agriculture

MDPI

Article

Integration of Precision Farming Data and Spatial Statistical Modelling to Interpret Field-Scale Maize Productivity

Guopeng Jiang [1,*], Miles Grafton [1], Diane Pearson [1], Mike Bretherton [1] and Allister Holmes [2]

[1] School of Agriculture and Environment, Massey University, Palmerston North 4410, New Zealand; m.grafton@massey.ac.nz (M.G); d.pearson@massey.ac.nz (D.P.); m.bretherton@massey.ac.nz (M.B.)

[2] Foundation for Arable Research, Christchurch 8441, New Zealand; allister.holmes@far.org.nz

* Correspondence: g.jiang@massey.ac.nz; Tel.: +64-022-198-5863

Received: 30 September 2019; Accepted: 30 October 2019; Published: 4 November 2019

Abstract: Spatial variability in soil, crop, and topographic features, combined with temporal variability between seasons can result in variable annual yield patterns within a paddock. The complexity of interactions between yield-limiting factors such as soil nutrients and soil water require specialist statistical processing to be able to quantify variability, and thus inform crop management practices. This study uses multiple linear regression models, Cubist regression and feed-forward neural networks to predict spatial maize-grain (*Zea mays*) yield at two sites in the Waikato Region, New Zealand. The variables considered were: crop reflectance data from satellite imagery, soil electrical conductivity, soil organic matter, elevation, rainfall, temperature, solar radiation, and seeding density. This exercise explores methods which may be useful in predicting yield from proximal and remote sensed data with higher resolution than traditional low spatial resolution point sampling using soil testing and yield response curves.

Keywords: data fusion; precision agriculture; arable; satellite imagery

1. Introduction

The arable industry is a small sector within the New Zealand farming system, primarily aimed at supplying the domestic requirements for cereals used in the milling, brewing and animal feed industries, and herbage seed production [1]. Maize is a major arable crop grown in New Zealand, producing grain for human or animal consumption, as well as silage as supplementary feed to livestock. It also tends to produce higher yields per hectare compared to other arable products such as wheat and barley, as seen in Figure 1. The Foundation for Arable Research (FAR) is the research organization for the arable sector. This research is funded by FAR, as they are looking for a way to utilize the historical data from precision farming technologies such as combine yield monitors and soil EC sensors to inform crop management decisions such as variable rate seeding and nitrogen applications.

The practice of precision farming in the New Zealand (NZ) arable sector began in the early 1990s. Since then, there has been wide-scale uptake of precision farming tools such as guidance systems and variable-rate irrigation. However, the commercial uptake of variable rate application (VRA) (which has the potential to improve farming efficiency) has been limited due to the lack of available information to estimate yield response [2].

With the increasing availability of regularly captured spatial data and publicly-available satellite imagery and climatic records, there is potential to use this information to inform farm management decisions. However, appropriate spatial analysis techniques are still limited for this type of application. Progress has been made on delineating potential management zones (MZs)

within-paddocks to represent similar yield-limiting factors based on a variety of spatial information (e.g., historical yield data, geo-referenced aerial photographs, soil and topography features) using spatial classification techniques [2–4]. From this, a different rate of an input (e.g., fertilizer, seeding rate) can be applied to each MZ. However, it is difficult to quantify spatial yield and temporal variability without understanding yield potential and crop response to specific variables (e.g., climate, soil type, management practices) [5,6].

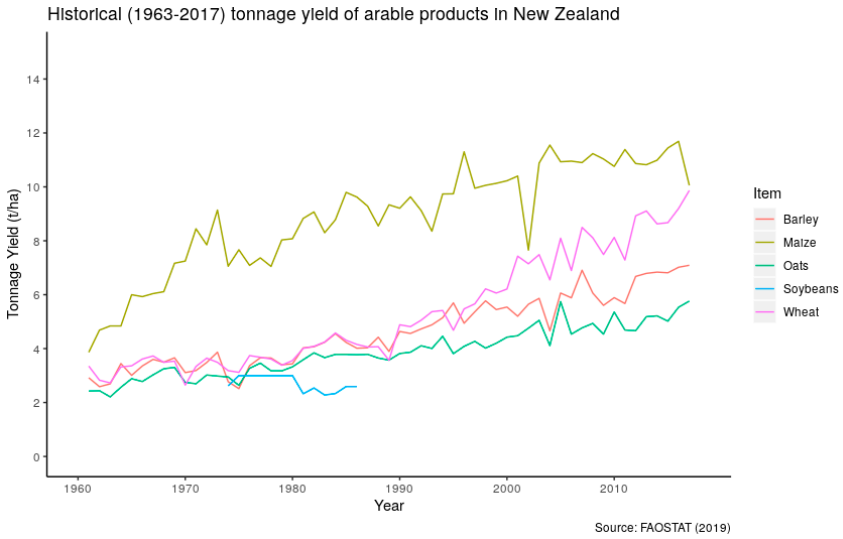


Figure 1. Historical (1963–2017) tonnage yield of arable products in New Zealand.

Other studies have attempted to use statistical modelling techniques to understand the relationship between crop yield potential and measured soil and site parameters using large, spatial, multivariate datasets. Correlation and other linear techniques (e.g., multiple linear regression [MLR] models) have been used in many previous studies to explore the relationship between crop or soil properties and the data derived from precision farming tools [7–9]. It can also provide insight into the linkages between precision farming data and crop yield spatial variability [5,10,11]. However, linear regression models assume that the relationships between the dependent and independent variables are linear. More accurate results have been reported with more complex machine learning techniques such as artificial neural networks on predicting crop yields [5,10,12].

Liu, et al. [13] used a feed-forward, back-propagation neural network (BPNN) model for predicting maize yield based on soil factors (e.g., soil pH, phosphorus, potassium, organic matter), management factors (nitrogen fertilizer), and monthly rainfall as inputs in their BPNN. The BPNN was able to model the interaction between rainfall and the rate of applied nitrogen fertilizer. It also predicted maize yields with 80% accuracy. Drummond, et al. [10] used three different algorithms (stepwise MLR, projection pursuit regression, BPNN) for predicting maize and soybean yield, with a number of soil fertility/topography variables (e.g., soil phosphorus, magnesium, potassium, pH, organic matter, topsoil depth). However, the back-propagation neural networks are computationally expensive and time consuming for handling large datasets. However, intensive soil testing is expensive. In cropping situations in New Zealand grid sampling is rarely undertaken at points less than one value per hectare.

The Cubist regression model is primarily used in remote sensing studies for handling large datasets [14,15]. Cubist is a rule-based decision tree algorithm, modified from Ross Quinlan's M5 model tree [16,17] and introduced to the R statistical software by Max Kuhn in the R development group [18]. As distinct from other types of regression trees such as the Classification and Regression Trees (CART) program (where values are predicted at their leaves), the Cubist model produces a set of rules ("if-then" statements) and each rule contains multivariate linear regression models at the terminal leaves. A specific set of predictor variables will then choose an actual prediction model

based on the rule that best fits the predictors [18,19]. Promising results have been reported when predicting continuous variables [14,15].

This paper aims to examine the use of statistical modelling techniques on precision farming data for estimating within-paddock maize-grain yield potential. The aspects looked into are:

- The uses of precision farming data including yield monitor data, spatial yield data, seeding density, high-precision elevation and soil EC;
- The uses of multispectral satellite imagery (NASA's Landsat-8 and Sentinel-2 ESA's missions) in crop management at the sub-paddock scale;
- The uses of statistical modelling techniques (stepwise MLR, Cubist regression and a feed-forward neural network) to predict within-paddock maize-grain yield potential;
- Agronomic interpretation and management implications.

The modelling output could be embedded into a GIS software program as a decision-making tool. It is hypothesized that such a modelling approach can help farmers modify crop management practices to maximize yield and minimize costs. Once the functional relationship between within-paddock yield potential and complementary variables is established, it should be possible to provide a more accurate management prescription to be uploaded to a variable-rate implement (e.g., liquid fertilizer applicator), enabling variable rates of an input (e.g., nitrogen fertilizer) to be applied automatically across the paddock based on the "input-response" function. This work suggests an approach which may be taken to develop within field crop management statistical algorithms.

2. Materials and Methods

2.1. Study Site

The Waikato region is the biggest producer of maize in New Zealand. Two rain-fed, non-irrigated sites in Waikato were chosen for this study, see Figure 2 because of their consistent within-site management histories, where the grain harvester was equipped with a yield monitor. Site 1 (FAR Northern Crop Research Site) (175.372 E, −37.835 S) is located at Tamahere, 10 km south of Hamilton. It is a 10-ha paddock, which has been dedicated to growing maize. Strip-tillage has been practiced for planting maize at Site 1 since 2013. Site 2 (175.487 E, −37.676 S) is a 23-ha paddock located 5 km southwest of Morrinsville.

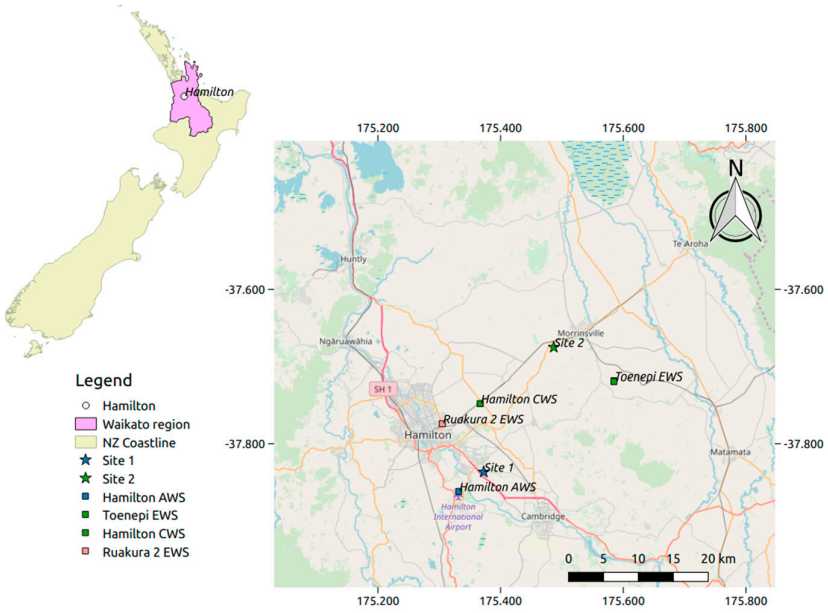


Figure 2. Study site location and the nearby weather stations (AWS. CWS, EWS).

The climate of the Waikato region tends to be warm, humid in summer and mild in winter. The average annual rainfall is 1250 mm, generally enough for crop production. However, 'dry spells' (Periods of fifteen days or longer with less than 1 mm of rain on any day from December to March) or drought events are common in the region [20].

Soils in the area between Hamilton and Cambridge have a high horticulture and cropping potential. Of note is the Horotiu soil, an Allophanic soil developing from a series of thin tephra (volcanic ash) layers overlying alluvium laid down about 18,000 years ago by the Waikato River. A key component of this soil is the amorphous mineral "allophane", which binds soil particles into stable and fine aggregates that allows free drainage but also retains moisture and is easy for roots to exploit [21]. The variability and banding in site 1 may be associated with texture variability related to old channels and levees of the adjacent stream (Figure 3).

The land-use in the region is primarily pastoral farming (58%), with less than 1% being used for crop farming. Around 5000 ha land in the region is used for growing maize, making up approximately 25% of the maize production in the North Island [1].

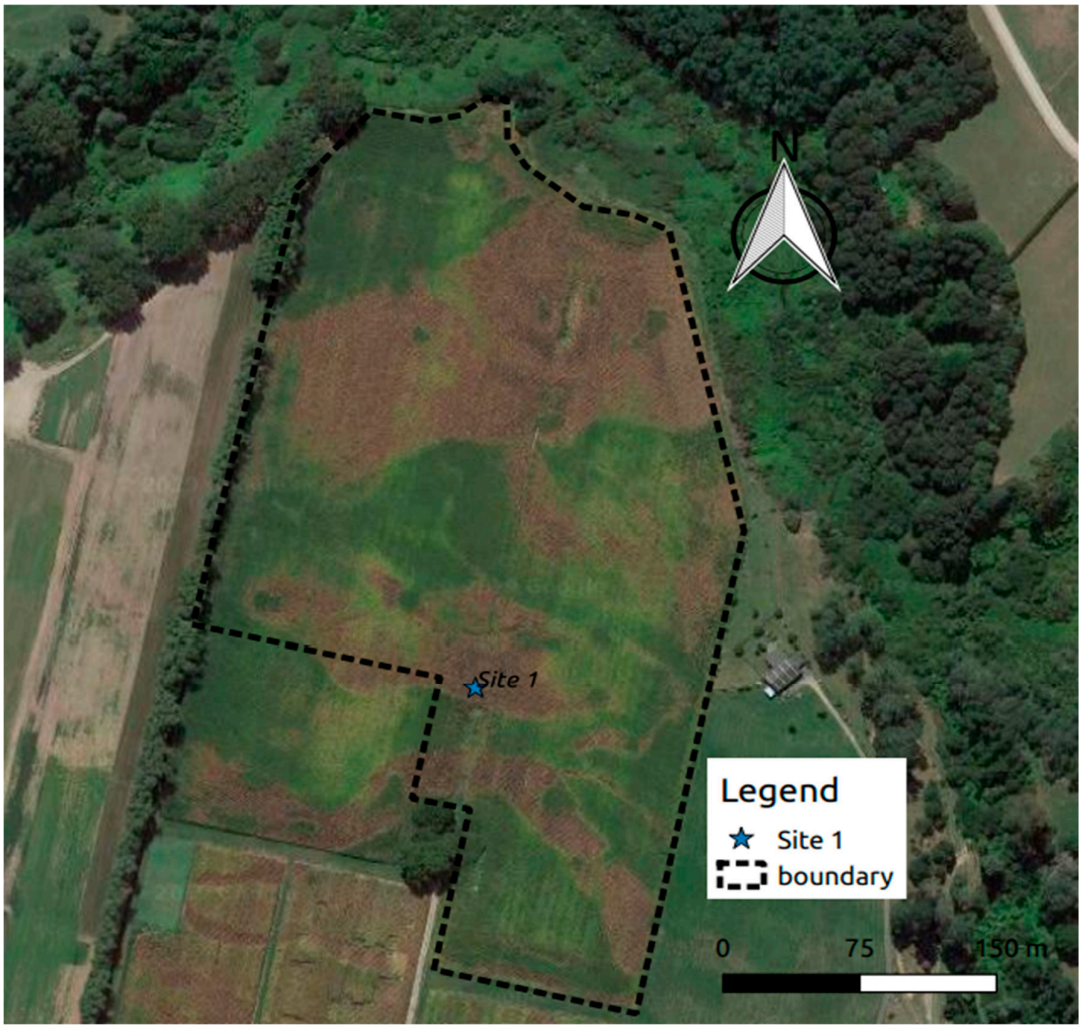


Figure 3. Google Earth image–true colour composite (site 1).

2.2. Data Acquisition

2.2.1. Geospatial Yield Monitor Data

For site 1, spatial yield data of maize-grain was collected over four years (2014, 2015, 2017, and 2018) by a yield monitor fitted on an 8-row (6.096 m swath) combine harvester with a GPS receiver. Three years of maize-grain yield data (2014, 2017 and 2018) were collected from site 2. Spatial data points were recorded at 1-s intervals during harvest.

2.2.2. Remote Sensing Imagery

The principle of crop reflectance sensing is based on the fact that healthy, actively growing plants can strongly absorb blue light (in the 0.4–0.5 μm wavelength) and red light (0.6–0.7 μm) for photosynthesis, while strongly reflecting NIR light (0.7–1.3 μm). Healthy, actively growing plants therefore display a large disparity between low red light and a high NIR light reflectance [22], which can be measured by a multispectral sensor on a satellite.

Images captured from a satellite often have a spatial resolution of 1–100 m. The spatial resolution currently used for agriculture has been in the range of 10–30 m. Sentinel-2 (available since June, 2015) was selected primarily because it is free for download and it provides reflectance data at the light spectrums (R, G, B, NIR) with sufficient spatial resolution (10 m) required to investigate crop performance at the sub-paddock scale. Sentinel-2 satellites have a 5-day revisit cycle, increasing the chance of obtaining cloud-free images at the necessary crop growth stage for the study sites. There has been much research interest in the use of Sentinel-2 data in agriculture to study crop nutrient status, biophysical variables and soil mapping in the local, regional and global scale [23–26]. There is also a potential to map sub-paddock variability and provide an indication of crop yield before harvest using Sentinel-2 data. For the acquisition of multispectral data from previous years, Landsat-8 images were downloaded. Available since April, 2013, Landsat-8 has a revisit cycle of 16 days and spatial resolution of 30 m for visible and NIR band. This resolution, however, may be too coarse for small paddocks as some pixels may contain "noise" such as bare soil and trees.

Remote sensing data (Sentinel-2 Level-1C and Landsat-8 Surface Reflectance Tier 1 collection) were analyzed using Google Earth Engine, which is an online platform that allows records of data from many satellite systems to be accessed and processed. The data were exported as a series of raster stacks, each containing reflectance data at four wavebands (R, G, B and NIR). Pixel values of each raster image were extracted at the coordinates of each yield data point.

For relatively coarse spatial-resolution satellite images such as Landsat-8 (30 m), the values of vegetation indices such as the ratio vegetation index and the normalized difference vegetation index (NDVI) can be influenced by soil background (caused by moisture differences, shadow, roughness, or organic matter) [27]. The band reflectance data were therefore converted into a soil adjusted vegetation index (SAVI) for removing soil-induced variations in soil brightness [28] using the equation:

$$\mathrm{SAVI} = \frac{(1+\mathrm{L})(\mathrm{NIR}-\mathrm{Red})}{(\mathrm{NIR} + \mathrm{Red} + \mathrm{L})} \quad (1)$$

where:

L is a canopy background adjustment factor (typically L = 0.5).

Images used for the analysis were presented in Figure 4.

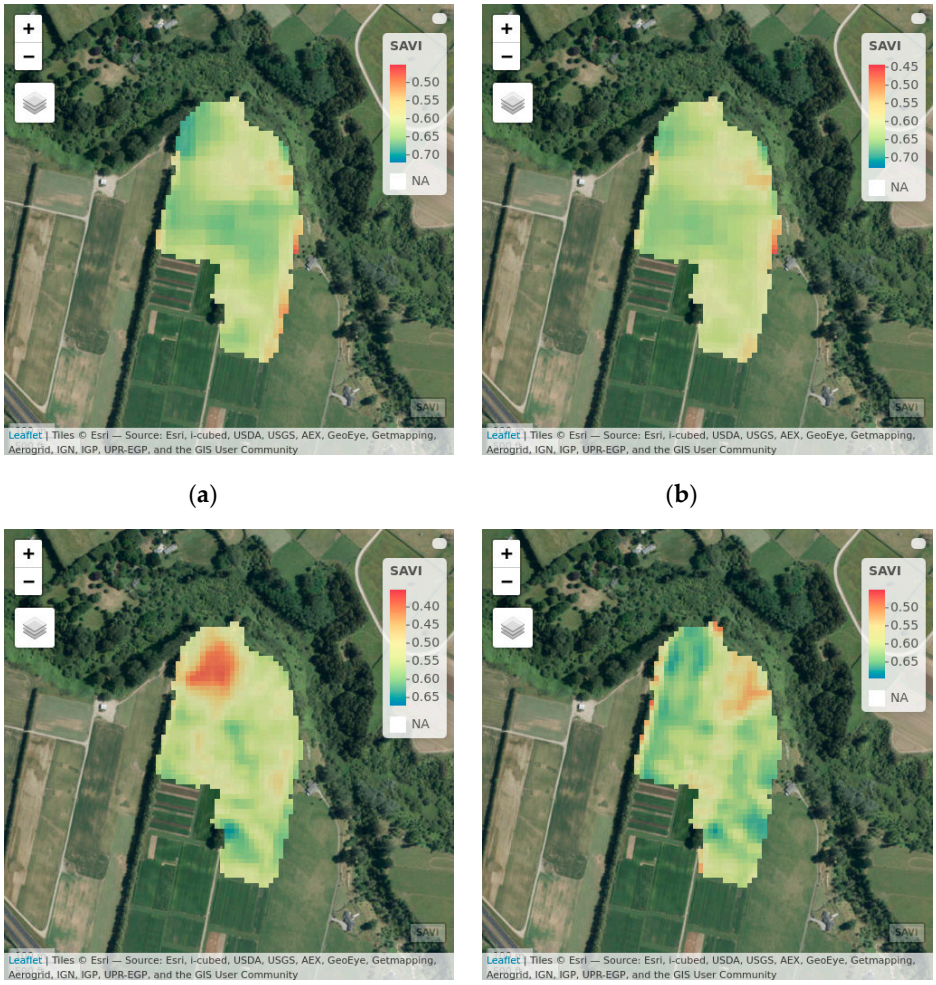


Figure 4. SAVI images captured in February (site 1): (**a**) SAVI image (9 February 2014); (**b**) SAVI image (12 February 2015); (**c**) SAVI image (23 February 2017); (**d**) SAVI image (23 February 2018). SAVI values typically range from 0 to 1. A higher SAVI indicates better plant health and larger biomass.

2.2.3. Soil Electrical Conductivity, Soil Organic Matter and Elevation

Apparent soil electrical conductivity (EC), soil organic matter (OM), and elevation (with RTK GPS) point data were collected using a Veris Mobile Sensor Platform. This system has five, disc coulter plates that measure electrical resistivity, which indicates how well the electrical current flows in soil. It also has a near-infrared LED that measures the light reflectance from the surface soil, which can be calibrated to indicate soil organic matter content [29,30]. The data were collected at 1-s intervals and 10 to 15 m apart between each transect. The spatial data point located within the closest proximity of a particular yield data point was selected. The values of soil EC, soil OM and elevation were aggregated to the location of yield as independent variables.

2.2.4. Meteorological Data

Meteorological data was downloaded from New Zealand's National Climate Database (Cliflo). For Site 1, the historical weather data were recorded by the Hamilton Aws weather station (175.332 E, −37.861 S), located within 5 km of the paddock. For Site 2, the weather data were recorded by three

different weather stations located within 15 km of the paddock: Toenepi Ews weather station (175.585 E, −37.720 S), Hamilton Cws (175.367 E, −37.748 S) and Ruakura 2 Ews (175.305 E, −37.774 S).

The data were recorded on a daily basis and can be retrieved as csv files from Cliflo. The daily rainfall, temperature, and solar radiation, critical for maize growth were considered for analysis [31]. Solar radiation data was not available for Site 1. The daily data were aggregated into 15-day intervals with the sum of rainfall, the average daily solar radiation and the average daily temperature calculated to represent crop growth stages [10]. The weather data is summarized in Table 1.

Meteorological data is important in establishing management zones as moisture stress inhibits plant growth and can lead to variations in yield based on the moisture holding capacity of the soil at various stages of crop development.

Table 1. Historical meteorological data for the crop growing season in 2014, 2015, 2017 and 2018 for site 1 (the sum of rainfall and the average temperature over 15-day intervals).

The First Day of a 15-Day Interval	Rain (mm)	T (°C)	Rain (mm)	T (°C)	Rain (mm)	T (°C)	Rain (mm)	T (°C)
	2014	2014	2015	2015	2017	2017	2018	2018
1 September	76.6	10.59	80.2	11.99	62.8	10.98	129.2	11.72
16 September	83.4	12.59	70.4	11.13	81.2	14.11	67.8	12.10
1 October	44.8	12.96	19	11.87	84.4	12.98	51.2	13.18
16 October	12.6	12.55	57.2	13.38	26.6	13.11	54.6	13.81
31 October	39.8	14.61	45	12.41	50	14.31	37.8	14.44
15 November	33.8	17.28	39.8	14.13	65	14.82	63.2	17.23
3 November	70.4	16.83	45.8	15.53	24	16.76	8.6	19.43
15 December	28.8	16.89	27.4	17.98	23.8	15.33	10	18.64
3 December	14.8	16.47	25.4	18.32	17	16.79	68.2	20.00
14 January	23.8	16.48	5.4	19.98	43	16.29	43.6	21.89
29 January	10.4	19.05	24	18.77	13.8	18.08	129.8	19.96
13 February	2	18.64	22.6	18.73	122	19.85	73.6	19.95
28 February	0.8	16.18	26	18.56	151.6	18.41	20	19.06
15 March	5.2	17.49	43.6	17.33	68	17.75	34.6	17.68
3 March	63.4	17.67	43.6	16.31	152.8	16.97	26.4	15.17
14 April	82.6	15.28	135.2	13.20	18	14.07	73.6	13.95
29 April	53	12.67	64.4	12.78	98.2	12.05	54.2	14.04
14 May	32.6	10.21	94.2	10.17	59	10.01	98.8	12.33
29 May	10.6	9.53	0.4	8.43	1.2	12.37	0.2	5.40

2.2.5. Maize Seeding Density

For the 2015/2016 and 2016/2017 seasons, the paddock was planted on 18 and 19 October using an 8-row (6.1 m swath) John Deere precision maize planter with seed meter and DGPS. Data points were recorded at 1-s intervals and about 2 m apart from each other. Again, the spatial data point located within the closest proximity to a particular yield data point was selected to aligning the value of seeding rate to the value of yield at the closest location. The values of seed rates applied were aggregated to the location of yield. The planting data for the 2013/2014 and 2014/2015 seasons were not available for Site 1. After consulting the grower, we assumed that a density of 100 thousand seeds/ha was uniformly applied across the paddock for those two seasons.

2.3. Spatial Data Filtering

Spatial yield monitor data were transferred from a USB stick from the yield monitor on the combine harvester to an office computer. The data was loaded into the Ag Leader SMS Advanced software (version: 19.2) (a GIS software program for precision farming data management) and organized under the farm property. A freeware Yield Editor 2 [32] was initially used to remediate data errors, which may have been caused by systemic errors (e.g., incorrect combine delay time,

inaccurate swath width, GPS positional errors, abrupt ground velocity changes and ramping at the crop edge) [33]. However, only Ag Leader and GreenStar Text files are accepted by the program. It is not straightforward to process the data collected from other yield monitors such as GUANTIMETER and RDS (which are volumetric sensors). Therefore, in this study, the spatial yield monitor data of each year was filtered at two stages, using generic methods and exported as a shapefile format.

The two steps include:

1. Outliers were identified in a boxplot as the points that were located outside the fences (whiskers) of the boxplot (outside 1.5 times the interquartile range above the upper quartile and below the lower quartile) and were removed from the dataset. Figure 5 demonstrated the automated procedure that was used for determining how to set the limits for the MIN-Y and MAX-Y parameters: Step 1) upper [Q2] and lower [Q1] quantiles of the yield distribution were computed at the yield maximum and minimum values. Step 2) the interquantile range was computed. Step 3) to compute MIN-Y and MAX-Y, the upper and lower quantile values were expanded outward by a percentage of the inter-quantile range indicated by the Y-scale.

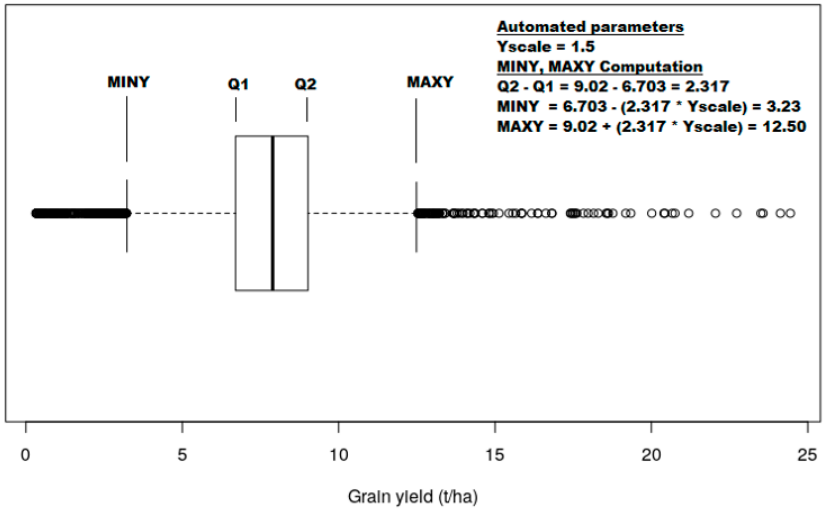


Figure 5. 2017 harvest yield boxplot and method for computing automated MIN-Y and MAX-Y filter parameter values.

2. Inliers were identified using a spatial filtering technique developed by Spekken, et al. 34. The coefficient of variation (CV) of a specific point was first calculated in relation to other points within a defined search radius (Figure 6). The points with CVs above 20% within a 5 m radius were identified. Then if the number of points inside each radius equals the number of points with high CV, that specific point was labelled as inliers and removed from the dataset. This technique requires minimal information and is effective in filtering a number of specific errors in yield monitor data (e.g., fill mode, incorrect set width and lag time) [34].

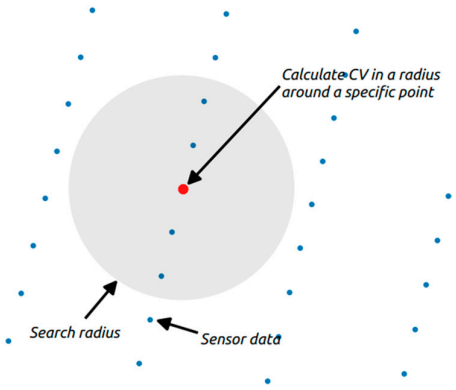


Figure 6. Inlier filtering process.

Geostatistical analysis was conducted to evaluate the performance of the spatial filtering. Geostatistics, based on the Matheron [35] regionalized variable theory, describes that the spatial dependency of near locations is greater than that of far locations [36]. The application of geostatistics to precision agriculture have been described in many studies [37–39].

A variogram (or semivariogram) quantifies spatial dependency using mathematical terms. This is expressed in Equation (2):

$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{j=1}^{m(h)} \left\{ Z_{x_j} + Z_{x_j+h} \right\}^2 \quad (2)$$

where:

$m(h)$ = the number of paired comparisons for the lag interval h;

$\hat{\gamma}(h)$ = semivariance for interval distance class h;

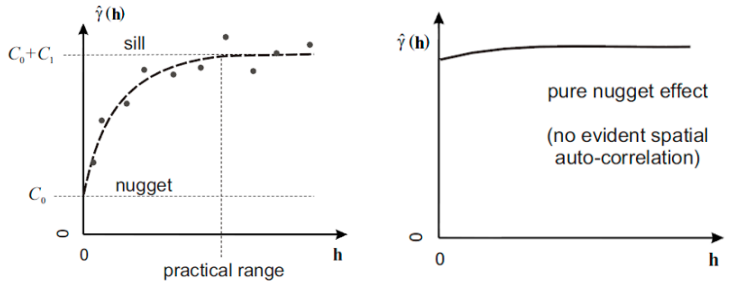
Z_{x_j} = measured sample value at point x_j;

Z_{x_j+h} = measured sample value at point $x_j + h$.

Figure 7a shows a typical variogram and its parameters. As the separation, or lag, distance h increases, semivariance $\hat{\gamma}(h)$ increases and then reaches a maximum at the level known as the sill $C_0 + C_1$. The practical range is defined as the separation distance beyond which two observations are spatially independent of each other.

The discontinuity at the origin is called the nugget C_0 effect, which often indicates measurement errors and sources of variation over distances less than the shortest sampling interval. Partial sill is the sill variance subtracted by the nugget [40]. The pure nugget effect often indicates that the sampling intervals are too large to adequately capture the spatial variation of the measured property or the number of samples is insufficient to compute variogram reliably (Figure 7b).

Therefore, if the nugget variance was reduced in relation to the sill variance, it was considered that the measurement errors were remediated [36,41].



(**a**) (**b**)

Figure 7. Variogram forms: (**a**) bounded variogram; (**b**) pure nugget (modified from [40]).

2.4. Spatial Data Modelling

The spatial data modelling analysis was conducted using R statistical software (R Core Team, version 3.5.3) with required packages: sp (version 1.2–5), rgdal (version 1.2–13), raster (version 2.5–8), caret (version 6.0-84), Cubist (version 0.2.2), nnet (version 7.3–12).

2.4.1. Eliminate Multicollinearity

Multicollinearity refers to the issue that an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation. Multicollinearity would increase the standard errors of some independent variables and make them statistically insignificant (while they should be statistically significant). A number of methods have been applied to handle multicollinearity such as applying orthogonal transformation to the data using Principal Component Analysis (PCA) or removing highly correlated variables indicated by the Variance Inflation Factor (VIF) [42]. However, PCA does not remove any variable but compresses the information into several independent principal components (PCs). VIF is generally acceptable for linear models, and when there are more samples than the predictors. It may not be appropriate otherwise [43].

In this study, an alternative method [43] was used for handling multicollinearity. First, the pair of predictors with the largest absolute correlation was determined. The linear correlation between two continuous variables was quantified using the Pearson correlation test. The coefficient of correlation (r) indicates how well one variable X is related to the variation in another variable Y. The correlation coefficient is expressed mathematically in Equation (3):

$$\mathrm{r} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{[\sum_{i=1}^{N}(x_i-\bar{x})^2 \sum_{i=1}^{N}(y_i-\bar{y})^2]^{1/2}}, \quad (3)$$

where:

N = Total number of observations;

x =Mean of the variable 'x';

y = Mean of the variable 'y'

As a rule of thumb, a correlation coefficient up to 0.20 is considered negligible, from 0.20 to 0.40 is low, 0.40 to 0.60 means moderate, 0.6 to 0.80 is substantial and from 0.80 to 1.0 is considered high to very high.

The average correlation between each predictor and all of the other variables was computed for both predictors. The variable with the largest mean absolute correlation was labelled for removal. This procedure was repeated until no correlations were above the defined threshold ($r > 0.8$).

2.4.2. Stepwise Multiple Linear Regression

Multiple Linear Regression (MLR) has been used to model the relationship between one response variable and many predictor variables by using parameters entered linearly and estimated by the least squares method. The linear regression is expressed mathematically in Equation (4):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon, \quad (4)$$

where:

β_i = coefficients;

X_i = the predictors;

Y = the response variable (maize-grain yield);

β_0 = a constant ($\beta_0 = 0$ as grain yield cannot be negative).

For MLR, the dependent variable (yield) needs to be normally distributed. To achieve a normal distribution, the historic yield was transformed using a natural logarithm. Then, an MLR was undertaken on the independent variables (e.g., soil EC, elevation, OM) with maize-grain yield as the dependent variable while simultaneously removing the variables that are not statistically important

in each step (stepwise MLR). The modelling results can help to identify the cause-and-effect relationships between maize-grain yield and predictor variables.

2.4.3. Cubist Tree Regression

Cubist regression was undertaken with ten-fold cross-validation performed to select the best model parameters (committees and neighbours) in order to optimise model performance. In the 10-fold cross validation, the data were randomly divided into 10 subsets of equal size. The regression technique was then repeated 15 times, with each repetition leaving out one of the validation subsets, and using only that subset to compute the root mean square error (RMSE). The parameters resulted in the smallest RMSE were selected for the model.

For example:

Model1: Trained on Fold1 + Fold2 + … Fold9 (90%), Tested on Fold10 (10%)
Model2: Trained on Fold2 + Fold3 + … Fold10 (90%), Tested on Fold1 (10%)
Model3: Trained on Fold1 + Fold3 + … Fold10 (90%), Tested on Fold2 (10%)

Figure 8 demonstrated this procedure—the reduction of RMSEs by tuning a set of parameters. Little gain in accuracy was found by increasing the number of committees.

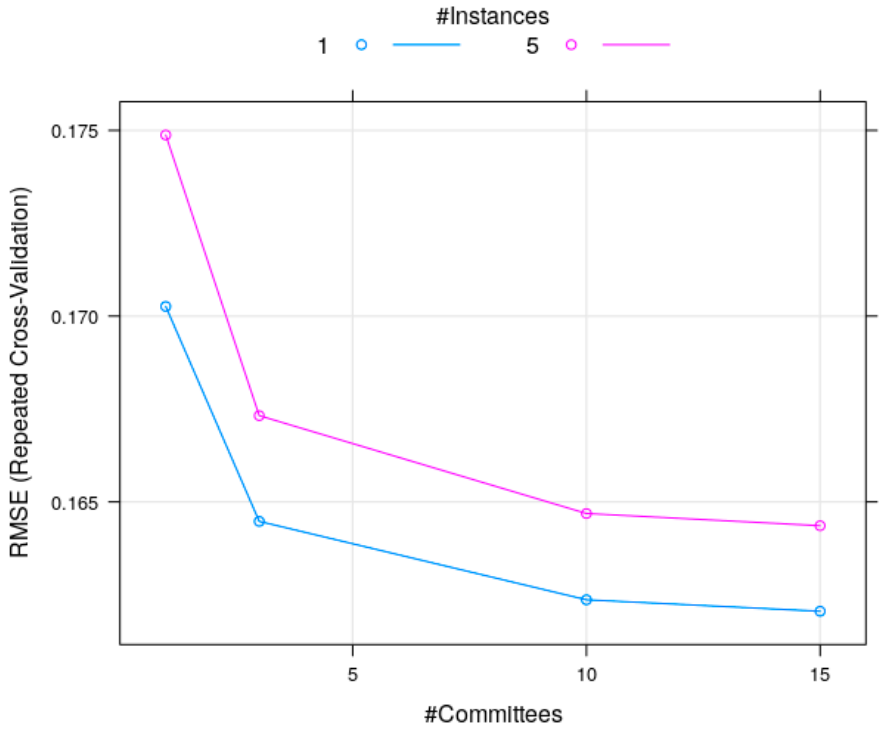


Figure 8. Computation of the prediction errors (RMSEs) for each validation set withheld in repeated cross-validation under various model parameters.

2.4.4. Feed-Forward Neural Network

Feed-forward, back-propagation neural network model has been used for predicting maize yield with soil factors, topography and meteorological data as inputs in a number of studies [10,13]. To benchmark their findings, a single hidden layer feed-forward neural network structure (Figure 9) was used to develop maize-grain spatial yield prediction models. Network topology was limited to a single hidden layer because of reportedly a faster execution and comparable prediction performance [44–46].

The architecture of the neural network was as follows:

- Number of layers = 3 (input, hidden, and output).
- Number of neurons in the hidden layer = 5 to 20.
- Type of activation functions = sigmoid for hidden layer, linear for output layer.
- Number of nodes in input layer = 7.
- Number of nodes in output layer = 1.
- Network error type = root mean square error (RMSE).

The mathematical form of neural network for regression can be expressed as:
From the input layer to the hidden layer:

$$H = h(\alpha_{0\mathrm{m}} + \alpha_m^T I), m = 1, \ldots, M, \quad (5)$$

where:

I = an input variable and $I^T = (I_1, I_2, \ldots, I_7)$

$\{\alpha_{0\mathrm{m}}, \alpha_{\mathrm{m}}; m = 1, 2, \ldots, M\}$ = the weights assigned to the network

$h(.)$ = the activation function, usually an S-shape (sigmoid) function $h(\mathrm{v}) = \frac{1}{1+e^{-sv}}$

From the hidden layer to the outputs (for regression):

$$O = \beta_0 + \beta^T H, \quad (6)$$

where:

$\{\beta_0, \beta\}$ = the weights assigned to the network.

For regression, a model was fitted to minimise the sum of squared error:

$$R(\theta) = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta^T h_i)^2 \quad (7)$$

Gradient descent back-propagation is the most common optimization algorithm to find the minimum of $R(\theta)$ by evaluating its derivatives with respect to weights in a network. However, it tends to run slower than other optimisation algorithms such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm and conjugate gradient [47]. In our study, the neural network analysis was conducted using the "nnet" package, which uses a general quasi-Newton optimization procedure–BFGS algorithm [48]. BFGS is a common optimisation technique for neural networks, which runs faster and is more reliable than the gradient descent [47,49].

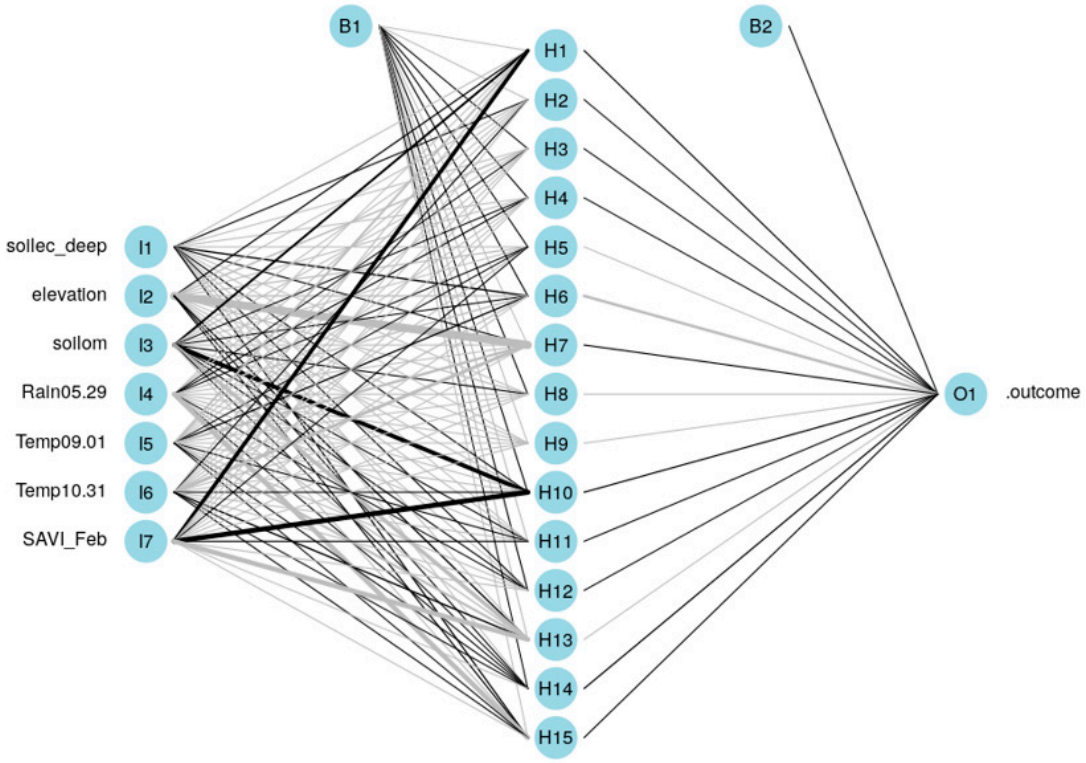


Figure 9. Illustration of a single hidden layer feed-forward neural network (the black lines are positive weights and the grey lines are negative weights. Line thickness is in proportion to the magnitude of the weight relative to all others. The input variables are shown in the first layer (labelled as I1-I8) and the response variable (yield) is shown in the far-right layer (labelled as O1). The hidden layer is labelled as H1–H15. B1 and B2 are bias layers that apply constant values to the nodes).

As with the procedure for the Cubist model, repeated ten-fold cross validation was conducted for the optimization of model parameters (size and weight decay) with the hidden layer (Figure 10). In this procedure, the training data was randomly divided into 10 folds. Each iteration left one subset (fold) as the validation set and the remaining subsets (folds) as the training set. Based on previous studies [10,13], four hidden layer sizes ranging from 5 to 20 hidden units were selected for optimization. Weight decay is a parameter to ensure the weight to decay in proportion to its size and. Four weight decays (0.001, 0.01, 0.1 and 1) were used for tuning the neural network model. Training on each data set continued until the maximum number of iterations (500) was reached; with both training and validation RMSE calculated. Within each site, the model that provided the minimal RMSE was determined, and the associated RMSE of validation was calculated with the optimised model.

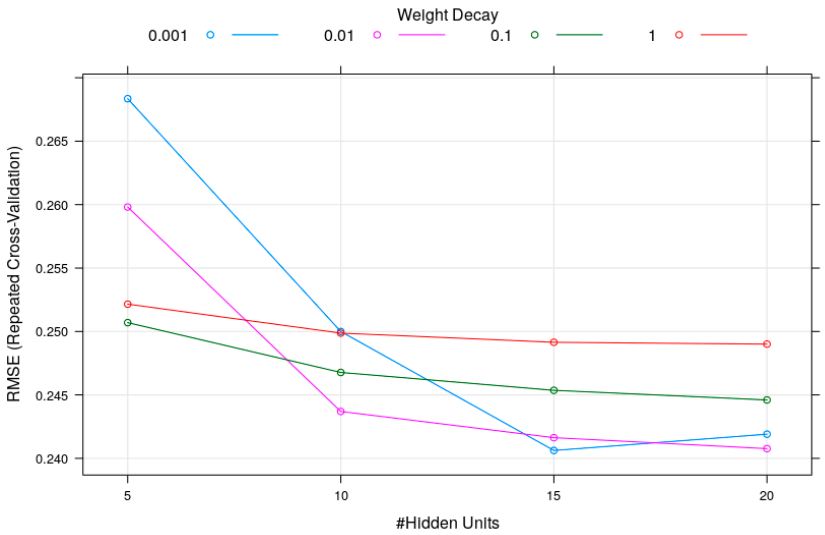


Figure 10. Hyper-parameter optimization for neural networks in repeated 10-fold cross validation.

2.4.5. Multiple Year Analysis

The "split-sample" approach was used to measure prediction accuracy, in which a subset (validation set) of the data is withheld from training. A measure of the accuracy of prediction on this validation set is then reported. In the multiple year analysis, data subsets were created from the maize data for all available years (2014, 2015, 2017, and 2018). Each training set consisted of 75% of the data, randomly sampled (with no replacement), and each validation set contained the remaining 25%.

2.4.6. Leave-out-One-Year Analysis

Due to the relatively small number of years, multiple year data was cross-validated by withholding one year of data as a validation set for each iteration, with all remaining years included in the training set. The training set was used to predict yields for the year that was held out as a validation set (Table 2). This process was iterated over the data for all the years and RMSEs were computed. This will provide an indication of the ability of the trained model to handle new information (i.e., yield data collected from an additional harvest).

Table 2. Datasets used in the leave-out-one-year analysis.

Model	Training Set	Validation Set
1	2014, 2015, 2017	2018
2	2014, 2015, 2018	2017
3	2014, 2017, 2018	2015
4	2015, 2017, 2018	2014

3. Results

3.1. Data Filtering

The results of variogram analysis (Table 3) show that the spatial filtering method used in this study has reduced the nugget/sill ratio from 0.18–0.91 to 0 for site 1; and from 0.08–0.58 to 0–0.05 for site 2 (Figure 11). After spatial filtering, the largest spatial variations in yield were found in 2015 for site 1, indicated by the partial sill/range ratio (0.34). The largest spatial variations in yield were found in 2014 for site 2. This may be attributed to the highest average yield for that individual year (Table

4). Figure 5 demonstrated that the large nugget effect in the un-filtered data due to short-distance measurement error was eliminated by the stepwise spatial filtering method.

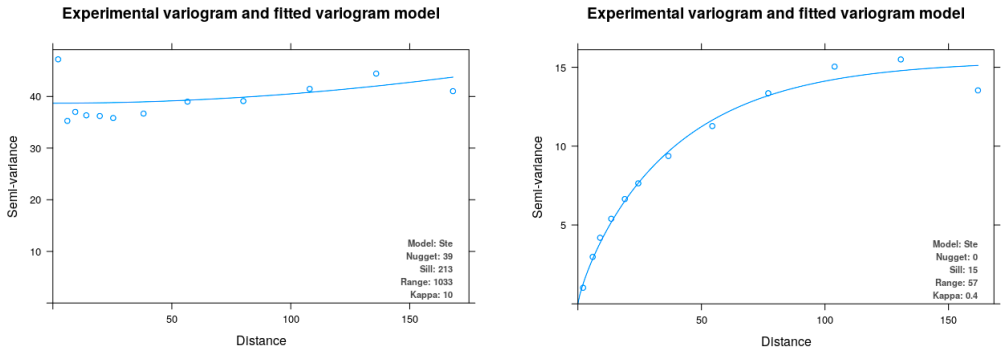
Table 3. Variogram statistics of yield monitor data values before and after-spatial filtering.

Year	Nugget/Sill		Partial Sill/Range	
Site 1				
	before	after	before	after
2014	0.18	0	0.17	0.26
2015	0.91	0	0.02	0.34
2017	0.43	0	0.02	0.13
2018	0.59	0	0.04	0.24
Site 2				
	before	after	before	after
2014	0.08	0.05	0.62	0.37
2017	0.58	0.04	0.06	0.13
2018	0.29	0	0.06	0.05

Table 4. Statistics of yield monitor data values before and after-spatial filtering.

Year	Mean		Sd		Min		Max		Obs	
Site 1										
	before	after	before	after	before	after	before	after	before	after
2014	7.13	8.54	6.3	3.6	0	1.09	97.99	19.68	11,452	6871
2015	13.39	13.5	8.3	3.18	0	3.22	101.5	22.96	6333	5555
2017	7.82	7.93	2.11	1.61	0.32	3.23	24.44	12.49	9017	8431
2018	10.96	11.23	2.68	1.84	0.38	5.83	24.99	16.38	8592	7773
All year	-	10.05	-	3.36	-	1.09	-	22.96	35,394	28,630
Site 2										
	before	after	before	after	before	after	before	after	before	after
2014	8.94	12.44	7.33	6.30	0.00	0.80	46.10	33.60	23,538	13,230
2017	9.65	10.15	3.55	2.72	0.32	0.95	25.02	18.67	18,063	14,980
2018	10.33	10.71	2.44	1.69	0.33	5.61	25.07	15.38	18,041	16,747
All year	-	11.03	-	4.01	-	0.80	-	33.60	59,642	44,957

Table 4 summarises yield monitor data values before and after-spatial filtering. Around 586 to 4581 points were removed for site 1 and 1294 to 10,308 points were removed from site 2. The extreme values of high yields such as 97.99 and 101.5 t/ha were removed using the filtering method. The values of 0 yields were also removed. The nil yields may have been caused by the status of an empty header (the combine harvester was turning around in the headland but not harvesting any crop).



(**a**) (**b**)

Figure 11. Experimental variograms of 2017 harvest yield at site 1. The best fitted variogram model was selected based on the smallest sum of square: (**a**) before spatial filtering (Model: Ste; Nugget: 39; Range: 1033); (**b**) after spatial filtering (Model: Ste; Nugget: 0; Range: 57).

3.2. Data Structure

Our constructed data for Site 1 consists of 46 continuous variables (including the year, four-year maize yield, two soil EC variables of shallow and deep, elevation, soil OM, seeding rates, four SAVI data captured in February, 19 rainfall variables and 19 temperature variables, collected at 15-day intervals) and totalling 28,630 observations. The data for Site 2 consists of 64 variables (including the year, three-year maize yield, two soil EC variables of shallow and deep, elevation, seeding rates, three SAVI data captured in February, 19 rainfall variables, 19 temperature variables, and 19 radiation variables, collected at 15-day intervals) and totalling 44,957 observations.

3.3. Model Outputs

For Site 1, after removing the highly correlated independent variables, 12 independent variables remain: soil EC deep, elevation, soil OM, and SAVI, seeding rates, the accumulated rainfall (29 May–12 June, and 15–29 March), the average temperature (1–15 September, 31 October–14 November, 16–30 November, 30 November–14 December, and 29 January–13 February). Seeding rates and the accumulated rain (March 15 to March 30) are not significant ($p > 0.05$) and are removed from the model. The average temperature (15 November to 14 December, and 29 January to 13 February) are estimated with "NA" as coefficients. This suggests those variables do not add any new information to the model and are also removed. The model is then run again with the remaining variables, including SAVI, soil EC deep, soil OM, elevation, and the accumulated rainfall (29 May–12 June), the average temperature (1 September–14 September, 31 October–14 November) (Table 5). For Site 2, seven significant variables are included in the final MLR model, including soil EC (shallow and deep), elevation, seeding rates, the accumulated rainfall (14 April to 28 April, 14 May to 28 May) and SAVI.

Table 5. Results of the multiple linear regression (MLR) models for the sites.

Model Variables	Estimated Coefficient	Standard Error	*T* Value	*p* Value
Site 1				
Soil EC deep	0.014	0.000	30.677	<0.001
Elevation	−0.035	0.002	−22.447	<0.001
Soil OM	0.546	0.033	16.452	<0.001
Rain05.29	−0.021	0.001	−25.130	<0.001
Temp09.01	0.163	0.006	25.478	<0.001
Temp10.31	−0.041	0.003	−15.184	<0.001
SAVI	2.628	0.045	58.442	<0.001
Site 2				
Soil EC shallow	0.008	0.001	9.299	<0.001
Soil EC deep	0.016	0.001	25.996	<0.001
Elevation	−0.014	0.001	−21.629	<0.001
Seeding rate	0.001	0.000	6.068	<0.001
Rain04.14	0.005	0.000	53.580	<0.001
Rain05.14	−0.003	0.000	−28.394	<0.001
SAVI	3.998	0.040	99.231	<0.001

3.3.1. Multiple Year Analysis

For both sites, both models demonstrate reasonable accuracy for predicting yield (Table 6), since the accuracy for training and validation was close. The prediction errors for the validation set (RMSEs) are smaller than the standard deviation (SD) of the multiple year predictions. For both sites, the cubist model showed that it was able to explain 70–80% of yield variation, indicated by the R^2 value. This is better than the neural network (which explained 20–60% of variation) and MLR (which

explained 30–50% of variation). The neural network produced better statistical prediction than MLR in training and validation for site 1. This, however, was not the case for validation at site 2. This suggested that the neural network model was not fully optimized for the site 2 data. The model still suffered the issue of "overfitting" (the accuracy of the training part is very high; however, the model cannot be applied successfully to the validation set). More iteration in the neural network may improve its statistical prediction accuracy in validation through this optimization process.

Table 6. Prediction errors (training and validation) provided by the MLR and Cubist model in the multiple year analysis.

	MLR Model		**Cubist Model**		**Neural Network**		**Observed Yield**	
				Site 1				
	RMSE	R^2	RMSE	R^2	RMSE	R^2	mean	SD
Training	0.27	0.47	0.16	0.81	0.24	0.58	10.06	3.36
Validation	2.41	0.51	1.47	0.82	2.12	0.61	10.05	3.36
				Site 2				
	RMSE	R^2	RMSE	R^2	RMSE	R^2	mean	SD
Training	0.34	0.29	0.22	0.69	0.31	0.41	11.03	4.02
Validation	3.37	0.31	2.13	0.72	3.57	0.22	11.03	4

3.3.2. Leave-out-One-Year Analysis

In the leave-out-one-year analysis, the results (Table 7) provided by the Cubist model are less accurate. The MLR model produced lower RMSEs (1.57–4.93) and higher R^2 (0.15–0.28) than the Cubist model (RMSE = 2.64–5.9; R^2 = 0.05–0.14) for all individual years for Site 1. The higher RMSEs in the Cubist model may be a result of skewed data distribution. More accurate results of statistical prediction were found in the multiple year analysis because the data was normally distributed. This suggests that the more complex machine learning models do not necessarily perform better at predicting within-paddock yield potential for a new harvest than a simple linear model, as the data distribution is often unknown. The poorest results, indicated by high RMSEs, were found for 2014 for both sites in the leave-out-one-year analysis. The distribution of the data appears to have two different trends (Figure 12a,c), which could indicate different soil types, management practices, or poor calibration of the yield monitor. Applying log-transformation to the dependent variable (yield) did not effectively transform the data into normal distribution (Figure 12b,d).

Table 7. Prediction results of the MLR and Cubist model in the leave-out-one-year analysis.

	MLR Model		**Cubist Model**		**Neural Network**		**Observed Yield**	
Site 1								
Withhold	RMSE	R^2	RMSE	R^2	RMSE	R^2	mean	SD
2014	4.93	0.28	5.9	0.14	6.97	0.14	8.54	3.6
2015	3	0.18	3.51	0.05	3.18	0.02	13.5	3.18
2017	1.57	0.31	2.92	0.08	3.03	0.22	7.93	1.61
2018	1.85	0.15	2.64	0.14	3.39	0.14	11.23	1.84
Site 2								
Withhold	RMSE	R^2	RMSE	R^2	RMSE	R^2	mean	SD
2014	9.09	0.3	7.33	0.09	6.52	0.25	12.44	6.3
2017	3.91	0.15	3.04	0.03	2.82	0.01	10.15	2.72
2018	3.15	0	2.22	0.05	3.83	0.00	10.71	1.69

For site 2, the best results were produced by the neural network (RMSE = 2.82–6.52) and an improvement for the year of 2014 and 2017. The Cubist model produced better results (RMSE = 2.22–7.33) than the MLR model (RMSE = 3.15–9.09). However, the R^2 provided by the Cubist and neural network model were lower, indicating substantially over- or under-estimated yield.

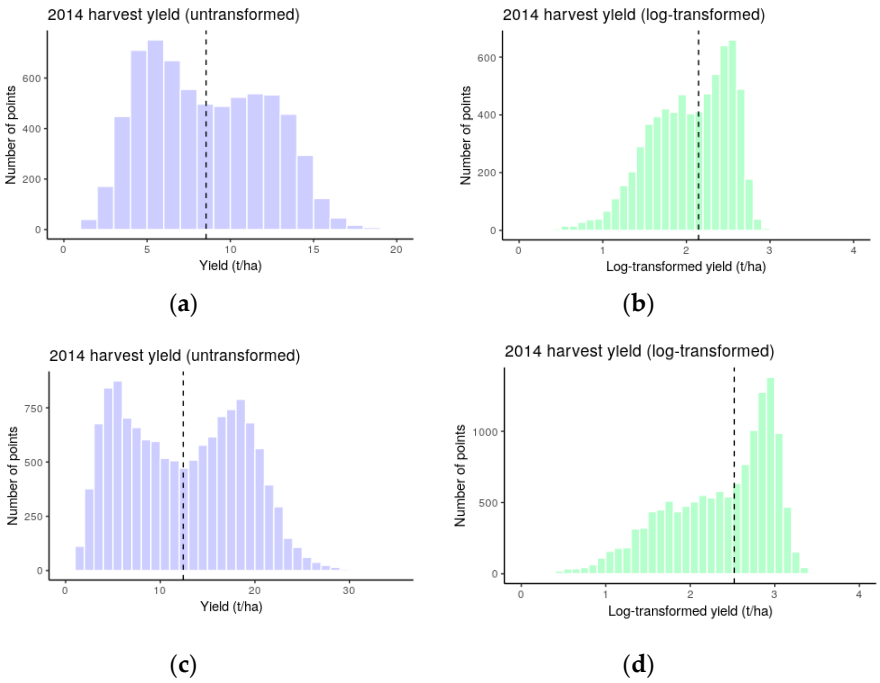


Figure 12. 2014 harvest yield data distribution (the vertical dash line represents the average value of the data distribution): (**a**) untransformed (site 1); (**b**) log-transformed (site 1); (**c**) untransformed (site 2); (**d**) log-transformed (site 2).

4. Discussion

The SAVI images captured in February were the most influential predictor in the MLR models for both sites in this limited data capture, producing the highest estimated coefficient (2.628 for site 1 and 3.998 for site 2, respectively). This supports Chang, et al. [50]. Their study estimated the maize yield variability for two 65-ha fields using multispectral and multi-date reflectance data measured by a digital camera mounted on an airplane and the IKONOS satellite. MLR and principal component regression were used. The models were able to explain 40%–60% of the yield variations. The methods used would also rank other factors such as meteorological data when moisture deficit is evident with some soil textures. The results from this study suggest the crop reflectance data can be used for mapping spatial maize yield variability in paddock before harvest. In our study, however, we were unable to acquire consistent multi-date data from the Landsat-8 and Sentinel-2 satellites at each growth stage for crop management in New Zealand due to cloud coverage at the point of capturing the images. The crop growth can be influenced by environmental conditions (e.g., the weather pattern of years, planting time, and time of the first onset), which makes it difficult to model yield if satellite images are not available for different growing stages. However, if a particular date in a particular year when the onset of grain filling is known (often shortly after max leaf area and beginning of senescence), an image can be obtained using a UAV (unmanned aerial vehicle) around this time frame to provide a good indication of final grain yield. This is also the point when decisions can be made to scout pests and diseases from the UAV image, and then apply patch spraying to improve yield or quality.

Soil organic matter (OM) was the second most influential predictor in the MLR models for site 1. Soil OM provides nutrients and habitat to micro-organisms living in the soil. It also binds soil particles into aggregates and improves the water holding capacity of the soil [51]. Our model suggested that a 1% increase in soil OM may result in a 1.72 t/ha increase in maize-grain yield for the site.

The average temperature (1–14 September) had a significant positive influence on maize-grain yield for site 1. Temperature and solar radiation are key factors that determine the potential

production of maize. Maize has a base temperature (the minimum temperature a plant requires to assimilate carbon dioxide through photosynthesis) range from 8–10 °C. Maize is ideally sown in early to mid-October in the northern North Island region of New Zealand once the soil temperature has increased to above 10°C [31]. A warmer soil temperature in spring is favorable for seed germination and seedling emergence. Our model suggested that an increase of 1 degree Celsius in the average temperature from September 1 may increase yield by 1.18 t/ha. With information on the soil temperature in September, the model may inform the optimal seeding rate to apply.

Soil EC deep has a positive influence on yield. For site 1, Soil EC shallow was highly correlated with soil EC deep (r = 0.9) and was removed due to multicollinearity. A higher soil EC value often indicates finer textured soil and higher soil water holding capacity [52]. Kitchen, et al [5] studied the relationship of soil EC and topographic features (elevation, slope and curvature) to grain yield for three contrasting soil-crop systems in the US. Their MLR models explained 1%-30% of the yield variation. The prediction was improved by adding the topographic features to the model. In our study, the correlations between soil EC/elevation and yield in the multiple years were very weak (r = 0.1). This may be related to the spatial resolution (10 m) of the data. In the study by Kitchen, et al. [5], kriging was applied to interpolate soil EC/elevation into 10-m grid, which reduced the variations and the size of the dataset to be modelled. Our study did not interpolate the spatial data because there were missing data points in different parts of the paddock from year to year. In the study by Blackmore, et al. [53], they also noted that the results of spatial yield prediction may be related to the spatial resolution of the predictors, as coarse spatial points tended to lose details from the resulting maps.

For Site 1, elevation, accumulated rainfall (29 May–12 June), and the average temperature (31 October–14 November) have negative impacts on the yield. For Site 2, elevation and rainfall (14 May–29 May) have negative influences on yield (Table 5). It was noted that the period of meteorological data such as rainfall and temperature may be too short (15 days) for studying their impacts on yield. However, there was no guideline on which interval to use, as different intervals were used previously such as a 15-day interval during the plant reproductive phase [10], or a monthly interval [13]. Elevation influences water movement and soil wetness potential in the paddock. Higher elevated soil may be more susceptible to nutrient runoff and erosion during rainfall events [54,55] and is likely to dry out earlier. The MLR model suggests that an increase of 1 m in elevation can cause a decrease of 1.04 t/ha in yield. There may also be other stronger influences (e.g., soil type) on the productivity. Based on our observation during a field visit in November, 2018 (as shown in Figure 13), it was hypothesized that the soil texture variations may have influenced the drainage and the performance of the seedling crop, differed within the paddock. Target sampling was undertaken with core samples taken at 0–90 cm depth and we are currently undertaking soil particle size analysis in the laboratory to examine this hypothesis.

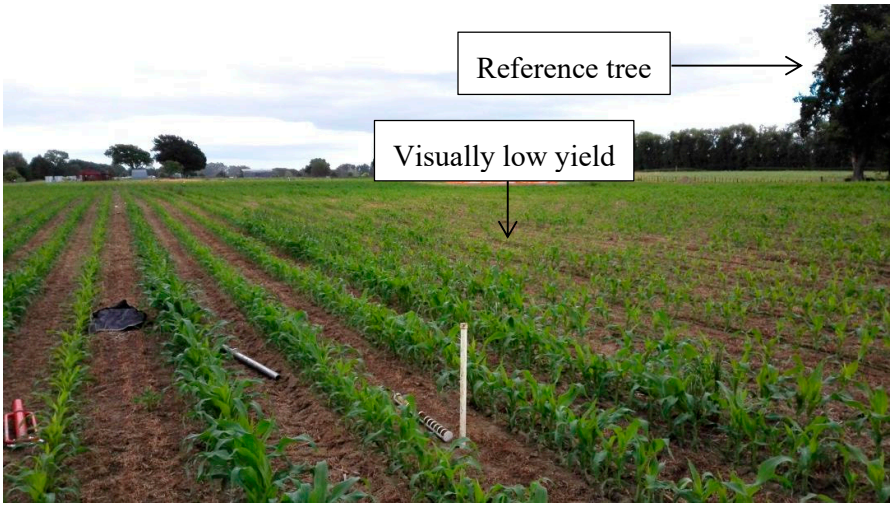


Figure 13. Site 1 (the photographer was standing facing South, with the tree on the right-hand side as an approximate reference of the location; visually low yield in the middle centre and middle right of the photo matched with the small rise in elevation).

Our results are consistent with Drummond, et al. [10]. The authors predicted maize and soybean yield on three sites in the US state of Missouri (ranging from 13 to 36 ha in size) using a feed-forward, back-propagation neural network (BPNN) model with a number of soil fertility parameters (e.g., soil pH, OM, phosphorus, calcium, magnesium, potassium) and topographic inputs (e.g., elevation, slope). The BPNN model generally provided better statistical predictions in the multiple year analysis than the other two models (MLR and projection pursuit regression). However, in the leave-one-site-year analysis, high prediction errors were produced for the individual site-year due to severe overfitting. The authors concluded that a much larger set of climatologically unique site–years would be required for these models to be used in a predictive manner. However, it is uncertain about the minimum number of years data required to produce a reasonable prediction of spatial yield. Even with additional years of yield data in the models, it is still challenging to acquire spatial data consistently for the task. For example, the precision planting data was only available after 2016, and the Sentinel-2 data was only available after 2015. The results may be distorted with additional years of yield data in the leave-out-one-year analysis if the consistency of historical management actions and yield data quality are not guaranteed.

Future research should perhaps focus on the use of crop reflectance sensors acquiring data at regular intervals to determine how predicted yield changes with the vegetation indices, and use this information to tailor the rates of input during crop growth such as applying nitrogen (N) fertilizer at the V5 stage (5th leaf vegetative stage when the tassel is initiated) of maize. This idea has been exploited by a number of commercial crop sensors such as Crop Circle™ (Holland Scientific), OptRx™ (Ag Leader) and GreenSeeker® (Trimble Navigation Limited), which calculate N-rates and apply N fertilizer automatically across the paddock by scanning crop canopy [56]. Satellite imagery in our study is perhaps more useful in other arable systems such as outdoor vegetables and wheat. For arable farmers who grow high protein wheat, a mid-season N application may improve the areas that might require more N and thus move the product to the premium range. The application of statistical modelling techniques can then overcome the limitations of yield monitor data or protein sensor [57] and provide the opportunity to improve production before harvest.

5. Conclusions

This paper examined several statistical models (stepwise MLR, Cubist regression and a feed-forward neural network) for predicting maize-grain yield potential at the sub-paddock scale using precision farming data including yield monitor data, spatial yield data, seeding density, high-precision elevation and soil EC, as well as publicly-free multispectral satellite imagery (NASA's Landsat-8 and Sentinel-2 ESA's missions). The results showed that among other statistical models, the Cubist model produced the best statistical performance for this limited data set.

Although the data is limited, this work demonstrates that there is potential to integrate statistical modelling techniques and spatiotemporal data for site-specific crop management. However, given the model responses, yield data for additional years, and inclusion of further relevant variables (e.g., soil fertility, soil texture) may improve the model. This should be possible when more data sets become available. Once the value proposition of calibrating yield monitors and collecting and storing the data is understood, then the opportunity to embed a yield model in GIS software for consultants to use to advise clients may be possible.

Data consistency is a potential problem in the acquisition of useful satellite imagery at an appropriate growth stage for crop management in New Zealand due to cloud coverage. Nevertheless, UAVs (unmanned aerial vehicles) are increasingly being used in agricultural applications, and may offer alternatives to currently available satellite imagery by providing more relevant scales of data capture and the ability to capture information at more appropriate times of year. Whilst acquiring better data to improve the model might remain a challenge in the near future,

the application of the approach used in this study offers advantages over techniques that use spatial data collected from intensive and expensive grid sampling. The minimal costs associated with the approach employed in this study are thus more likely to be of commercial interest to New Zealand farmers. By predicting within-paddock yield potential, this study provided a statistical basis for delineating management zones for precision crop management decisions.

In the future, we will incorporate the models into on-farm GIS software and evaluate the financial viability of the management information through on-farm trials.

Author Contributions: A.H. supplied the data, G.J. and M.B. undertook site visits and collected samples, G.J. undertook the statistical analysis with the guidance of M.G., D.P. and M.B. All authors contributed to the writing.

Acknowledgments: The authors would like to express their thanks to the Foundation for Arable Research (FAR) for providing funding and data for conducting this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Millner, J.P.; Roskruge, N.R.; Dymond, J. The New Zealand arable industry. In *Ecosystem Services in New Zealand: Conditions and Trends*; Manaaki Whenua Press, Landcare Research: Lincoln, New Zealand, 2013; pp. 102–114.
2. Holmes, A.; Jiang, G. In Increasing Profitability & Sustainability of Maize using Site-Specific Crop Management in New Zealand. In Proceedings of the 14th International Conference on Precision Agriculture, Montreal, QC, Canada, 24–26 June 2018.
3. Khosla, R.; Westfall, D.; Reich, R.; Mahal, J.; Gangloff, W. Spatial variation and site-specific management zones. In *Geostatistical Applications for Precision Agriculture*; Springer: Berlin, Germaqny, 2010; pp 195–219.
4. Hedley, C.; Ekanayake, J.; McCarthy, A. In Precision irrigation: Trials to assess impacts on crop yield. In Proceedings of the 18th Australian Society of Agronomy Conference, Ballarat, Australia, 24–28 September 2017; pp 24–28.
5. Kitchen, N.; Drummond, S.; Lund, E.; Sudduth, K.; Buchleiter, G. Soil electrical conductivity and topography related to yield for three contrasting soil–crop systems. *Agrono. J.* **2003**, *95*, 483–495.
6. Guastaferro, F.; Castrignanò, A.; De Benedetto, D.; Sollitto, D.; Troccoli, A.; Cafarelli, B. A comparison of different algorithms for the delineation of management zones. *Precis. Agric.* **2010**, *11*, 600–620.
7. Blasch, G.; Spengler, D.; Itzerott, S.; Wessolek, G. Organic matter modeling at the landscape scale based on multitemporal soil pattern analysis using RapidEye data. *Remote Sens.* **2015**, *7*, 11125–11150.
8. Schirrmann, M.; Gebbers, R.; Kramer, E.; Seidel, J. Soil pH mapping with an on-the-go sensor. *Sensors* **2011**, *11*, 573–598.
9. Stadler, A.; Rudolph, S.; Kupisch, M.; Langensiepen, M.; van der Kruk, J.; Ewert, F. Quantifying the effects of soil variability on crop growth using apparent soil electrical conductivity measurements. *Eur. J. Agron.* **2015**, *64*, 8–20.
10. Drummond, S.T.; Sudduth, K.A.; Joshi, A.; Birrell, S.J.; Kitchen, N.R. Statistical and neural methods for site–specific yield prediction. *Trans. Asae* **2003**, *46*, 5.
11. Wang, X.; Miao, Y.; Dong, R.; Chen, Z.; Guan, Y.; Yue, X.; Fang, Z.; Mulla, D.J. Developing Active Canopy Sensor-Based Precision Nitrogen Management Strategies for Maize in Northeast China. *Sustainability* **2019**, *11*, 706.
12. Sudduth, K.; Drummond, S.; Birrell, S.J.; Kitchen, N. Analysis of spatial factors influencing crop yield. *Precis. Agric.* **1996**, *3*, 129–139.
13. Liu, J.; Goering, C.; Tian, L. A neural network for setting target corn yields. *Trans. Asae* **2001**, *44*, 705.
14. Aviv, T.; Lundsgaard-Nielsen, V. In Ensemble of Cubist models for soy yield prediction using soil features and remote sensing variables. In Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017.
15. Noi, P.; Degener, J.; Kappas, M. Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. *Remote Sens.* **2017**, *9*, 398.

16. Quinlan, J.R. In Learning with continuous classes. In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Tasmania, 16–18 November 1992; World Scientific: Singapore, 1992; pp. 343–348.
17. Quinlan, J.R. In Combining instance-based and model-based learning. In Proceedings of the Tenth International Conference on Machine Learning, Amherst, MA, USA, 27–29 July 1993; pp. 236–243.
18. Kuhn, M.; Weston, S.; Keefer, C.; Coulter, N. Cubist models for regression. R package Vignette R package version 0.0 2012. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.398.3360&rep=rep1&type=pdf (accessed on 12 May 2019).
19. Walton, J.T. Subpixel urban land cover estimation. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 1213–1222.
20. Chappell, P.R. *The Climate and Weather of Waikato*; NIWA: Auckland, New Zealand, 2013; p. 40.
21. Molloy, L. Soils in the New Zealand Landscape: The Living Mantle, 2nd ed.; In *New Zealand Society of Soil Science*; Mallinson Rendel Publishers Ltd.: Wellington, New Zealand, 1998; pp. 33–50.
22. Whelan, B.; Taylor, J. *Precision Agriculture for Grain Production Systems*; Csiro Publishing: Clayton, Australia 2013.
23. Martínez, M.; Joel, L. Relationship between crop nutritional status, spectral measurements and Sentinel 2 images. *Agron. Colomb.* **2017**, *35*, 205–215.
24. Frampton, W.J.; Dash, J.; Watmough, G.; Milton, E.J. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS J. Photogramm. Remote Sens.* **2013**, *82*, 83–92.
25. Castaldi, F.; Chabrillat, S.; van Wesemael, B. Sampling strategies for soil property mapping using multispectral Sentinel-2 and hyperspectral EnMAP satellite data. *Remote Sens.* **2019**, *11*, 309.
26. Csillik, O.; Belgiu, M. In Cropland mapping from Sentinel-2 time series data using object-based image analysis. In Proceedings of the 20th AGILE International Conference on Geographic Information Science Societal Geo-Innovation Celebrating, Wageningen, The Netherlands, 9–12 May 2017.
27. Almutairi, B.; El battay, A.; Ait Belaid, M.; Musa, N. Comparative study of SAVI and NDVI vegetation indices in Sulaibiya Area (Kuwait) using Worldview satellite imagery. *Int. J. Geosci. Geomat.* **2013**, *1*, 2052–5591.
28. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309.
29. Lund, E.; Maxton, C. In Proximal sensing of soil organic matter using the Veris® OpticMapper™. In Proceedings of the 2nd Global Workshop on Proximal Soil Sensing, Montreal, QC, Canada, 15–18 May2011; pp. 15–19.
30. Hurst, C.; Lovell, S.; Lund, T.; Holmes, A. Precise surveying of soil productivity indicators using on-the-go soil sensors. In *Moving Farm Systems to Improved Attenuation*; Currie, L.D., Burkitt, L.L., Eds.; Occasional Report, Massey University: Wageningen, the Netherlands, 2015. Available online: http://flrc.massey.ac.nz/publications html (accessed on 20 December 2018).
31. Booker, J.W.; *Production, Distribution and Utilisation of Maize in New Zealand*; Lincoln University: Lincoln, UK, 2009.
32. Sudduth, K.A.; Drummond, S.T.; Myers, D.B. In Yield editor 2.0: Software for automated removal of yield map errors. In Proceedings of the 2012 ASABE Annual International Meeting Sponsored by ASABE Hilton Anatole, Dallas, TX, 29 July–1 August 2012; American Society of Agricultural and Biological Engineers: St. Joseph, MI, USA, 2012; p. 1.
33. Blackmore, S. Remedial correction of yield map data. *Precis. Agric.* **1999**, *1*, 53–66.
34. Spekken, M.; Anselmi, A.; Molin, J. A simple method for filtering spatial data. In *Precision Agriculture'13*; Springer: Berlin, Germany, 2013; pp. 259–266.
35. Matheron, G. Principles of geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266.
36. Webster, R.; Oliver, M.A. Sample adequately to estimate variograms of soil properties. *J. Soil Sci.* **1992**, *43*, 177–192.
37. Chung, S.; Sudduth, K.; Drummond, S. Determining yield monitoring system delay time with geostatistical and data segmentation approaches. *Trans. Asae* **2002**, *45*, 915.
38. Mulla, D.J. Using geostatistics and spectral analysis to study spatial patterns in the topography of southeastern Washington State, USA. *Earth Surf. Process. Landf.* **1988**, *13*, 389–405.

39. Pullman, W. 2 Mapping and Managing Spatial Patterns In Soil Fertility and Crop Yield. In *Soil Specific Crop Management*; Robert, P., Larson, W., Rust, R., Eds.; American Society of Agronomy: Madison, WI, USA, 1993; pp.15–26.
40. Oliver, M. An overview of geostatistics and precision agriculture. In *Geostatistical Applications for Precision Agriculture*; Springer: Berlin, Germany, 2010; pp. 1–34.
41. Maldaner, L.F.; Corrêdo, L.P.; Tavares, T.R.; Mendez, L.G.; Duarte, C.; Molin, J.P. Identifying and filtering out outliers in spatial datasets. In Proceedings of the 14th International Conference on Precision Agriculture, Montreal, QC, Canada, 24–27 June 2018.
42. Licht, M.A.; Lenssen, A.W.; Elmore, R.W. Corn (*Zea mays* L.) seeding rate optimization in Iowa, USA. *Precis. Agric.* **2017**, *18*, 452–469.
43. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26.
44. Nakama, T. In Comparisons of single-and multiple-hidden-layer neural networks. In *International Symposium on Neural Networks*; Springer: Berlin, Germany, 2011; pp. 270–279.
45. Hayashi, Y.; Sakata, M.; Gallant, S.I. Multi-layer versus single-layer neural networks and an application to reading hand-stamped characters. In Proceedings of the International Neural Network Conference, Vienna, Austria, 17–19 September, 2019; Springer: Berlin, Germany, 1990; pp 781–784.
46. Al-kaf, H.A.G.; Chia, K.S.; Alduais, N.A.M. A comparison between single layer and multilayer artificial neural networks in predicting diesel fuel properties using near infrared spectrum. *Pet. Sci. Technol.* **2018**, *36*, 411–418.
47. Dao, V.N.; Vemuri, V. A performance comparison of different back propagation neural networks methods in computer network intrusion detection. *Differ. Equ. Dyn. Syst.* **2002**, *10*, 201–214.
48. Bergmeir, C.N.; Benítez Sánchez, J.M. In *Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS*; American Statistical Association: Alexandria, VA, USA, 2012.
49. Ripley, B.D.; Hjort, N. *Pattern Recognition and Neural Networks*; Cambridge university press: Cambridge , UK, 1996.
50. Chang, J.; Clay, D.E.; Dalsted, K.; Clay, S.; O'Neill, M. Corn (*Zea mays* L.) yield prediction using multispectral and multidate reflectance. *Agron. J.* **2003**, *95*, 1447–1453.
51. Six, J.; Conant, R.; Paul, E.A.; Paustian, K. Stabilization mechanisms of soil organic matter: Implications for C-saturation of soils. *Plant Soil* **2002**, *241*, 155–176.
52. Lund, E.; Christy, C.; Drummond, P. Practical applications of soil electrical conductivity mapping. *Precis. Agric.* **1999**, *99*, 771–779.
53. Blackmore, S.; Godwin, R.J.; Fountas, S. The analysis of spatial and temporal trends in yield map data over six years. *Biosyst. Eng.* **2003**, *84*, 455–466.
54. Changere, A.; Lal, R. Slope position and erosional effects on soil properties and corn production on a Miamian soil in central Ohio. *J. Sustain. Agric.* **1997**, *11*, 5–21.
55. Yuan, M.; Fernández, F.G.; Pittelkow, C.M.; Greer, K.D.; Schaefer, D. Tillage and Fertilizer Management Effects on Phosphorus Runoff from Minimal Slope Fields. *J. Environ. Qual.* **2018**, *47*, 462–470.
56. Craigie, R.; Yule, I.; McVeagh, P. *Crop. Sensing for Nitrogen Management*; Foundation for Arable Research: Christchurch, New Zealand, 2013.
57. Long, D.S.; Engel, R.E.; Carpenter, F.M. On-combine sensing and mapping of wheat protein concentration. *Crop. Manag.* **2005**, *4*, doi:10.1094/CM-2005-0527-01-RS.