Predicting the Yields of Field Vegetable Using the Multiple Functional Regression Model

Wanhyun Cho
Department of Statistics
Chonnam National University
Gwangju, South Korea
whcho@chonnam.ac.kr

Myung-Hwan Na
Department of Statistics
Chonnam National University
Gwangju, South Korea
mhn@chonnam.ac.kr

Yuha Park
Department of Statistics
Chonnam National University
Gwangju, South Korea
yuhapark0@gmail.com

***Abstract*—**In this paper, we propose a new statistical method that can analyze the yield data of the field vegetable using a functional regression model with multiple functional covariates and scalar response. From the experimental results, we could see the following results. First, through the descriptive statistical analysis, we can see that CO_2 level is positively correlated with yield, while humidity has a strong negative correlation with the yield of vegetable, but temperature has a weak negative correlation with yield. Second, through functional regression analysis of three environmental variables, we confirmed that CO_2 and humidity have a very significant effect on the yield of vegetable at a significance level of 5% or less. Third, we can see that the vegetable yields is maximum when the CO_2 level is approximately 870 to 900 (ppm), and the vegetable yield is the maximum when the humidity is 93 to 95 (%), but when the temperature is between 16 and 18 (℃), there is no significant difference in vegetable yields.

Keywords—Functional data analysis (FDA); Multiple functional regression model; Yields of vegetable, Enoviromental factors; Opimal conditions; R-packages;

I. INTRODUCTION

Mushrooms are one of the most popular vegetables for Koreans, and this is the most healthful food and is being used in various dishes. In addition, mushrooms are produced in many parts of Korea because they are regarded as high value - added items in institutional farms. Therefore, farmers are interested in farming methods that can improve the yield of mushroom. With these problems, we are trying to study various environmental variables that affect the yield of mushrooms.

Before we begin research to address these problems, let us briefly review past research results. Manikandan and Vethamoni [1] recently published a review paper on the results of a study on cultivation models in which various environmental variables in vegetable crops can be used to determine how to cultivate vegetable crops. In this paper they talked about modeling techniques that can be useful to define research priorities and understanding the basic interactions of the soil-plant-atmosphere system. Sher et al. [2] investigated to examine the suitability of Oyster mushroom cultivation and to compare the growth and yield of Oyster mushroom in the two different areas with different ecological conditions. Villers [3] has published his master's thesis on the predicting tomato crop yield from weather data using statistical learning techniques. He described how to use various statistical techniques, such as multiple regression analysis, lasso, regression tree, to see how the weather phenomenon affects the yield of tomato using the planting and harvest data observed for seven years in his master thesis. Keita and colleagues [4] proposed a consolidated methodology for estimating vegetable crops area, yield and production addressing the main methodological issues and taking into account lessons learnt from past experiences, from country practices and analysis of data from case studies in pilot countries in the context of African countries.

In general, yields of field vegetables such as mushrooms are highly affected by various environmental factors. Also, measurements of these factors are given differently from day to day or from week to week. Therefore, we need a way that can analyze the variable functional data that changes with time. Here, we will consider the functional data analysis that is to be the best method for analyzing the data of the time-varying functions.

Two important contributions of our study are as follows. The first contribution is to discriminate what environmental factors affect the yield of vegetable using the multiple functional regression analysis. The second contribution is to find out what levels of these environmental variables can maximize the yield of vegetable grown in fields.

II. DATESETS AND METHOD

A. Datasets

Here, we consider the problem of cultivation methods that can improve the yields of oyster mushroom, which is popular among Koreans and has a high perennial value. Generally, the mushroom cultivation process takes about 15 days from seeding to the harvest in 5 stages. Fig. 1 shows the process of cultivating mushrooms in stages.

Fig. 1. Mushroom cultivation process

At this time, environmental variables affecting growth of mushrooms include temperature, CO_2 level, and humidity. Thus, the data set used in this study includes the yield of mushroom and measurements of three environmental variables. This dataset is collected from six facility farms (smart farming) from the first week of August, 2016 to the 4st week of May, 2018 in South Korea. We measured the values of three environmental variables every hour from day 1 to day 15 for 34 mushrooms in 12 rooms and measured yields on the last day. TABLE 1 shows partially the mushroom yields of 34 chambers and some of the values of the three environmental variables measured during the 360-hour growth period for them.

TABLE 1. Mushroom yields and the values of the three environmental variables.

Factor	Measurement Unit	Measurement Interval
Yield	g/unit	34 chambers
CO_2	ppm	360-hour
Temperature	℃	360-hour
Humidity	%	360-hour

B. Multiple Functional Regression Model

A statistical model to determine whether p environmental factors affect the yield of n mushrooms over time is given by the scalar response regression model with p functional covariates [5-7]. This model is formulated as follows:

$$y_i = \alpha + \sum_{k=1}^{p} \int_{I_s} x_{ik}(s)\beta_k(s)ds + \epsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

where α is the mean function, p is the number of functional covariates, n is the number of observations, $\beta_k(s)$ is the regression function for the k–th covariate and ϵ_i is a random error function.

Here, to estimate the functional parameters $\beta_k(s)$ of this model, we can first consider the centered covariates and response variables to eliminate the functional interceptα. They are centered by

$$y_i^c = y_i - \bar{y}, \qquad x_{ik}^c(s) = x_{ik}(s) - \bar{x}_k(s), k = 1, \cdots, p. \tag{2}$$

And we can also use a measure of least minimum integrated squared residual (LMISE), which is defined as follows to derive an estimate of the regression coefficient.

$$\text{LMISE} = ||\mathbf{y} - \hat{\mathbf{y}}|| = \sum_{i=1}^{n} \left(y_i - \left(\alpha + \sum_{k=1}^{p} \int_{I_s} x_{ik}(s)\beta_k(s)ds \right) \right). \tag{3}$$

In order to derive an estimate of the functional regression parameters that can minimize the measure LMISE defined above, we can use functional principal component analysis. We first expand the $x_{ik}^c{}'s$ in a basis $\phi_{jk}{}'s$ and the y_i^c in a basis ψ_l, to give

$$x_{ik}^c = \sum_{j=1}^{J} c_{ijk}\phi_{jk} = \boldsymbol{c}_{ik}^T \boldsymbol{\phi}_k, k = 1, \cdots, p, \tag{4}$$

and

$$y_i^c = \sum_{l=1}^{L} d_{il}\psi_l = \boldsymbol{d}_i^T \boldsymbol{\psi}\,.$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are the J– and L– vectors of the respective basis functions. We denote the matrices of coefficients by $\mathbf{C}_k$ and $\mathbf{D}$, so that we can write these expressions in the function form

$$\mathbf{X}_k^c = \mathbf{C}_k\boldsymbol{\phi}, \quad \boldsymbol{y}^c = \mathbf{D}\boldsymbol{\psi}\,. \tag{5}$$

We consider the expression of $\boldsymbol{\beta}_k$ as a double expansion

$$\beta_k(s) = \sum_{j=1}^{J}\sum_{l=1}^{L} b_{jlk}\phi_{jk}(s)\psi_l = \boldsymbol{\phi}_k^T(s)\mathbf{B}_k\boldsymbol{\psi}, \tag{6}$$

where $\mathbf{B}_k$ is a $(J \times L)$ matrix of coefficients b_{jlk}, or, more compactly, as $\boldsymbol{\beta}_k = \boldsymbol{\phi}_k^T\mathbf{B}_k\boldsymbol{\psi}$. Define $\mathbf{J}_{\boldsymbol{\phi}_k}$ and $\mathbf{J}_{\boldsymbol{\psi}}$ to be the matrices of inner products between the elements of the $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}$ bases, respectively. Thus,

$$\mathbf{J}_{\boldsymbol{\phi}_k} = \int_{I_s} \boldsymbol{\phi}_k(s)\boldsymbol{\phi}_k(s)^T ds, \quad \mathbf{J}_{\boldsymbol{\psi}} = \boldsymbol{\psi}\boldsymbol{\psi}^T \tag{7}$$

Substitute the basis expansions of x_{ik} and $\boldsymbol{\beta}_k$ into (1) to give

$$\hat{\boldsymbol{y}}^c(t) = \sum_{k=1}^{p} \int_{I_s} \mathbf{C}_k\boldsymbol{\phi}_k(s)\boldsymbol{\phi}_k^T(s)\mathbf{B}_k\boldsymbol{\psi}ds = \sum_{k=1}^{p} \mathbf{C}_k\,\mathbf{J}_{\boldsymbol{\phi}_k}\mathbf{B}_k\boldsymbol{\psi}. \tag{8}$$

If we let $\widehat{\mathbf{D}}$ be the matrix of coefficients of the basis expansion of the vector of predictors $\hat{\boldsymbol{y}}^c$ (corresponding to the matrix $\mathbf{D}$ for the vector $\boldsymbol{y}^c$), we obtain the matrix form of the model

$$\widehat{\mathbf{D}} = \sum_{k=1}^{p} \mathbf{C}_k\,\mathbf{J}_{\boldsymbol{\phi}_k}\mathbf{B}_k \tag{9}$$

Now we can get an expression for the integrated squared residual:

$$\| \hat{\boldsymbol{y}}_i - \boldsymbol{y}_i \|^2 = \| \hat{\boldsymbol{y}}_i^c - \boldsymbol{y}_i^c \|^2 = \left(\left(\widehat{\mathbf{D}} - \mathbf{D}\right)\mathbf{J}_{\boldsymbol{\psi}}\left(\widehat{\mathbf{D}} - \mathbf{D}\right)^T \right)_{ii} \tag{10}$$

and, finally,

$$\text{LMISE}(\mathbf{B}_k) = \text{trace}\left(\left(\widehat{\mathbf{D}} - \mathbf{D}\right)\mathbf{J}_{\boldsymbol{\psi}}\left(\widehat{\mathbf{D}} - \mathbf{D}\right)^T \right) \tag{11}$$

is given by a sum of quadratic forms in the unknown coefficient matrices $\mathbf{B}_k$.

Furthermore, let us consider the minimization of the quantity LMISE($\mathbf{B}_k$) as given in (11). In the case where $\mathbf{J}_{\boldsymbol{\phi}_k}$ and $\mathbf{J}_{\boldsymbol{\psi}}$ are identity matrices, the matrix $\mathbf{B}_k$ will minimize (11) if and only if

$$\mathbf{C}_k^T\mathbf{C}_k\mathbf{B}_k = \mathbf{C}_k^T\mathbf{D}$$

so that

$$\mathbf{B}_k = (\mathbf{C}_k^T\mathbf{C}_k)^{-1}\mathbf{C}_k^T\mathbf{D}. \tag{12}$$

The matrix $\mathbf{B}_k$ is easily found by using the SVD (singular value decomposition) of $\mathbf{C}_k$. Write $\mathbf{C}_k = \mathbf{U}\boldsymbol{\Delta}_{C_k}\mathbf{V}^T$ where $\boldsymbol{\Delta}_{C_k}$ is a diagonal matrix with strictly positive diagonal elements and $\mathbf{U}$ and $\mathbf{V}$ have orthogonal columns. Then,

$$\mathbf{C}_k{}^T\mathbf{C}_k = \mathbf{V}\boldsymbol{\Delta}_{C_k}^2\mathbf{V}^T, \tag{13}$$

and hence the Moore-Penrose g-inverse of $\mathbf{C}_k{}^T\mathbf{C}_k$ is $\mathbf{V}\boldsymbol{\Delta}_{C_k}^{-2}\mathbf{V}^T$. Substituting it into (12) gives

$$\mathbf{B}_k = \mathbf{V}\boldsymbol{\Delta}_{C_k}^{-1}\mathbf{U}^T\mathbf{D}. \quad (14)$$

Therefore, if we substitute $\mathbf{B}_k$ into $\beta_k(s)$ in Eq. (6), we get the following regression estimates.

$$\hat{\beta}_k(s) = \boldsymbol{\phi}_k^T(s)\hat{\mathbf{B}}_k\boldsymbol{\psi}. \quad (15)$$

Finally, we get the following predictions for response values.

$$\hat{y} = \bar{y} + \sum_{k=1}^{p}\int_{I_s} x_{ik}(s)\hat{\beta}_k(s)ds\ . \quad (16)$$

III. Experimental Results

A. Graphical Anaysis

First, we plotted scatter plots to verify graphically the relationship between yield and three environmental factors. We have plotted two-dimensional scatter plots between mushroom yields and mean values of the three environmental factors during the growing season. The Fig. 2 show the relationship between mushroom yields and three environmental factors. From the given graphs, we can see that the mushroom yields is maximum when the CO_2 level is approximately 870 to 900 (ppm), and the mushroom yield is the maximum when the humidity is 93 to 95 (%), but when the temperature is between 16 and 18 (℃), there is no significant difference in mushroom production.

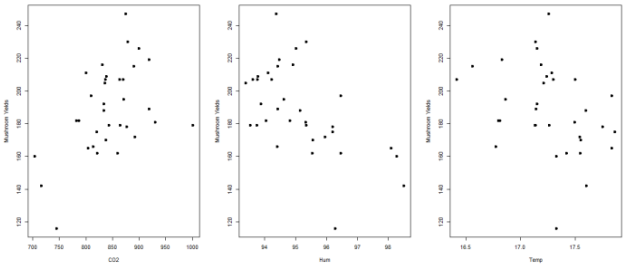


Fig. 2. Relationship between mushroom yields and three environmental factors.

We conducted a statistical test through correlational analysis of the relevance that we had so far visually confirmed. From TABLE 2 we can see that CO_2 level is positively correlated with yield, while humidity and temperature are negatively correlated with yield.

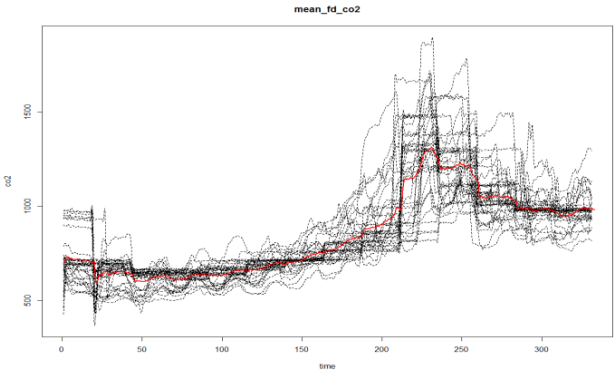
TABLE 2. Correlation coefficients between mushroom yields and three environmental factors.

ρ	yield	CO_2	Hum	Temp
yield	1.00	0.47	-0.57	-0.34
CO_2	0.47	1.00	-0.43	-0.19
Hum	-0.57	-0.43	1.00	0.59
Temp	-0.34	-0.19	0.59	1.00

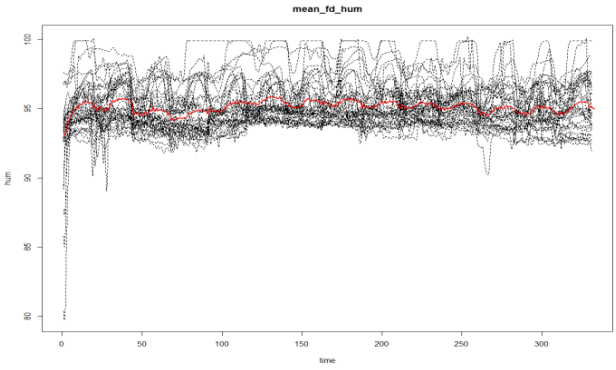
B. Fuctional Data Analysis for mushroom data

We performed multiple covariate functional regression analysis to see how the three environmental factors affect mushroom yield. We also performed a functional regression analysis on mushroom data using the "fda" package in the R-statistical software.

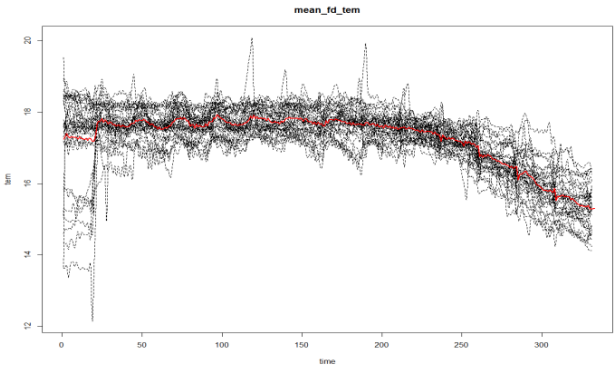
We first averaged the values measured in the 34 chambers for the three environmental variables to produce data suitable for functional regression analysis. Fig. 3 shows the line graphs for the values as well as their mean values observed at 34 chambers of the three environmental factors, CO_2, humidity and temperature.



(a) Line graph for CO_2 levels



(b) Line graph for humidity values



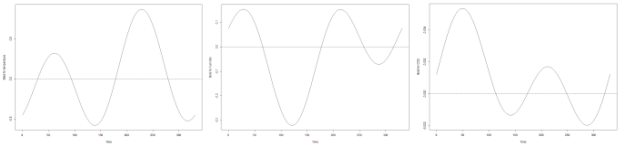
(c) Line graph for temperature values

Fig. 3. The line graph for values and their means of three environmental factors

Second, we set up a functional data objects for the 34 chambers of the three environmental factors, CO_2, humidity and temperature, called timeco2fd, timehumfd and timetemfd. To keep things simple and the computation rapid, we will use 61 basis functions without a roughness penalty. To calculate the estimator of the regression coefficient $\boldsymbol{\beta}$ of the functional linear model on the created function data, we will use the basis expansion of the regression coefficient $\boldsymbol{\beta}$. This can be done

using in the R function "**fRegress**", which requires at least three arguments: yfdPar, xfdlist, betalist. Here, yfdPar contains the value of the response variable, xfdlist contains the functional covariate function, and finally betalist represents the list object created in R with the same length as xfdlist.

Fig. 4 below shows the estimator functions of the regression coefficient **β** to predict the mushroom yield using the three environmental factors.



(a) Temperature (b) Humidity (c) CO_2

Fig. 4. Estimated **β** for three environmental factors

Here, we can see that humidity and CO_2 affect the mushroom yield in the early stage, while the temperature has a great influence on the mushroom yield in the post-intermediate stage. Concretely, first, the level of CO_2 has a positive influence on the mushroom yield from the first to the middle of the growth stage, and it has almost no influence after the middle of the growth stage. Second, to improve the mushroom yield, we can see that the humidity should be raised sufficiently during the second growth stage of mushroom. Third, the temperature is generally kept constant during the growth period, and the temperature is lowered in the latter half of the period, but in order to improve mushroom yield during this period, the temperature must be lowered slowly.

Third, we conducted goodness-of-fit tests for three environmental factors to statistically determine how the three environmental factors affect mushroom yield. We can see the following results from the output of the R program. First, the squared multiple correlation of the temperature is not explained as much as 0.31, and the test result showed that F statistic was 2.57, meaning that the effect of the temperature is not significant at the significance level 5%. This result is consistent with the results of correlation analysis. Second, the squared multiple correlation of the humidity is 0.51, and the F statistic is 5.73, which the effect of the humidity is significant at the significance level of 5%. This result is also consistent with the results of correlation analysis. Third, the squared multiple correlation of CO_2 is 0.53, which showed the highest explanatory power. As a result, the F statistic is 6.43, which the effect of the CO_2 is significant at the significance level of 1%. This result is also consistent with the results of correlation analysis.

IV. Conclusion and Discussion

In this study, we have considered the functional regression analysis that can be the best method for analyzing the mushroom data of the time-varying functions. To do this, we performed two tasks. The first object is to identify the environmental factors that affect mushroom yield and the second one is to find the optimal condition of environmental factors that can maximize mushroom yield.

From the experimental results, we could see the following results. First, through the descriptive statistical analysis, we can see that CO_2 level is positively correlated with yield, while humidity has a strong negative correlation with the yield of mushroom, but temperature has a weak negative correlation with yield. Second, through functional regression analysis of three environmental variables, we confirmed that CO_2 and humidity have a very significant effect on the yield of mushroom at a significance level of 5% or less. Third, we can see that the mushroom yields is maximum when the CO_2 level is approximately 870 to 900 (ppm), and the mushroom yield is the maximum when the humidity is 93 to 95 (%), but when the temperature is between 16 and 18 (℃), there is no significant difference in mushroom yields.

Future research plans will use the same functional regression analysis to study the environmental variables that affect the yield of other agricultural products such as tomatoes and strawberries.

Acknowledgment

This work was partially supported by the Research Program of Rural Development Administration (Project No. PJ0138672019), Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET), funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (319002-01), and the Korea National Research Foundation (Project No. 2017R1D1A1B03028808) of Korea Grant funded by the Korean Government.

References

[1] K Manikandan and P Irene Vethamoni, "A review: Crop modeling in vegetable crops," Journal of Pharmacognosy and Phytochemistry, vol. 6(4), pp. 1006-1009, 2017.

[2] H. Sher, M. Al-Yemeni, Ali H. A. Bahkali, and H. Sher, "Effect of environmental factors on the yield of selected mushroom species growing in two different agro ecological zones of Pakistan", Saudi Journal of Biological Sciences, vol. 17, pp. 321-326, 2010.

[3] Margaret de Villers, "Predicting tomato crop yield from weather data using statistical learning techniques", Thesis for the degree of Master, Department of Statistics and Actuarial Sciences, University of Stellenbosch, South Africa.

[4] N. Keita, E. Ouedraogo, and U. E. Nyamsi, "Measuring Area, Yield and Production of Vegetable Crops," Proceedings of ICAS VII: Seventh International Conference on Agricultural Statistics, Rome 24-26 October, 2016.

[5] J.-L. Wang, J.-M. Chiou, and H.-G. Muller, "Review of functional data analysis," Annual Review Statistics, vol. 7, 1-41, 2015.

[6] J. S. Morris, "Functional Regression," Annual Review of Statistics and its Applications, 2, 321–59, 2015.

[7] S. Grieven and F. Scheipl, "A general framework for functional regression modeling," Statistical Modeling, vol. 17, 1-35, 2017