Regular Paper

Development of Yield Prediction Model Generation Process for Fruit Vegetables in Plant Factories

Yuki Todate*, Michiko Oba**, and Mitsuru Takamori***

*Graduate School of Systems Information Science, Future University Hakodate, Japan
** School of Systems Information Science, Future University Hakodate, Japan
*** Apure Inc, Japan
{g2119026, michiko}@fun.ac.jp, takamori@agricc.biz

Abstract – Plant factories in Japan have become increasingly popular in recent years. In these factories, data are collected using production management systems and both external and internal environmental sensors. However, predicting yields for fruit and vegetable crops is difficult because these crops have unique biological characteristics, and their growth depends on weather conditions. Thus, the purpose of this study is to develop a yield model generation process for fruits and vegetables in plant factories. By defining a process for generating predictive models, we aim to improve the efficiency and accuracy of their development, as well as to make them applicable to various fruit and vegetable crops and to various facilities. This paper reports on the results of our application of the proposed predictive model generation process to develop a model for mini cucumbers and mini tomatoes, which we interpreted as being representative of other fruit and vegetable crops. Experimental results show that the proposed model generation process can be applied to various crops. In addition, it was confirmed that the tsfresh Python package, which we used to automatically extract features from time-series data, improved the prediction accuracy.

Keywords: Yield Prediction, Plant Factory, Fruit Type Vegetables, Statistical Modeling, tsfresh

1 INTRODUCTION

In recent years, there has been great interest in the promising field of protected horticulture, including next-generation horticulture. This new approach involves equipping sites with advanced environmental control devices utilizing information and communications technology (ICT) and other technologies, aiming to integrate facilities and efficiently use local resources and energy [1]. To extend the shipping period for vegetables, horticulture facilities in Japan have been upgraded from plastic tunnels and rain shelters to greenhouses and then to plant factories with highly controlled greenhouse environments.

Plant factories are a form of horticultural agriculture that enables year-round, planned production of vegetables and other plants through advanced environmental control and growth forecasting [2]. They collect and accumulate large quantities of data using production management systems and environmental sensor data inside and outside the factory. Plant factories can be broadly classified into two types: (1) the “sunlight-based” type, in which plants are cultivated in greenhouses, or similar structures using sunlight, supplemented by artificial light and technology to prevent high temperatures in summer and (2) the “fully artificial light-based” type, in which plants are cultivated in a closed environment without sunlight. The main crops grown in the artificial light type are leafy greens, such as spinach and lettuce, and fruit crops such as tomatoes and cucumbers are commonly grown in facilities of the sunlight type. Plant factories in Japan are still developing, and there are still major problems such as low yield per farm area [3][4] and low labor productivity [5][6]. Therefore, it is necessary to establish a new Japanese-style cultivation platform, standardize facilities, and promote research benefiting the average of cultivation expertise.

One of the challenges faced by plant factories is the difficulty of predicting crop yield [7][8]. Because fruit vegetables are mainly grown in sunlight-type plant factories, they are subject to variations in sunlight, temperature, and other seasonal weather conditions, significantly affecting their growth. Given these factors and because of their biological characteristics, yield prediction for fruit vegetables is difficult. Compared with leafy vegetables, fruit vegetables are more affected by environmental conditions for a longer duration as they have both flowering and fruiting periods. Moreover, as fruit vegetables can be harvested from a single seedling several times, predicting their yield is a challenge. Yield prediction is important to match market demand and prevent overproduction.

Yield prediction and the relationship between yield and environmental factors have been extensively studied for crops grown in open fields or greenhouses with simple environmental controls and major crops such as paddy and wheat. However, we have not found any studies on fruit and vegetable crops grown in plant factories.

Therefore, the purpose of this study is to propose a model generation process for fruit vegetables grown in plant factories. The model generation process comprised six sequential steps: (1) data selection for model generation, (2) data preprocessing, (3) data visualization, (4) feature design, (5) selection of prediction methods, and (6) model optimization. By defining the process for generating predictive models, we aim to improve the efficiency and accuracy of their development as well as to make them applicable to various fruit vegetable crops and other facilities.

2 RELATED RESEARCH

2.1 Prediction Method

Many studies have investigated yield prediction using various machine learning techniques, such as artificial neural networks and boosted regression trees [9]-[12]. Such techniques have achieved high prediction accuracy; however, machine learning requires a large amount of training data, which limits its application.

Apart from machine learning, statistical modeling has also been used as a prediction method. Related studies include those by Hoshi et al. [7] and Okuno et al. [13]. In 2000, Hoshi et al. predicted the daily yields of tomatoes grown in a greenhouse using topological case-based modeling (TCBM) [14] and multiple regression analysis. TCBM was the most accurate, with an average absolute error of 26 %. In 2018, Okuno et al. proposed combining machine learning methods and statistical modeling for asparagus yield prediction. Using a Bayesian network as a machine learning method and multiple regression analysis for statistical modeling, the authors generated a regression model that can be effectively used as a sound basis for prediction results and estimated yields by capturing increasing and decreasing trends. However, there is insufficient information on the actual operation due to the limited number of prediction methods and crops evaluated.

2.2 Feature Generation

In model development, it is important to consider the features to be incorporated. Many of the abovementioned related studies incorporated features related to environmental data, such as temperature, humidity, light, and precipitation. In general, only the most basic statistics, such as average, maximum, and minimum values over a period of time, were used to generate features in the relevant literature. However, this method has the drawback that finding factors related to yield is difficult because of the small variation in features, and the accuracy of the generated models is limited.

2.3 Research Tasks

As mentioned previously, the model generation process formulated in this study consisted of six subprocesses, performed in the following order: (1) data selection for model generation, (2) data preprocessing, (3) data visualization, (4) feature design, (5) selection of prediction methods, and (6) model optimization. Based on the issues discussed in related studies, designing the features and selecting a prediction method are the major challenges. Therefore, the two research tasks in this study were:

(1) Selection of a method for building a predictive model

Preparing a large dataset is difficult for plant factories due to significant environmental and cultivar variations caused by the equipment. Therefore, it is necessary to build a predictive model that can be used even with a relatively small amount of data. In other cases, it is also necessary to select a method whereby the process by which the final prediction result is calculated is easily understood. It is important to develop models that plant factory employees are able to clearly understand in practical use because unconvincing prediction models are generally unacceptable to farmers.

(2) Selection of feature extraction method

Actual yield data are time-series data. A defining characteristic of time series data is that the datasets are closely related; therefore, it is necessary to generate features that take this into account. Thus, it is necessary to calculate various features, such as median, variance, standard deviation, Fourier transform, and autocorrelation coefficients.

3 PROPOSED METHOD

3.1 Approaches

The approaches employed to carry out the research tasks described in Section 2.3 are as follows.

(1) We selected a prediction model that works even with a small amount of data and has a comprehensible model structure.

(2) We also selected methods for extracting various features from time-series data.

A more detailed description of these approaches is given in Section 3.2. The model generation process that incorporates all these approaches is also described.

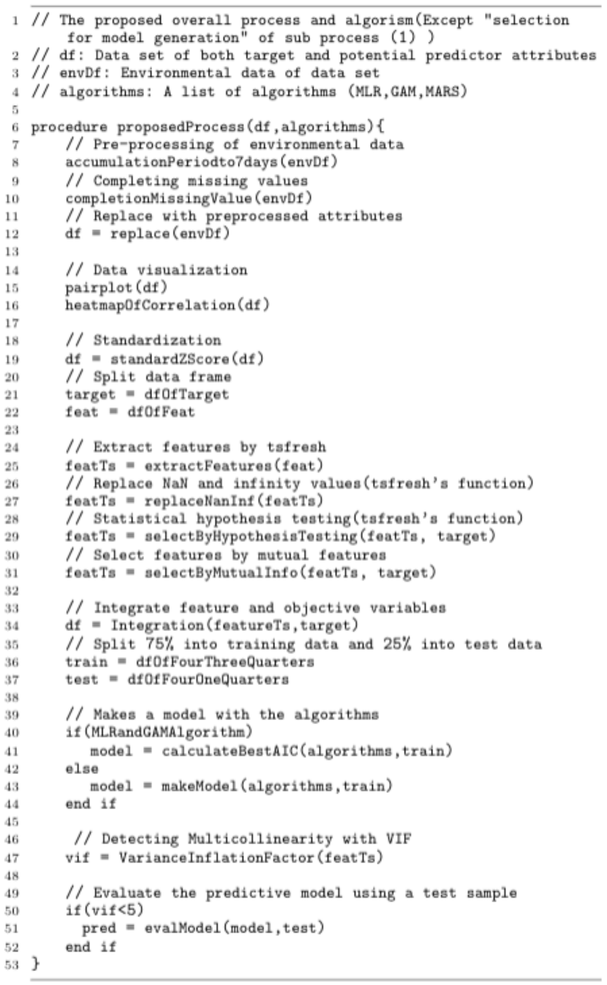
3.2 Model Generation Process

As mentioned, the model generation process involves six subprocesses. In this research, these processes were formulated as simple arithmetic procedures to identify factors related to yield and generate a model that can accurately predict yield according to various facility-specific conditions. In each subprocess, a representative fruit was used to develop the model. For subprocess (1) to (6), we show the process of searching for an optimal model generation process for the clauses corresponding to the numbers. Figure 1 shows the algorithm of the yield prediction model generation process, which is the proposed method.

3.2.1 Data Selection

In this study, we decided to use three major types of data as candidate features of two crops: actual data of past yields, environmental data in the facility, as well as outdoor weather data. There are two main reasons for this.

The first reason is that the data are continuously collected and recorded in plant factories. Some related studies have used data on crop appearances, such as the normalized difference vegetation index (NDVI) and crop stem diameter measurement data [9]-[12]. However, this method requires the installation of new equipment for sensing the appearance data, which results in high data collection costs. On the other hand, environmental data inside and outside structures and actual yield data are generally obtained and recorded in plant factories, thus collecting this data is easy.



```
// The proposed overall process and algorism(Except "selection
    for model generation" of sub process (1) )
// df: Data set of both target and potential predictor attributes
// envDf: Environmental data of data set
// algorithms: A list of algorithms (MLR,GAM,MARS)

procedure proposedProcess(df,algorithms){
    // Pre-processing of environmental data
    accumulationPeriodto7days(envDf)
    // Completing missing values
    completionMissingValue(envDf)
    // Replace with preprocessed attributes
    df = replace(envDf)

    // Data visualization
    pairplot(df)
    heatmapOfCorrelation(df)

    // Standardization
    df = standardZScore(df)
    // Split data frame
    target = dfOfTarget
    feat = dfOfFeat

    // Extract features by tsfresh
    featTs = extractFeatures(feat)
    // Replace NaN and infinity values(tsfresh's function)
    featTs = replaceNanInf(featTs)
    // Statistical hypothesis testing(tsfresh's function)
    featTs = selectByHypothesisTesting(featTs, target)
    // Select features by mutual features
    featTs = selectByMutualInfo(featTs, target)

    // Integrate feature and objective variables
    df = Integration(featureTs,target)
    // Split 75% into training data and 25% into test data
    train = dfOfFourThreeQuarters
    test = dfOfFourOneQuarters

    // Makes a model with the algorithms
    if(MLRandGAMAlgorithm)
       model = calculateBestAIC(algorithms,train)
    else
       model = makeModel(algorithms,train)
    end if

     // Detecting Multicollinearity with VIF
    vif = VarianceInflationFactor(featTs)

    // Evaluate the predictive model using a test sample
    if(vif<5)
      pred = evalModel(model,test)
    end if
}
```

Figure 1: Algorithm of the proposed method

The second reason is that these three types of data were reported to be effective as features in a related study that predicted the yield of fruit and vegetable crops in greenhouses [7][9][14]. Greenhouses and plant factories differ in terms of environmental control methods, but they share the primary characteristic of growing crops “in the facility.” Therefore, we use these attributes because we believe they are likely to be effective for fruit and vegetable crops in plant factories. We also use the weekly yield as the objective variable and the “yield one week ago” relative to this objective variable as an attribute related to the data of past actual yields. We adopted this approach based on the results of a study previously conducted by the author [15], in which a single correlation analysis of weekly yield and past actual yield data showed that the yield one week previous had the highest correlation with the objective variable. Because multicollinearity may occur if multiple similar attributes of past yields are included and it is necessary to reduce the dimensionality of the features, we decided to use only the yield of one week before based on the results of the previous correlation analysis.

For these two reasons, environmental data inside and outside the greenhouse and past actual yield data were used.

3.2.2 Data Preprocessing

Data preprocessing consisted of three major components.

First, we process the integration period of environmental data for the attributes, created by using the accumulated value from 7 day prior to the integration up to the morning of the reference date (the day before the forecast date). This period corresponded to the fruiting period, which is the period from flowering to harvest for both mini cucumbers and mini tomatoes. As environmental factors during this period have been shown to have a significant impact on yields [16], the data during this period were used. In addition, we create attributes for daytime hours only as plants grow mainly when there is sunlight for photosynthesis. Because the times of sunrise and sunset change throughout the year, we use the annual average of the time period of daylight in the region where the facility is located.

The second component is the replacement of the missing values. Missing values due to malfunctioning sensor devices were supplemented by the average value for the three days before and after the missing value occurred.

The third component is the integration of data scales. To do this, we standardize the process using a robust z-score. Robust z-scores were chosen as their median is not affected by the shape of the distribution, and the quartiles are not affected by the outliers at both ends of the distribution (statistic of variability).

3.2.3 Data Visualization

Data visualization facilitates the analysis of relationships between objective variables and features and identifies outliers. Scatter plots can be used to visualize the data. A scatter plot takes one feature on the x-axis and another on the y-axis and plots dots at each data point. In this study, as there were more than three features, a scatter plot matrix (paired plot) was used to plot all possible feature combinations. In addition, to obtain a clearer picture of the correlations between the attributes, a heat map of the correlation matrix, as well as a scatter plot diagram, was used.

3.2.4 Feature Design

The design of features involves two steps: extraction and selection.

First, we discuss the feature extraction method. Various features are generated by extracting the features of time series data, as described in Section 3.1. In time series data, the observed values and the times of observation are recorded, and it is necessary to generate features that capture the characteristics of the order of the data and the backward/forward relationship. To extract various features specific to time series data, we used a suitable Python package called tsfresh (Time Series FeatuRe Extraction on the basis of Scalable Hypothesis tests) [17], which includes feature extraction and feature selection algorithms for time series analysis. Figure 2 shows a schematic of our process flow using this package. tsfresh provides 63 characterization methods for the extraction of 794 distinct features. For example, there are mean, maximum, minimum, number of peaks, median, standard deviation, and

Fourier transform features, as well as features using autocorrelation coefficients and time-reversal symmetry features. We used tsfresh because it can generate a comprehensive set of features specific to time series data, and the extraction and selection processes can be parallelized to significantly reduce execution time. In addition, tsfresh has been used in many research papers and in various applications such as disease prediction, machine fault detection, and traffic volume prediction.

Next, we describe four methods of feature selection.

The first method uses the select_features module provided by tsfresh, which selects features using statistical hypothesis testing so that only features that are likely to have statistically significant differences are selected. The significance testing methods for features included Fisher's exact test of independence, the Kolmogorov–Smirnov (KS) test for binary features, the KS test for continuous features, and a Kendall rank correlation coefficient test, depending on the type of supervised machine learning problem (classification or regression) and the type of feature (categorical or continuous) [18]. The result of validation by these methods is a vector of p-values, which quantifies the importance of each feature for predicting labels and targets. In Fig. 2, this corresponds to the third stage, "p-values." These p-values were evaluated based on the Benjamini–Yekutieli procedure to determine which features to retain. This stage corresponds to the last stage, "Selected features," in Fig. 2.

The second method is based on mutual information (MI). MI calculates the dependence of the product of the simultaneous distribution P(X,Y) and the individual distributions P(X)P(Y) between one feature X and another feature Y. If they are independent of each other, MI is zero. According to the MI, the number of features should be chosen to be "the number of samples in the training data/10." This is d to the amount of training data is not appropriate in terms of noise pattern learning and learning speed.

The third method is the stepwise method based on the Akaike Information Criterion (AIC), which is a common statistical variable selection method. The AIC is used to find the best combination of features in the prediction models of multiple regression analysis and generalized additive models, which will be explained in the next section. The multivariate adaptive regression spline, which is another prediction because having too many features compare method used, automatically performs feature selection in the algorithm, so this feature selection method is applied to both multiple regression analysis and generalized additive models

In the fourth method, we use the variance inflation factor (VIF) to check whether multicollinearity occurs. If multicollinearity is present, the corresponding variable is deleted and the feature is selected.

The mentioned above, we use four methods of feature selection.

3.2.5 Prediction Method Selection

Our methods, requiring only a small amount of data and involving a comprehensible model structure, are described in this section. The selection criteria for the prediction methods are as follows: (i) the regression structure of the features can be searched and evaluated, and (ii) a model with high prediction accuracy and low data requirement can be constructed. Three modeling methods that satisfied both criteria were selected for the study.

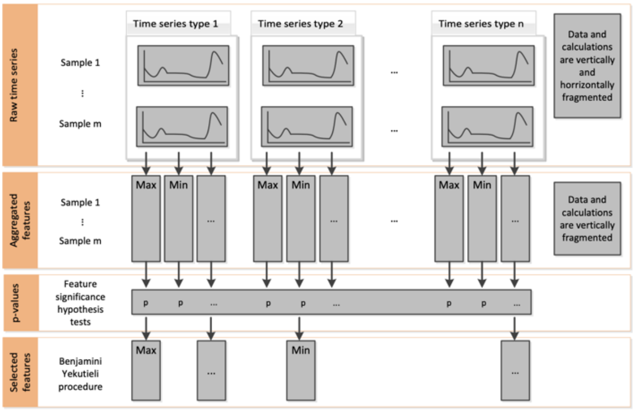

Raw time series
Sample 1
Sample m
Time series type 1
Time series type 2
Time series type n
Data and calculations are vertically and horrizontally fragmented
Aggregated features
Sample 1
Sample m
Max
Min
...
Data and calculations are vertically fragmented
p-values
Feature significance hypothesis tests
p
p
...
Selected features
Benjamini Yekutieli procedure
Max
Min
...

Figure 2: Data processing flow from feature extraction to selection (adapted from [18])

The first method was multiple linear regression (MLR), which is a general statistical method for predicting continuous values of objective variables using two or more features. MLR is widely used to predict yields of various crops, including fruits and vegetables [7][13][16]. MLR was selected because it has some degree of prediction accuracy. To determine the best feature combination, a stepwise method based on the Akaike information criterion (AIC), a common statistical variable selection method, was implemented in our MLR approach. The second method used was the generalized additive model (GAM), selected because it can predict with an accuracy similar to that of a machine learning model. It also retains the advantage of a linear model, where the relationship between objective variables and features can be easily determined. Similar to MLR, AIC was also implemented in GAM. The third method used was multivariate adaptive regression splines (MARS). Compared to GLM, MARS can explicitly represent the interaction between features, including tipping points in tree structures [14].

3.2.6 Model Optimization

The target data were divided into training data (75 %) and test data (25 %). The datasets were evaluated using the leave-one-out method. The correlation coefficient (R) and the mean absolute error (MAE) were used as performance indicators of the regression model, where R measures the linear relationship between the predicted value and the measured value, and MAE is the average value (in physical units) of the difference between the predicted values. As percent yield varies among crops, MAE was expressed as a percentage relative to the average yield.

We evaluated the prediction accuracy of the developed model using R and MAE, and the method with the highest prediction accuracy was selected.

4 EXPERIMENTS

4.1 Target Facility

The experimental facility used in this study was a solar-powered plant factory located in Hakodate, Hokkaido (hereafter referred to as "Plant Factory A"). Plant Factory A owns two greenhouses that produce and sells hydroponically grown fruits and vegetables (7 fruits and 17 leafy vegetables). Apart from collecting environmental data, such as temperature, humidity, CO2 concentration, and nutrient concentration in the hydroponic solution, Plant Factory A collects external weather data, such as temperature, humidity, precipitation, and light intensity. Monitoring both external and internal environmental parameters is useful for controlling the environment of the facilities.

We confirmed the demand for yield prediction of fruits and vegetables in Plant Factory A from interviews with employees in charge of management and employees in charge of actual production.

4.2 Target Crops

We used mini cucumber (Larino White) as the base case, and mini tomato (Aiko) as the target crop to verify the versatility of the proposed forecasting model generation process. These two crops are among the top three major crops cultivated in horticulture facilities in Japan. Because they are also the main crops grown in Plant Factory A, their cultivation area is large, and they are shipped almost every day. Therefore, there is a high demand for the development of a yield prediction model for these crops.

4.3 Target Data and Problem Setting

In this subsection, we describe the details of the data used to construct the model and the problem setting based on interviews with plant factory employees.

4.3.1 Target Data and Preprocessing

The data used for model building and evaluation were all 66 weeks of the period from February 2018 to June 2019. The selection of attributes was based on the approach described in Section 3.2.1. Table 1 lists the candidate attributes for the feature values. As explained in Section 3.2.2, preprocessing such as changing the integration period of the environmental data was performed for the attributes in Table 1.

Table 1: Potential predictor attributes

Attribute code name	Attribute name and description
YW	Yield weight (kg)
PYW	Past yield weight a week ago (kg)
AT	Air temperature (°C)
AH	Air humidity (%)
SR	Solar radiation (kWh m^{-2})
FT	Temperature in the facility (°C)
FCD	Carbon dioxide concentration in the facility (ppm)
FS	Saturation humidity in the facility, in free water vapor capacity per 1 m^3 of air (g/m^3)

4.3.2 Problem Setting

From the interview with the management and production employees, we set the yield for a single week from the day following the day on which the yield was predicted (hereafter, the prediction date) as the objective variable, hereafter referred to as the weekly yield and predicted in this study using environmental and production data up to the forecast date. Plant Factory A had not been able to predict the yield of fruit and vegetable crops at all, so it is necessary to bring them to the stage of utilization as soon as possible. Based on these circumstances and the general accuracy standards for the practical use of prediction models, we aim to achieve a prediction accuracy of R = 0.8 or better and MAE = 20.0 % or less.

4.4 Data Visualization

4.4.1 Visualization of Weekly Yield Trends

Figure 3 shows a graph of the weekly yields of the predicted targets. The average weekly yield was 135 kg for mini cucumbers and 24.5 kg for mini tomatoes. Compared to mini tomatoes, mini cucumbers showed a larger variation in yield. Trend pattern analysis of the time series data showed that the yield of mini cucumbers fluctuated seasonally. In Fig. 3, the period of large increase in the yield of mini cucumbers corresponds to the summer season, indicating that the yield is at its maximum during the summer when the light intensity increases.

On the other hand, the yield of mini tomatoes showed cyclic variation and only a slight trend of seasonal variation. Fig. 3 shows that there were several cycles of large yield increase and decrease, which may be due to the biological characteristics of fruit crops: new fruits are produced after a period of time. To investigate the reason for the lack of seasonal variation in the tomatoes, we interviewed the employees of the plant factory and found that they had been neglecting the cultivation management of the tomatoes because they had been focusing on the cultivation of other crops during the period. Specifically, they left some plants long after they should have been discarded, and were late in responding to pest damage.

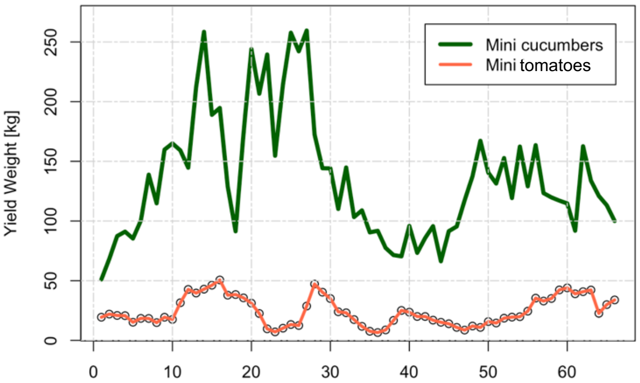


Figure 3: Weekly yield of the two crops

These results suggest that when cultivation management is properly carried out, the tendency of seasonal variation is such that the yield is highest in the summer season, as in the case of mini cucumbers.

4.4.2 Visualization of Attribute Relationships

The relationships between the objective variables and features and are described in this section. Figures 4 and 5 show a heat map table of the correlation coefficients. Spearman's rank correlation coefficient was used to calculate the single correlation coefficient.

Eight attributes, including the objective variables, were used in the analysis. Preprocessing was carried out, such as changing the integration period of environmental data, processing of missing values, and standardization. Table 2 lists the attributes created for both crops.

The strongest correlation for mini cucumbers was found for the weekly yield of the previous week with a single correlation coefficient of 0.79, followed by carbon dioxide concentration (−0.71) and air temperature (0.69). The strongest correlation for mini tomatoes was found for the weekly yield of the previous week, with a single correlation coefficient of 0.83, followed by carbon dioxide concentration (−0.34) and air temperature (0.33).

The same attributes were found to be strongly correlated with yield in both mini cucumbers and mini tomatoes. By adding these highly correlated attributes into the predictive model, the accuracy of the model would most likely be improved. Based on the values of the correlation coefficients for each attribute, there were significant differences between the two crops. Therefore, we inferred that the degree of influence of environmental factors varies with the crop.

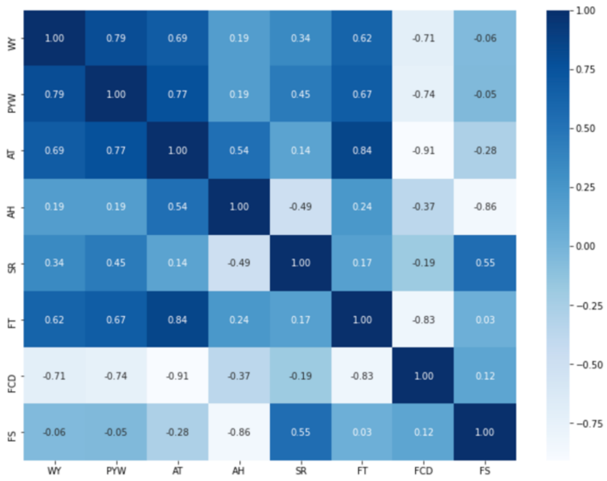


Figure 4: Heat map table
of correlation coefficients for mini cucumbers

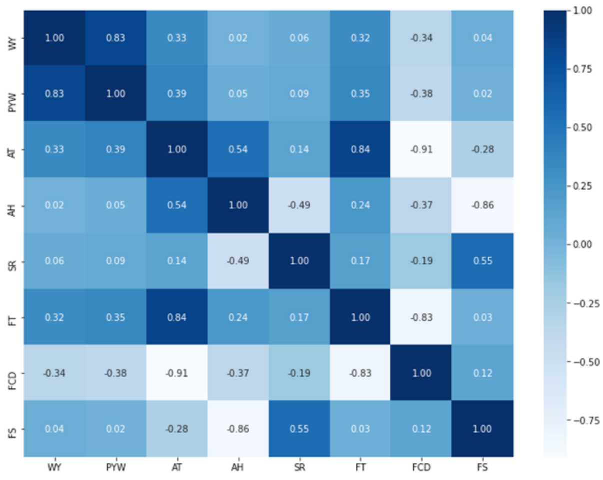


Figure 5: Heat map table
of correlation coefficients for mini tomatoes

4.5 Feature Design

First, we used the extract_features module of tsfresh to extract features reflecting the characteristics of the time series data from the attributes in Table 1. As a result, a combined total of 5278 features were extracted for mini cucumbers and mini tomatoes. Next, by statistical hypothesis testing using the tsfresh select_features module, we narrowed down the number of features to 82 for mini cucumbers and 69 for mini tomatoes. Next, using mutual features, we selected five features each for mini cucumber and mini tomato. Finally, the VIF values were calculated, and because all the features were less than 5, the suspicion of multicollinearity was low, and we decided to use all five features.

The features thus obtained are shown in Tables 2 and 3. The feature names are composed of the following three elements: (1) the time series attribute from which the feature is extracted, (2) the name of the feature calculator used to extract the feature, and (3) the key-value pairs of the parameters that make up the feature calculator:

[kind] _ [calculator] _ [parameterA] _ [valueA] _ [parameterB] _ [valueB]

For example, the feature name carbon-DioxideConcentration _cwt_coefficients_widths_(2, 5, 10, 20)_coeff_0_w_10 for mini cucumbers in Table 2 has kind = carbon-DioxideConcentration, calculator = cwt_coefficient, parameter width = (2, 5, 10, 20), parameter coeff = 0, and parameter w = 10, which represents the continuous wavelet transform of the Ricker wavelet of the yield of one week ago (PYW).

Next, we describe the selected features. First, we detail the features that reflect the characteristics unique to time series data. These are C_CDC_TS, C_SR_T, T_PYW_TS, T_HT_TS, and T-CDC_TS, which use the one-dimensional discrete Fourier transform and the Ricker wavelet continuous wavelet transform, shown in Tables 2 and 3. These features represent periodic changes. This periodicity was observed in both crops as seasonal and cyclic variation during the analysis

Table 2: Mini cucumber features extracted by tsfresh

Notation	Feature name	Feature description
C_PYW_TS	pastYieldWeight_average_period_7	Cumulative value of past yields from 7 days before the forecast date
C_AT_TS	airTemp_average_period_7	Cumulative value of air temperature from 7 days before the forecast date
C_CDC_TS	carbonDioxideConcentration _cwt_coefficients_widths_(2, 5, 10, 20)_coeff_0_w_10	Previous week's yield of the objective variable, Calculates a continuous wavelet transform for the Ricker wavelet, also known as the Mexican hat wavelet
C_SR_TS	solarRadiation_fft_coefficient_coeff_0_attr_angle	Fourier coefficients of the one-dimensional discrete fast Fourier transform about solar radiation
C_HT_TS	houseTempAve_average_period_7	Cumulative value of temperature in the facility from 7 days before the forecast date

Table 3: Mini tomato features extracted by tsfresh

Notation	Feature name	Feature description
T_PYW_TS	pastYieldWeight_cwt_coefficients_widths_(2, 5, 10, 20)_coeff_0_w_2	Previous week's yield of the objective variable; calculates a continuous wavelet transform for the Ricker wavelet
T_AT_TS	airTemp_average_period_7	Cumulative value of air temperature from 7 days before the forecast date
T_SR_TS	solarRadiation_abs_energy	Sum over the squared values of solar radiation
T_HT_TS	houseTemp_fft_coefficient_coeff_0_attr_angle	Fourier coefficients of the one-dimensional discrete fast Fourier transform about temperature in the house
T_CDC_TS	carbonDioxideConcentration_fft_coefficient_coeff_0_attr_angle	Fourier coefficients of the one-dimensional discrete fast Fourier transform about carbon dioxide concentration

of weekly yield trends, which explains the selection of these features. Other features selected were C_PYW_TS, C_AT_TS, C_HT_TS, and T_AT_TS, which use cumulative values from 7 day before the forecast date, and T_SR_TS, which represents the addition of squared values.

Subsequent model tuning and selection of the best prediction method were performed using these features.

4.6 Model Optimization

Of the 66 available data, 50 (about 75 %) were used as training data and 16 (about 25 %) were used as test data (Fig. 6).

First, MLR, GAM, and MARS methods were trained on the data. We searched for the best combination of features and hyperparameters in the data. Tables 4 and 5 show the combinations of features and hyperparameters with the highest accuracy. Tables 4 and 5 summarize the results of the most accurate feature combinations and optimal hyperparameters. The feature values selected differed depending on the target crop and prediction method. For each feature, the attributes that were found to have high correlation with yield in the correlation analysis tended to be selected more frequently.

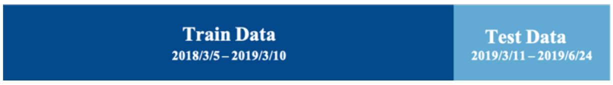


Figure 6: Data partitioning method

Table 4: Model tuning results (feature selection)

Crop	Feature		
	MLR	GAM	MARS
Cucumber	C_PYW_TS C_CDC_TS	C_PYW_TS C_CDC_TS	C_PYW_TS
Tomato	T_PYW_TS T_SR_TS T_CDC_TS	T_PYW_TS	T_PYW_TS T_SR_TS

Table 6 shows the prediction accuracy results for the test data using the tuned features and hyperparameters. For mini cucumbers, the highest prediction accuracy was achieved by MARS, and for mini tomatoes, it was achieved by GAM.

To evaluate the effectiveness of tsfresh, we compared the results of feature extraction using tsfresh and the results of feature extraction using the features in Table 2 without tsfresh. The prediction method that showed the highest accuracy for each crop was used for comparison. The results are shown in Tables 7 and 8. As a result, we confirmed that using tsfresh improved the prediction accuracy, except for the MAE in the test data for mini cucumbers and mini tomatoes

4.7 Discussion

For both mini cucumbers and mini tomatoes, the highest prediction accuracy was achieved when only the past yields (C_PYW_TS, T_PYW_TS) were considered. Conversely, it was found that incorporating environmental data into the features significantly reduced the accuracy. The results of this

Table 6: Evaluation results for each accuracy index

Crop	R			MAE(%)		
	MLR	GAM	MARS	MLR	GAM	MARS
Cucumber	-0.267	0.067	**0.279**	43.1	46.9	**23.8**
Tomato	0.755	**0.790**	0.720	42.0	**19.8**	42.6

Table 5: Model tuning results (hyperparameter)

Crop	Hyperparameter		
	MLR	GAM	MARS

Cucumber	none	Select = TRUE
		Method = GCV.Cp
			degree=1
			nprune=2

Tomato	none	Select = TRUE
		Method = REML
			degree=1
			nprune=12

Table 7: Accuracies with and without using TSFRESH in feature extraction（mini cucumber）

	R		MAE (%)	
	Train	Test	Train	Test
Using tsfresh	**0.796**	**0.279**	**20.7**	23.8
Without tsfresh	0.762	0.156	22.3	**17.3**

Table 8: Accuracies with and without using TSFRESH in feature extraction (mini tomato)

	R		MAE (%)	
	Train	Test	Train	Test
Using tsfresh	**0.851**	**0.790**	25.7	**19.8**
Without tsfresh	0.838	0.770	**19.7**	19.9

experiment showed that past yields had the highest contribution to the prediction of crop yields in plant factories and that environmental data were probably noise. Environmental data made a significant contribution to yield predictions in open-field crops [8]-[13]. On the other hand, this tendency was not observed in fruit and vegetable crops in plant factories, which may be because plant factories are subject to various influences other than the environment. As shown in the visualization of the yield trend of mini tomatoes in Section 4.4.1, the influence of human factors such as production management methods and shipping adjustment methods is particularly high. In addition, we received similar comments from the employees of the plant factory. In the future, we will try to improve the accuracy of the prediction model by adding features related to human factors.

In terms of accuracy, the prediction accuracy of mini tomatoes was R = 0.790, MAE = 19.8, which was higher than that of mini cucumbers. This accuracy generally met the accuracy targets of R = 0.8 or higher and MAE = 20.0% as described in 4.3.2. This result indicates that the proposed model generation process is applicable to other crops. On the other hand, the accuracy for mini cucumbers was R = 0.279 and MAE = 17.3. MAE reached the target value, but R did not reach the target value and was lower than 0.300, which is generally considered to be correlated. This difference in accuracy is largely due to the variation in average yield. The average yield of mini cucumbers was 135 kg, whereas that of mini tomatoes was 24.5 kg, a difference of more than five times. Therefore, the error value is more likely to be larger for mini cucumbers. In the future, it will be necessary to improve the prediction model generation process so that a certain level of accuracy can be achieved even in the case of crops with large fluctuations in average yield and weekly actual yield, such as mini cucumbers.

We asked the employees at the plant factory to confirm the accuracy of the test results, and they said that the yield prediction accuracy for the mini tomatoes was practically viable, but using the approach for mini cucumbers was not yet practicable with the current accuracy. They pointed out that to achieve a practical level, it is more important to know the increasing or decreasing trend of yield compared with the prediction error. In the future, we will improve the accuracy with the goal of identifying yield trends to meet the needs of employees in the plant factory.

5 CONCLUSION

5.1 Summary

The purpose of this study is to develop a model generation process for fruit vegetables in plant factories. The aim is to improve the efficiency and accuracy of plant factory crop yield prediction models and to make them applicable to various fruit and vegetable crops as well as facilities by having a defined process for model generation. The model generation process comprises (1) data selection, (2) data preprocessing, (3) data visualization, (4) feature design, (5) selection of prediction methods, and (6) model optimization. This paper reports on the results of applying the proposed predictive model generation process to mini cucumbers and mini tomatoes, which we interpreted as being representative of other fruit and vegetable crops. Application of the proposed model generation process to mini tomatoes resulted in R = 0.790 and MAE = 19.8, indicating that the proposed method is applicable to other crops. In addition, it was found that the tsfresh package, which automatically generated hundreds of features from the time series data used in the feature extraction process, can be used to enhance the accuracy of the predictions.

In terms of effective features, the prediction model that considered only actual past yield data was the most accurate for both crops. The results of this experiment showed that past yields had the highest contribution to accurate predictions and the environmental data were probably noise. We aim to further improve the accuracy of the models in future works by utilizing other prediction methods such as time series analysis.

5.2 Future Works

(1) Examination of feature quantities

As mentioned in the discussion, we are considering incorporating features related to human factors such as production management and shipment adjustment methods into the prediction model.

(2) Improvement of prediction accuracy by applying time series analysis method

Experiments showed that the contribution of past yields was extremely high, suggesting that time series analysis methods that specialize in predicting time series data, which can capture autocorrelation and cyclical variation, are more suitable. For the time series analysis method, the use of an auto regressive integrated moving average, which applies to nonstationary processes, will be considered.

REFERENCES

[1] Japan Greenhouse Horticulture Association, "Large-scale horticulture and plant factory survey and case studies", https://www.maff.go.jp/j/seisan/ryutu/engei/sisetsu/pdf/daikibo.pdf [Accessed December 10, 2020](*in Japanese*).

[2] Ministry of Agriculture Forestry and Fisheries, "Explanation of the Plant Factory, Japan Center for Social Development Research" , http://www. maff. go. jp/j/heya/sodan/1308/01. html [Accessed December 10, 2020](*in Japanese*).

[3] Ministry of Agriculture Forestry and Fisheries, "Status of installation of horticultural facilities (2019)" , https://www.maff.go.jp/j/seisan/ryutu/engei/sisetsu/haipura/setti_30.html[Accessed December 10, 2020](*in Japanese*).

[4] Ministry of Agriculture, Forestry and Fisheries, "Vegetable Production and Shipment Statistics(2018)", https://www.maff.go.jp/j/tokei/kikaku/book/seisan/attach/pdf/30_yasai- 1.pdf [Accessed December 10, 2020](*in Japanese*).

[5] Ministry of Agriculture Forestry and Fisheries, "Management Statistics by Farming Type(2016)" , https://www.maff.go.jp/j/tokei/kikaku/book/seisan/attach/pdf/30_yasai- 1.pdf [Accessed December 10, 2020](*in Japanese*).

[6] WageningenUR, "Quantitative Information on Dutch grennhouse horticulture (2017)", https://www.wur.nl/en/newsarticle/Quantitative-Information-for-Greenhouse- Horticulture-KWIN-2016-2017.html[Accessed December 10, 2020](*in Japanese*).

[7] T. Hoshi, T. Sasaki, and H Tsutsui, "A daily harvest prediction model of cherry tomatoes by mining from past averaging data and using topological case-based modeling", Computers and Electronics in Agriculture, Vol. 29, No. 1, pp. 149-160 (2000).

[8] A. Bashar and S. Pearson, G. Leontidis, and S.Kollias, "Using Deep Learning to Predict Plant Growth and Yield in Greenhouse Environments"(2019).

[9] W. C. Lin, and B. D. Hill, "Neural network modelling to predict weekly yields of sweet peppers in a commercial greenhouse", Canadian Journal of Plant Science, Vol. 88, pp. 531–536 (2008).

[10] M. Stas, J. Van Orshoven, Q. Dong, S. Heremans, and B. Zhang, "A comparison of machine learning algorithms for regional wheat yield prediction using NDVI time series of SPOT-VGT", IEEE, pp. 258–262 (2016).

[11] Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R. and Mouazen, A., "Wheat yield prediction using machine learning and advanced sensing techniques", Computers and Electronics in Agriculture, Vol. 121, pp. 57 – 65 (2016).

[12] H. Li, Z. Chen, W. Wu, Z. Jiang, B. Liu, and T. Hasi, "Crop model data assimilation with particle filter for yield prediction using leaf area in- dex of different temporal scales", 2015 Fourth International Conference on Agro-Geoinformatics, pp. 401–406 (2015).

[13] G. Okuno, and S. Niiya, "Yield estimation of asparagus using a combination of machine learning and statistical modeling", Japan Social Data Science Society, Vol.2, No.1, pp.14-18(2018) (*in Japanese*).

[14] W. C. Lin, D. Frey, G. D. Nigh, and C. Ying, "Combined Analysis to Characterize Yield Pattern of Greenhouse-grown Red Sweet Peppers", HortScience, Vol. 44, pp. 362–365 (2009).

[15] Y. Todate, M. Oba, and M.Takamori, "Prediction of weekly yields for fruit and vegetable crops in plant factories", Information Systems and Social Environment (IS) 149th Research and Presentation Meeting , Vol.151,No.12,pp.1-8 (2020).

[16] A. González-Sanchez, J. Frausto-Solis, and W. Ojeda, "Predictive ability of machine learning methods for massive crop yield prediction", SPANISH JOURNAL OF AGRICULTURAL RESEAR(2014).

[17] M. Christ, N. Braun, J. Neuffer, and A.W. Kempa-Liehr, "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh -- A Python package)". Neurocomputing, Vol. 307, pp. 72-77(2018).

[18] M. Christ, A.W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications", Asian Machine Learning Conference (ACML) 2016, Workshop on Learning on Big Data (WLBD), Hamilton (New Zealand)(2016).

(Received October 30, 2020)
(Accepted February 15, 2021)

Yuki Todate received his B.E. and M.E. degrees in infor-mation science from Future Univer-sity Hakodate, Japan in 2019 and 2021. Hers research interests include smart agriculture, plant factory, and statistical modeling. She currently works in Nomura Research Institute, Ltd..

Michiko Oba received her B.S. in physics from Japan Women's University in 1982, and Ph.D. in engineering from Osaka University, Japan, in 2001. She worked in the Systems Development Laboratory and the Software Division of Hitachi Ltd. Presently, she is a professor in the Department of Media Architecture, Future University Hakodate, Japan.Prof. Oba is a member of IEEE Computer Society, the Information Processing Society of Japan (IPSJ), the Institute of Electrical Engineers of Japan (IEEJ). She became a Council Member of Science Council of Japan in 2020. She received IPSJ fellow in 2020.

Mitsuru Takamori is president of IT consulting company called Agri. Connections Corp. from 2013 and he is doing research to utilize ICT and IOT in the field of agriculture.
1988-2013 Project manager in IBM Japan 2012-2016 Project Professor in Future University Hakodate.