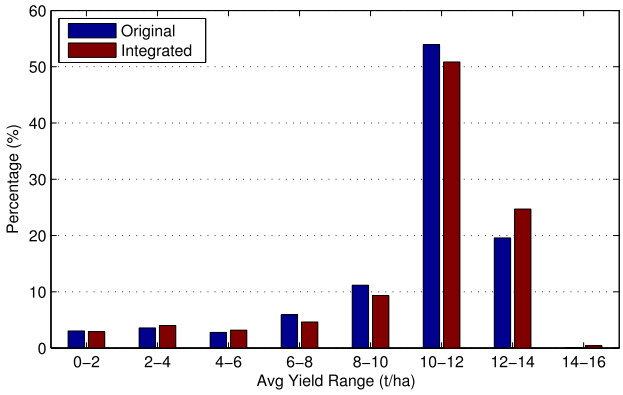
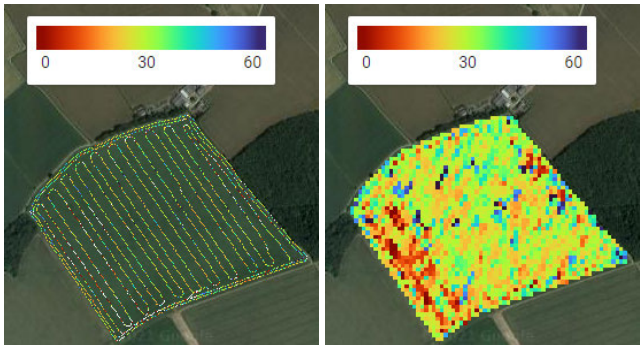
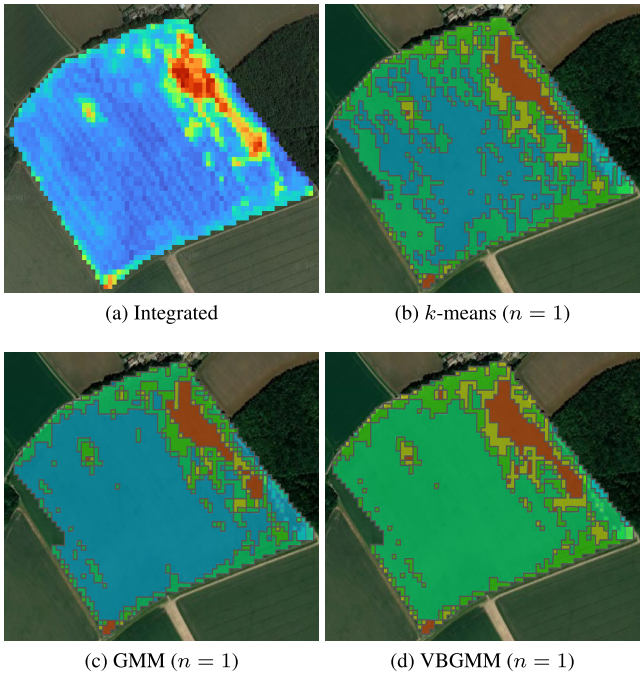


FIGURE 7. Data Distribution of 2016 Yield Map.



(a) Original (b) Integrated

FIGURE 8. EC Data (Sparse Data) Integration.



(a) Integrated (b) k-means ($n = 1$)

(c) GMM ($n = 1$) (d) VBGMM ($n = 1$)

FIGURE 9. NDVI clustering with different algorithms.

The Dense Point Data Integration (1) method is harnessed to integrate six years of yield map data for the same field. Figure 6 visualises both the original data and integrated data.

The field was planted with winter wheat for three years, then beans for one year (years 2014 and 2018) for crop rotation purposes. Since the average yield of beans is less than winter wheat, the legends of 2014 and 2018 are from 0 to 8, while others are from 0 to 16. It can be observed that the integration method preserves the original data features and distributions very well. There are a few white grids in some edge areas where the value is 0. These grids have minimal overlap within the field and no original data located within them. However, more white grids occur in non-edge areas in the year 2015. This might be a data collection issue, for example, a faulty yield monitoring sensor but more data is needed to infer the actual reason.

Data distribution for 2016 is evidently visible across the figures. Figure 7 illustrates the result. The x-axis displays the data value range, and the y-axis represents the percentage of a value range. From this figure, it can be seen that the data distribution trend after integration is pretty much as the original. The percentage of the greatest concentration range [10, 12] changes from 53.9% to 50.9% and the second greatest concentration range [12, 14] changes from 19.6% to 24.7%. Although there is a slight shift in data distribution, the overall trend is pretty much similar with the largest change less than 5%.

In assessing sparse data integration, soil EC is utilised. Figure 8 illustrates a comparison of original data and integrated data. The selected data set is the shallow measurement (30 centimetres) of EC. This data was produced in 2013, and there are no subsequent measurements available for the field under consideration. As can be seen from the colour change, the weighted average value calculation (2) is reliable in integrating sparse data.

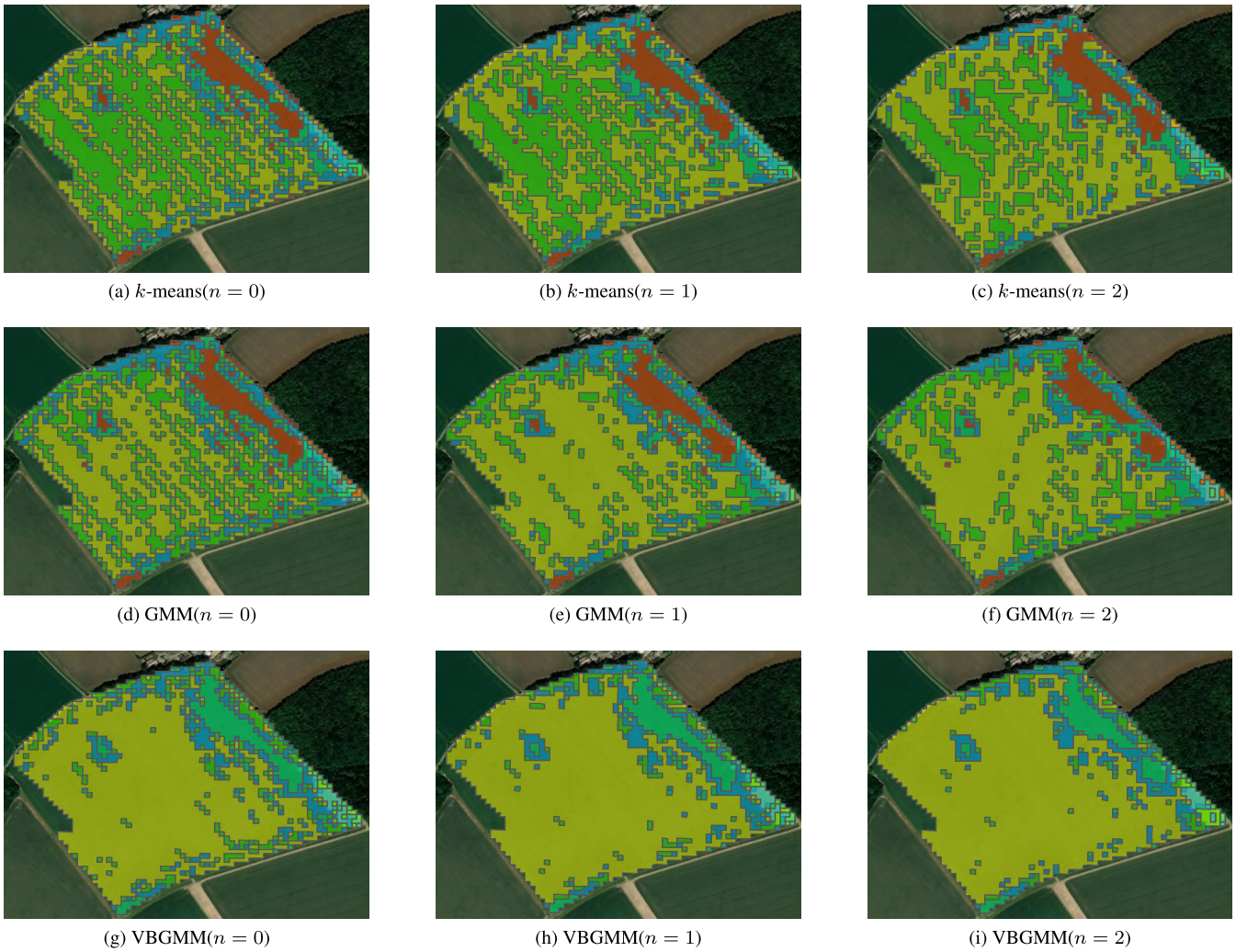
C. CLUSTERING

This section assesses the effectiveness of the clustering and n-DFS methods. The experiments are based on a yield map data set from 2016 and a corresponding satellite NDVI image of 2016/07/06. The number of clusters is set to 5, which means $k = 5$. Firstly, the k-means, GMM and VBGMM are applied to the NDVI image with $n = 1$. Secondly, the three algorithms are applied separately to the yield data set, and different n values of n of 0, 1, and 2 are used. Descriptive statistics for each resultant cluster - mean value, variance, and items count, were captured for each clustering algorithms.

1) NDVI CLUSTERING

Figure 9 illustrates the integrated NDVI image for before harvest (2016/07/06) and the clustering results of the three different algorithms. Figure 9(a) shows the integrated image with the yield map of 2016. A high correlation between satellite images and crop growth is clearly visible. Note that the value range of this NDVI image is between 0.46 (red) and 0.75 (blue).

Figure 9(b) shows the k-means clustering result, Figure 9(c) shows the GMM clustering result, and Figure 9(d) shows the VBGMM clustering result. The experimental



(a) k-means($n = 0$) (b) k-means($n = 1$) (c) k-means($n = 2$)

(d) GMM($n = 0$) (e) GMM($n = 1$) (f) GMM($n = 2$)

(g) VBGMM($n = 0$) (h) VBGMM($n = 1$) (i) VBGMM($n = 2$)

FIGURE 10. Yield map clustering with *k*-means, GMM and VBGMM algorithms and *n*.

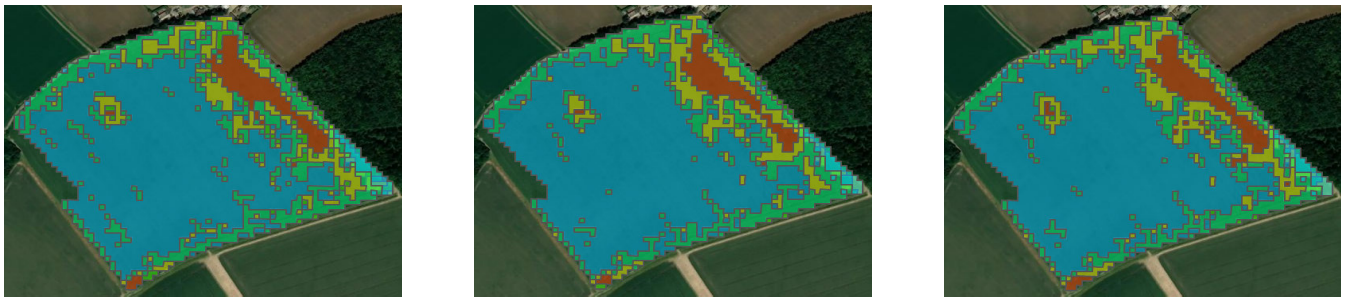
results show some differences between the three algorithms. However, all three algorithms can recognise the visible differences. For example, the red area is almost the same in all three figures.

2) YIELD MAP CLUSTERING

Figure 10 displays the clustering results of yield map data of 2016, and Table 1 shows the corresponding statistical information. Comparison is possible between these clustering results with Figure 6(g) and Figure 6(h) since they are all based on the same data set. From Figure 10, it can be seen that even though different clustering algorithms give different results, significant differences are clearly distinguishable from each other. In comparison with the other two algorithms, the k-means provides more details. The cluster item count data in Table 1 shows that the k-means tends to distinguish maximum values from the majorities, while VBGMM is on the opposite end of this continuum and the GMM is situated between the approaches. From Table 1, it can be seen that cluster 1 has the mean values of 12.08, 11.83 and 11.67 separately, and items' counts are 908, 1367 and 1626, and cluster 2 has the mean values of 11.07, 10.61 and 9.81 separately, and items' counts are 808, 388, 261.

Reflecting upon the figures, the k-means clusters are more dispersed, and the VBGMM clusters are very concentrated. Note that the VBGMM only returns four groups and one empty group when the k is set to 5. From the perspective of variance, GMM is relatively concentrated. Its minimum variance is 0.21, and the maximum variance is 1.15. In comparison, k-means' minimum variance is 0.13, and maximum variance is 1.24, while VBGMM's minimum variance is 0.31 and maximum variance is 1.79. However, we cannot generally tell which one is better since it may be suitable for different situations. The correlation of algorithms to suitable situations will form the basis of future work.

One can also obverse the affection of n in zone generation from Figure 10. The purpose of n is to reduce isolated points, thereby improving adaptation for use in VRA. From the figures, the isolated points reduce as n increases. However, the search sequence of clusters can affect the result.



(a) k-means($n = 1$) (b) GMM($n = 1$) (c) VBGMM($n = 1$)

FIGURE 11. Mixture data clustering with different algorithms.

Conclusions can be inferred from the Figures: 1) $n = 2$ obtained the best result when using k-means clustering. If n is less than two, excessive fragmentation rendering the resultant map unusable in agricultural machinery may result. 2) When $n = 1$, an acceptable result using GMM clustering is obtained. 3) A value of $n = 0$ is good enough for VBGMM clustering, but a larger value may cause a loss of edge details.

D. MIXTURE DATA CLUSTERING

As discussed in Section II, both yield map and satellite image could guide variable-rate seeding and variable-rate fertilising. Besides, NDVI can also inform the variable-rate application of many products, such as pesticides and growth regulators. However, other data such as topology and weather also affect these operations. Thus more comprehensive data is needed for each VRA activity.

Here, a combination of yield data and NDVI data form a basis for clustering. Both the satellite imagery and the yield map are considered correlated data. Satellite data can be used to estimate yield. The closer the acquisition time of the satellite image is to the actual harvest date, the more accurate the prediction of yield. Under normal circumstances, the weight matrix will differ according to the different growth stages. However, in this case, the satellite image was captured on the 6th July 2016 when the crop was approaching maturity. Thus, equal weights are set for both datasets. In zone merging, the n is set to 1. The clustering results are illustrated in Figure 11. From this figure, it can be seen that the three clustering algorithms have very similar results, indicating that a mixture of data may help in eliminating data bias.

Comparing to Figure 10 and Figure 11, it is concluded that a mixed-use of correlated data could improve the clustering reliability and stability in VRA. This is easily understood when it is considered that sensor data usually exhibit measurement errors. Such errors can be reduced by multiple measures according to statistical theory. Even in cases where sensors can only measure once, a judicious mixed-use of different sensor data can serve as an effective proxy for multiple measurements.

VII. RELATED WORK

Precision agriculture is widely envisaged as a key application scenario of the Internet of Things (IoT) going forward [6], [7]. Moreover, increased integration with other advanced ICT technologies such as artificial intelligence and Cloud-enabled weather services will also become a popular trend [8]–[11]. Progress has been made in smart farm and VRA technologies, such as smart irrigation [12] and smart fertilizer [13]. However, those works emphasise the importance of sensors but with single sources of data.

Grisso *et al.* [14] see VRA as an essential tool of precision farming, and they summarised the VRA methods and systems in their work. Later work [15], [16] demonstrates that VRA can result in environmental and economic benefits. Zhang *et al.* [17] developed a web-based tool called ZoneMAP to map zones according to satellite images or user input data for VRA.

Yield mapping is an increasingly important tool for smart agriculture, enabling consideration of both quality and quantity of harvests [18]. However, capturing accurate yield data is fraught with difficulty. Yield monitors mounted on harvesters may be inaccurate due to calibration errors; for example, cut-width, lag-time, and header settings may be incorrectly specified [19]. Sensor drift is also a potential source of error. A variety of approaches have been proposed to improve the accuracy of yield maps. Combining in-situ sensor data from the harvester with Earth Observation (EO) data has been proposed [20]. Data may be cleaned using software, either manually or through the use of automated filters [21]. Such filters might incorporate approaches for identifying and removing outlier data [22]. The approach outlined in this research is holistic in that it can be harnessed to conflate diverse data sources such that potential causes of yield variation within a particular field might be better understood, and thereby remedied.

Clustering is fundamental in many data-driven application domains [23]. However, there are many different kinds of clustering algorithms that can be divided into different catalogues, partitioning-based, hierarchical-based, density-based, grid-based and model-based [24]. All the algorithms have advantages and disadvantages and may be suitable

for different situations. Rodriguez *et al.* [25] performed a systematic comparison of 9 well-known clustering methods with artificial normally distributed data. However, the data used for VRA is complicated, and we cannot expect one clustering method can do everything well, hence three different clustering algorithms were considered in this paper as baselines.

This research introduces a novel data integration framework that can integrate different data into a common granularity thus making it possible to opportunistically consume different datasets when clustering. Benefits from utilising standardized information exchange formats accrue enabling the viewing of zone partitions on a suite of smart devices such as smart phones or VR/AR devices [26]. Moreover, the zone map can be seamlessly interpreted by VRA-enabled machinery.

VIII. CONCLUSION

VRA is a crucial dimension of precision agriculture as it allows an on-demand application of agricultural inputs to a specific management zone in a farm or field, resulting in sustainable economic and environmental practices. Automatically identifying application zones based on heterogeneous data is the foundation of effective VRA implementation but certain challenges must be first overcome. The first is to harmonise multi-dimensional spatio-temporal data for consistent feature extraction, and the second is to classify the data for coherent zone creation according to the different data features available. This research proposed novel grid-based methods by which to integrate dense point data, sparse point data, and raster data in spatial and temporal dimensions to resolve the data harmonization problem. A suite of real-world datasets are harnessed to validate the data integration methods. Statistical analyses indicate that the difference between original data and integrated data is less than 5%. To address the second challenge, this research proposed a clustering-based data classification method and an n-looking forward graph traverse method for coherent zone creation. Experimental results indicate different n settings benefit different clustering algorithms, and a mixture of correlation data could improve the clustering reliability and stability. Finally, this research employed internationally accepted, open, standardized information representations that enable interoperability with compatible agriculture machinery. Real-world, heterogeneous data was used throughout thus confirming the efficacy and efficiency of the proposed approach.

Zone creation is only the first step towards achieving automatic decision making for VRA, however. The next step is to identify effective strategies for identifying appropriate application implementation rates. Such rates will obviously depend upon operations being planned – seeding, fertilizer, and so forth. Machine learning techniques offer significant potential. Other potential applications for management zone identification will be further investigated.

REFERENCES

[1] P. Sanjeevi, S. Prasanna, B. S. Kumar, G. Gunasekaran, I. Alagiri, and R. V. Anand, "Precision agriculture and farming using Internet of Things based on wireless sensor network," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 12, Dec. 2020, Art. no. e3978.

[2] V. P. Kour and S. Arora, "Recent developments of the Internet of Things in agriculture: A survey," *IEEE Access*, vol. 8, pp. 129924–129957, 2020.

[3] Z. Ullah, F. Al-Turjman, and L. Mostarda, "Cognition in UAV-aided 5G and beyond communications: A survey," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 3, pp. 872–891, Sep. 2020.

[4] Yara. *N-Sensor ALS to Variably Apply Nitrogen*. Accessed: Sep. 6, 2021. [Online]. Available: https://www.yara.co.U.K./crop-nutrition/farmers-toolbox/n-sensor/

[5] Sentinelhub. *Api*. Accessed: Nov. 6, 2021. [Online]. Available: https://www.sentinel-hub.com/develop/api/

[6] A. Khanna and S. Kaur, "Evolution of Internet of Things (IoT) and its significant impact in the field of Precision agriculture," *Comput. Electron. Agricult.*, vol. 157, no. 1, pp. 218–231, 2019.

[7] N. Ahmed, D. De, and I. Hussain, "Internet of Things (IoT) for smart precision agriculture and farming in rural areas," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4890–4899, Dec. 2018.

[8] Y. Mekonnen, S. Namuduri, L. Burton, A. Sarwat, and S. Bhansali, "Review—Machine learning techniques in wireless sensor network based precision agriculture," *J. Electrochem. Soc.*, vol. 167, no. 3, Jan. 2020, Art. no. 037522.

[9] Y.-Y. Zheng, J.-L. Kong, X.-B. Jin, X.-Y. Wang, and M. Zuo, "CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture," *Sensors*, vol. 19, no. 5, p. 1058, Mar. 2019.

[10] N. Zhu, X. Liu, Z. Liu, K. Hu, Y. Wang, J. Tan, M. Huang, Q. Zhu, X. Ji, Y. Jiang, and Y. Guo, "Deep learning for smart agriculture: Concepts, tools, applications, and opportunities," *Int. J. Agricult. Biol. Eng.*, vol. 11, no. 4, pp. 32–44, 2018.

[11] M. O'Grady, D. Langton, F. Salinari, P. Daly, and G. O'Hare, "Service design for climate-smart agriculture," *Inf. Process. Agricult.*, vol. 8, no. 2, pp. 328–340, Jun. 2021.

[12] L. García, L. Parra, J. M. Jimenez, J. Lloret, and P. Lorenz, "IoT-based smart irrigation systems: An overview on the recent trends on sensors and IoT systems for irrigation in precision agriculture," *Sensors*, vol. 20, no. 4, p. 1042, Feb. 2020.

[13] A. Peerlinck, J. Sheppard, and B. Maxwell, "Using deep learning in yield and protein prediction of winter wheat based on fertilization prescriptions in precision agriculture," in *Proc. Int. Conf. Precis. Agricult. (ICPA)*, 2018, pp. 1–13.

[14] R. D. Grisso, M. M. Alley, W. E. Thomason, D. L. Holshouser, and G. T. Roberson, "Precision farming tools: Variable-rate application," Virginia Cooperat. Extension, VA, USA, Tech. Rep. PUBLICATION 442-505, 2011.

[15] S. Stamatiadis, J. S. Schepers, E. Evangelou, A. Glampedakis, M. Glampedakis, N. Dercas, C. Tsadilas, N. Tserlikakis, and E. Tsadila, "Variable-rate application of high spatial resolution can improve cotton N-use efficiency and profitability," *Precis. Agricult.*, vol. 21, no. 3, pp. 695–712, Jun. 2020.

[16] M. Gatti, M. Schippa, A. Garavani, C. Squeri, T. Frioni, P. Dosso, and S. Poni, "High potential of variable rate fertilization combined with a controlled released nitrogen form at affecting cv. Barbera vines behavior," *Eur. J. Agronomy*, vol. 112, Jan. 2020, Art. no. 125949.

[17] X. Zhang, L. Shi, X. Jia, G. Seielstad, and C. Helgason, "Zone mapping application for precision-farming: A decision support tool for variable rate application," *Precis. Agricult.*, vol. 11, no. 2, pp. 103–114, Apr. 2010.

[18] C. L. Bazzi, M. R. Martins, B. E. Cordeiro, L. Gebler, E. G. de Souza, K. Schenatto, P. L. de Paula Filho, and R. Sobjak, "Yield map generation of perennial crops for fresh consumption," *Precis. Agricult.*, pp. 1–14, Sep. 2021, doi: 10.1007/s11119-021-09855-2.

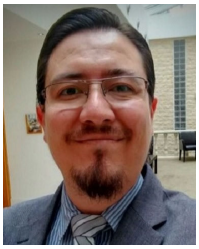
[19] J. P. Fulton. (2015). *Improving Yield Map Quality by Reducing Errors Through Yield Data File Post-Processing*. [Online]. Available: https://www.yara.co.U.K./crop-nutrition/farmers-toolbox/n-sensor/

[20] M. Karampoiki, A. Heiß, G. Sharipov, S. Mahmood, L. Todman, A. Murdoch, H. Griepentrog, and D. Paraforos, "Producing grain yield maps by merging combine harvester and remote sensing data," in *Precision Agriculture'21*. Wageningen The Netherlands: Academic, 2021, pp. 377–402.

[21] K. A. Sudduth and S. T. Drummond, "Yield editor: Software for removing errors from crop yield maps," *Agronomy J.*, vol. 99, no. 6, pp. 1471–1482, Nov. 2007.

[22] L. F. Maldaner, L. P. Corrêdo, T. R. Tavares, L. G. Mendez, C. Duarte, and J. P. Molin, "Identifying and filtering out outliers in spatial datasets," in *Proc. 14th Int. Conf. Precis. Agricult.*, Montreal, QC, Canada, 2018, pp. 24–27.
[23] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39501–39514, 2018.
[24] A. C. Benabdellah, A. Benghabrit, and I. Bouhaddou, "A survey of clustering algorithms for an industrial context," *Proc. Comput. Sci.*, vol. 148, pp. 291–302, Jan. 2019.
[25] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. D. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PLoS ONE*, vol. 14, no. 1, 2019, Art. no. e0210236.
[26] M. Zheng and A. G. Campbell, "Location-based augmented reality *in-situ* visualization applied for agricultural fieldwork navigation," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Oct. 2019, pp. 93–97.

JIWEI XU received the B.S. and M.S. degrees from Wuhan University, in 2008 and 2010, respectively, and the Ph.D. degree in computer software and theory from the University of Chinese Academy of Sciences, in 2016. He currently works as a Postdoctoral Researcher with the School of Computer Science, University College Dublin. His research interests include distributed systems, software engineering, virtualization technology, cloud computing, and the IoT.



NESTOR VELASCO BERMEO received the Ph.D. degree in artificial intelligence in 2014; his research work based on the implementation of Semantic Web technologies to the product lifecycle management. He currently works as a Postdoctoral Researcher with the School of Computer Science, UCD. He has participated in various European H2020 Projects as a Project and Technical Manager. His current work focuses on data models for big data analysis and interoperable intelligent applications.

MENGYA ZHENG received the bachelor's degree in software engineering from University College Dublin, where she is currently pursuing the Ph.D. degree in computer science. Her research explores context-aware AR data visualization metaphors to present explainable decision support data for non-expert users who may lack enough domain knowledge to independently comprehend system-generated advice.

DAVID LANGTON works at Origin Enterprises PLC and has over 33 years' experience in crop agronomy research managing research teams, and he is a member of several U.K. industry body committees. His current focus is on the development of digital decision support tools for Origins customers and advisors. These include crop growth models, pest and disease decision support tools as well as being closely involved in a number of multidisciplinary collaborative projects involving machine learning and big data analysis to derive new insights and develop new decision support tools.

MICHAEL O'GRADY (Senior Member, IEEE) is currently a Senior Research Fellow at University College Dublin. He has over 30 years of experience both in industry and academia; his expertise has been internationally acknowledged by the ACM and IEEE through senior membership, as well as by international funding agencies as an expert reviewer. He has published in a wide range of international fora, contributing to over 176 peer-reviewed publications, resulting in a H-index of 28. His specific research interests include the applicability of intelligent systems in context and opportunistic service provision, as well as participatory action research.

GREGORY M. P. O'HARE (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the University of Ulster. He is currently a Professor of artificial intelligence and the Head of the School of Computer Science and Statistics, Trinity College Dublin. Prior to joining Trinity College Dublin, he was a Professor of computer science at University College Dublin (UCD). He was the Head of the Department of Computer Science, UCD (2001–2004). Prior to joining UCD, he has been on the Faculty of the University of Central Lancashire (1984–86) and The University of Manchester (1986–1996). From 2008 to 2009, he secured a Visiting Research Fellowship at the University of Oxford. In 2010, he was awarded a Fulbright Scholar visiting position at the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT). In 2018, he was a Visiting Research Professor at Queens University Belfast (QUB). He has published over 490 refereed publications in journals and international conferences, seven books, and has won significant grant income (ca €75.00M). He is an established researcher with international repute. His research interests include the areas of multi-agent systems (MAS), mobile and ubiquitous computing, and precision agriculture. He currently is the Lead PI for CONSUS (Crop Optimisation through Sensing, Understanding and Visualisation) SFI Strategic Partnership Programme €17.65 Million.