ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 52, No.4: 767-784, December 2024
https://dx.doi.org/10.7469/JKSQM.2024.52.4.767

시설 오이 재배를 위한 핵심 요인 분석과 수확량 예측 모델 개발

강소라* · 이혜림** · 나명환*†

* 전남대학교 수학/통계학과
** 농촌진흥청

Analysis of Key Factors and Development of Yield Prediction Model for Greenhouse Cucumber Cultivation

Kang, So Ra* · Lee, Hyerim** · Na, Myung Hwan*†

* Department of Mathematics and Statistics, Chonnam National University
** Rural Development Administration

ABSTRACT

Purpose: The purpose of this study is to examine the relationship between yield and growth factors and environmental factors, and to propose a model suitable for predicting cucumber yield by reflecting these characteristics and to derive important factors.

Methods: Using smart farm cucumber data, correlation analysis and dynamic time warping (DTW) are used to analyze the relationship between yield and factors, and three regression models, MLR (multiple linear regression), PLSR (partial least square regression), and SVR (support vector regression), are used for the yield prediction model.

Results: The results of this study are as follows; correlation analysis showed that stem thickness and leaf number were highly correlated with yield, and dynamic time warping showed that the increase in the number of nodes and leaf length showed a similar pattern to yield. The relationship between yield and factors can be interpreted differently from the independent influence of a single variable and the perspective of multivariate interaction. In general, environmental management of temperature and humidity during the day plays an important role in improving yield. The SVR model is the most suitable model for predicting cucumber yield because it is advantageous in nonlinear and highly variable data compared to the MLR and PLSR models.

Conclusion: This study is expected to expand the applicability of smart farm technology and contribute to optimizing crop growth and improving productivity through data-based predictive modeling.

● Received 15 November 2024, 1st revised 26 November 2024, accepted 3 Dember 2024
† Corresponding Author(nmh@chonnam.ac.kr)

This research was supported by the research fund from the Rural Development Administration (RDA) (RS-2023-00215013).
* 본 연구는 농촌진흥청 연구비(RS-2023-00215013)의 지원을 받아 수행되었습니다.

Key Words: Greenhouse Cucumber Cultivation, Optimal Cultivation Enviromental Factor, Yield Prediction System, Support Vector Regression

1. 서 론

1.1 연구 배경

기후 변화와 자연재해가 빈번해지면서 농업 생산은 점점 더 많은 불확실성과 위험에 직면하고 있다. 이러한 문제는 개방된 환경에서 재배되는 농작물의 생육에 부정적인 영향을 미치며, 특히 기후에 민감한 채소류의 생산성과 품질을 크게 위협하고 있다. 이에 따라 외부 환경의 변동성을 줄이고 작물 생육을 효율적으로 관리할 수 있는 시설 재배의 중요성이 높아지고 있다. 시설 재배는 온실을 통해 작물의 생육 환경을 인위적으로 조절함으로써, 기후 변화에 대한 저항성을 강화하고 농작물의 안정적인 생산성을 확보할 수 있다.

이 중에서도 오이(cucumber)는 환경 조건에 민감한 작물로, 온도, 습도, 일조량, 이산화탄소 농도 등의 변화에 따른 생육 반응이 뚜렷하게 나타나 시설 재배의 필요성과 가능성을 연구하기에 적합한 대상이다. 각 환경 요인은 오이의 생리적 반응에 영향을 미쳐 생장 속도, 품질, 수확량에 직결되기 때문에, 이러한 요인들을 최적화하는 것은 오이 생산의 효율성을 높이는 핵심 과제이다. 따라서 오이의 생육 및 생산량을 극대화하기 위해서는 이러한 환경 요인들을 효과적으로 관리할 수 있는 최적화된 시스템이 필요하다.

최근 스마트팜 기술의 발전으로 다양한 센서를 통해 온실 내 환경 데이터를 실시간으로 모니터링하고 제어하는 것이 가능해졌다. 환경 데이터는 오이의 생육 특성에 중요한 영향을 미치는 변수들이므로, 빅데이터와 다양한 알고리즘을 사용하여 분석하고 예측 모델을 개발함으로써 작물의 수확량을 예측하는 새로운 가능성이 열리고 있다. 특히, 회귀분석과 같은 통계적인 방법은 환경 요인과 수확량 간의 관계를 모델링하는 데 유용하며, 이를 통해 최적의 생육 조건을 도출할 수 있다.

1.2 선행 연구

시설 재배 환경에서 오이의 생육과 수확량을 최적화하는 연구는 지속적으로 이루어져 왔으며, 스마트팜 기술과 데이터 기반의 예측 모델링이 농업 분야에 적용되면서 이에 대한 관심이 더욱 높아지고 있다. 최근 연구들은 센서와 빅데이터, 머신러닝 및 딥러닝 기술을 통해 온실 내부 환경을 실시간으로 감지하고, 이를 분석하여 작물 생육에 최적화된 환경을 제어하려는 시도가 늘어나고 있다.

첫째, 오이 생육에 영향을 미치는 환경 요인을 분석한 연구들이 있으며, 생산량 향상에 있어서 생육과 환경 관리의 중요성을 강조하였다. Lee et al.(2012)들은 기상 요인이 농산물 생산에 미치는 영향을 분석하여 습도, 일사량, 풍속이 주요 기상 요인임을 확인하였고, 이러한 기상 요인에 대한 정보가 최적화된 재배 계획을 통해 농산물 생산량 극대화에 기여할 수 있음을 제시하였다. Lee et al.(2018)들은 고온과 토양수분 조건이 오이 생육에 미치는 영향을 분석하였으며, 온도가 높을수록 생육이 저하되고 특히 고온과 건조 조건에서 수량 감소가 두드러짐을 확인하였다. Jeon et al.(2019)들은 오이 절간장을 10~15cm로 유지하는 것이 수량 증가에 유리하며, 특히 시설 내 습도가 절간장에 중요한 영향을 미치는 환경 요소임을 확인하였다.

둘째, 작물의 생장과 수확량을 예측하기 위해 회귀분석과 머신러닝 및 딥러닝 기법을 활용한 연구가 진행되고 있

다. Na et al.(2017)들은 스마트팜 토마토 농가에서 수집된 데이터를 바탕으로 시차 개념을 적용해 수확량과 환경 변수 간의 연관성을 분석하였으며, 다중 회귀분석을 통해 선택된 환경 지연변수를 이용하여 토마토의 수확량을 조절할 수 있음을 제시하였다. Lee et al.(2019)들은 다양한 환경 요인들(내부온도, 내부습도, 내부대기압, 외부온도, 외부습도, 외부대기압, 강수량, 토양온도, 토양습도, 토양수분함량, 토양전도도, 이슬점, 토양수분장력) 중 온실에서 재배된 오이의 성장과 생산성에 영향을 미치는 주요 환경 요인을 식별하고, 지역과 품종을 고려한 회귀분석을 통해 오이의 생산성을 예측하였다. Hong et al.(2020)들은 토마토의 생육, 환경, 경영 정보를 이용하여 1주 후, 2주 후의 생산량과 생장량을 예측하고, 다중선형회귀, 랜덤포레스트, ConvLSTM 알고리즘을 비교 분석하였다. 그 결과, ConvLSTM 모델이 가장 높은 성능을 보였으나, 학습 모델 구축을 위한 데이터의 부족 문제를 지적하였다. Kim and Kim(2021)은 토마토의 생육, 환경 등 여러 정보를 이용하여 상관분석과 Boruta 알고리즘으로 주요 요인을 선정한 후, 세 가지 머신러닝 알고리즘(릿지 회귀, 랜덤 포레스트, XGBoost)으로 총 출하량을 예측하였다. 또한, 병해충 정보 등 추가 요인이 예측 모델의 정확성을 높일 수 있으며, 데이터 품질 관리를 위한 표준화된 가이드가 필요하다고 언급하였다. Jang and Park(2023)은 파프리카와 오이의 환경 및 생육 정보를 활용하여 생산량 예측을 위한 MLP, LSTM, GRU, Transformer 등의 딥러닝 모델을 적용하였으며, 그 중 GRU와 Attention 기반 GRU 모델이 좋은 성능을 보였으나 데이터셋의 크기 부족 문제를 해결할 필요성을 언급하였다.

1.3 연구 목적

본 연구는 온실 환경에서 오이 생장에 영향을 미치는 주요 생육 요인과 환경 요인을 식별하고, 이를 활용하여 수확량을 예측하는 최적의 모델을 제시하는 것을 목적으로 한다. Na et al.(2017)의 연구에서는 특정 시점의 환경 요인이 수확량에 미치는 영향을 분석하기 위해 지연 변수(lagged variable) 개념을 사용하였다. 이러한 접근은 단일 시점에서의 환경 요인과 수확량 간의 관계를 평가하는 데 효과적이지만, 환경 요인의 지속적 변화나 누적된 영향을 충분히 반영하지 못한다는 한계가 있다. 또한, 최근 연구에서 주로 사용된 머신러닝 및 딥러닝 기법은 대규모 데이터셋을 필요로 하며, 데이터가 제한적인 상황에서는 예측 성능이 저하될 가능성이 있다.

본 연구에서는 이러한 한계를 보완하기 위해 환경 요인이 수확량에 미치는 시간적 영향을 여러 주에 거쳐 반영한 가중 시차(weighted lag) 개념을 도입하였다. 이 개념은 특정 기간 동안의 환경 변화가 수확량에 미치는 효과를 가중치로 계산하여, 시간적 지연 효과와 누적된 영향을 동시에 모델링하는 방식이다. 가중 시차 개념은 작물의 생육 반응이 환경 변화에 즉각적으로 나타나지 않고, 대사 활동, 세포 성장, 광합성과 같은 생리적 과정을 거쳐 일정 시간이 지난 뒤에 발현되는 특성을 고려하여 설계되었다. 또한, 데이터셋이 제한적인 상황에서도 안정적이며 신뢰성 있는 예측 성능을 제공하기 위해 세 가지 일반 회귀 모형을 활용하였다. 이를 통해 수확량에 영향을 미치는 주요 요인을 파악하고, 각 재배 환경 요인과 수확량 간의 관계를 정량적으로 분석하여 실제 스마트팜 운영에 적용 가능한 예측 시스템을 구축하고자 한다.

연구는 다음의 절차를 거쳐 수행되었다. 첫째, 온실 내부의 다양한 환경 요인들의 수준을 센서를 통해 측정하고, 여러 생육 특성을 실측하여 자료를 수집하였다. 둘째, 수집된 환경 자료를 가중 시차 개념을 적용하여 시간적 영향을 반영한 형태로 변환하였다. 셋째, 변환된 자료를 기반으로 상관분석과 동적 시간 워핑(dynamic time warping, DTW)을 활용하여 수확량을 극대화할 수 있는 재배환경 요인들의 최적 수준을 도출하였다. 넷째, 다중 선형 회귀(multiple linear regression, MLR), 부분 최소 제곱 회귀(partial least square regression, PLSR), 서포트 벡터 회귀(support vector regression, SVR) 등의 일반 회귀 모형을 사용하여 주요 생육·환경 요인들을 추출하고, 수확량을 정확히 예측할 수 있는 예측 모델을 구축하였다. 다섯째, RMSE(root mean squared error)와 MAE(mean absolute

error)의 오차 기반 지표를 통해 구축한 모델을 평가하였다(Lee et al., 2024).

2. 자료 수집 및 연구방법

2.1 연구 자료 및 전처리

본 연구 자료는 경기도 지역에서 스마트팜 오이를 토경 재배하고 있는 네 농가로부터 수집된 생육 정보, 환경 정보, 생산성 정보로 구성된다. 조사 대상 농가는 품종과 재배 시기가 각기 다르며, 2021년 1작기 동안 단동 및 연동 비닐온실(1,000평에서 3,085평 사이)에서 재배를 진행하였고, 평균 온실 규모는 약 1,684평이었다. 재배 기간은 최소 7주에서 최대 17주까지로, 첫 수확부터 마지막 수확까지의 단위면적당 총 오이 과중은 최소 $1{,}634g/3.3m^2$에서 최대 $9{,}502g/3.3m^2$로 기록되었다. 생육 정보는 조사원에 의해 주 1회 주기로 기록되었으며, 생산량 정보는 대략 일주일 간격으로 여러 개체에 대해 다양한 마디에서 수확이 이루어졌다. Figure 1은 주차별 단위면적당 오이의 수확량 증가량과 누적 수확량의 분포를 보여주는 박스플롯으로, 주차에 따른 변동성과 증가 패턴을 시각적으로 확인할 수 있다. 주간 수확량은 전반적으로 변동폭이 크고 주차별로 일정하지 않은 증가 양상을 보이며, 누적 수확량은 초기에는 완만히 증가하다가 특정 시점 이후 급격히 비선형적으로 증가하여 S자 형태의 패턴을 나타낸다.

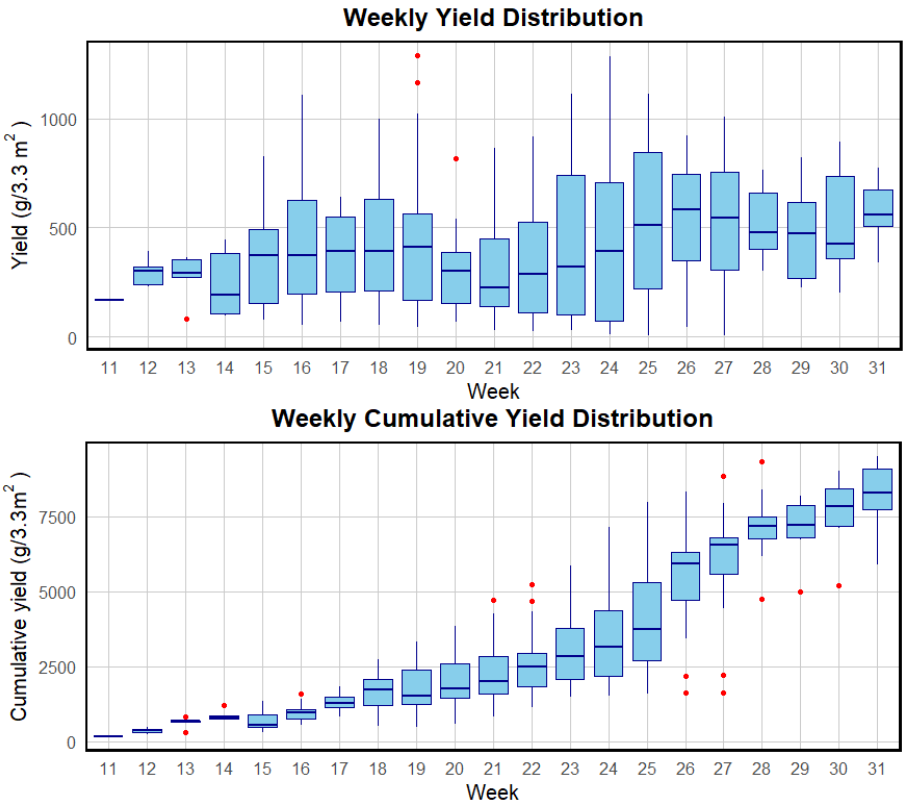

Figure 1. Weekly Yield Increase and Cumulative Yield Distribution of Cucumbers

환경 정보는 ICT 기반의 환경 제어 시스템을 통해 매 10분마다 내부온도, 상대습도, 토양온도, CO_2 농도가 측정되어 수집되었다. 수집된 환경 정보는 생육 조사 일자에 맞추어 주 단위로 평균, 최소, 최대, 주간, 야간, 일출전후 1시간 평균, 일출부터 정오까지 평균, 일교차, GDD(growing degree days) 등 각 환경 요인별로 다양한 값을 산출하여 수확량과의 상관성을 분석하고 생육 상태 변화를 파악할 수 있도록 가공되었다. 환경 변수는 평균, 최고, 최저, 주간, 야간, 일출전후 1시간 평균, 일교차, GDD 등 내부온도 관련 8가지 요소, 평균, 최고, 최저, 주간, 야간 등 상대습도 관련 5가지 요소, 평균, 최고, 최저 등 토양온도 관련 3가지 요소, 평균, 최고, 최저, 일출부터 정오까지 평균 등 CO_2 농도 관련 4가지 요소의 다양한 형태로 활용되었다. 생육 정보는 초장 증가량, 마디수 증가량, 줄기굵기, 엽장, 엽폭, 엽수, 평균 절간장 증가량, 암꽃수, 착과수 등이 조사되었다. 생산량 정보, 생육 정보, 환경 정보에 대한 자세한 설명은 Table 1에 제시되어 있다.

Table 1. Description of Variables in the Analysis Data

Classification		Explanation	Variable name	Unit
Response variable		Weekly cumulative yield per unit area	Yield	g/3.3m^2
Experiment variable	Growth factors	Increase in plant height	Plant	cm
		Increase in stem node	Node	cm
		Stem diameter	Stem	cm
		Leaf length	Leaf_len	cm
		Leaf width	Leaf_w	cm
		Leaf number	Leaf_num	count
		Increase in average internode length	Internode	cm
		Number of female flowers	P_f	count
		Fruit number	Fruit	count
	Environmental factors	Internal temperature	Temp	℃
		Soil temperature	Stemp	℃
		Internal relative humidity	Hum	%
		CO_2 concentration	CO2	ppm

오이 수확량 예측 모델에서는 주별 단위면적당 누적 수확과중(weekly cumulative yield per unit area)을 반응변수로 설정하고, 매주 수확된 누적 과중을 단위면적 기준으로 환산하여 사용하였다. 설명변수로는 9가지 생육 요인과 파생된 변수를 포함한 20가지 환경 요인을 활용하였다. 각 농가에서 6개체씩 조사하였으며, 재배 기간이 충분하지 않아 수확 패턴과 수확량의 변동성이 클 수 있는 재배 기간 4주 이하의 개체는 제외하여 총 288개의 데이터를 사용하였다. 전체 23개 개체 중 누적 수확량이 높은 경우와 낮은 경우를 고르게 포함하여 19개 개체를 훈련용으로, 나머지 4개 개체를 테스트용으로 선정하였다. 훈련 자료는 5-폴드 교차 검증(5-fold cross validation)을 10회 반복하여 모델을 학습하고 검증하는 데 사용하였으며, 학습된 모델의 일반화 성능은 독립적인 테스트 데이터를 통해 최종적으로 평가하였다.

2.2 환경 요인의 반응 시차를 고려한 자료 변환

본 연구에서는 환경이 작물의 생장 및 수확량에 미치는 시간적 영향을 반영하기 위해 가중 시차(weighted lag) 개념을 도입하였다. 오이는 정식 후 약 30일이 지나면 첫 수확이 가능하며, 개화 후에는 약 7~10일 후에 수확할 수 있다. 이러한 생육 특성을 고려하여 5주간($k=5$)의 환경 조건을 수확량에 대한 주요 반응 시기로 설정하였다.

먼저, 각 t시점에서 과거 $k+1$개 시점(즉, $t-0, t-1, \cdots, t-k$)의 환경 변수와 t시점의 반응변수 간의 상관성을 평가하기 위해 표본 상관계수 r_{t-i}를 계산하였다. 계산된 상관계수는 식 (2.1)을 통해 가중치 w_{t-i}로 변환되었으며, 이는 t시점의 반응변수와 각 과거 환경 변수 간의 상대적 중요도를 나타낸다. 여기서 n은 개체별 자료의 개수(즉, 총 주차의 수)이고, k는 최대 시차를 의미한다.

$$w_{t-i} = \frac{r_{t-i}}{r_{t-0} + r_{t-1} + \cdots + r_{t-k}}, \; t = 1, \cdots, n, \; i = 0, 1, \cdots, k \tag{2.1}$$

이 가중치를 활용하여, 각 t시점의 p번째 환경 변수 $x_{p,t}$에 대해 과거 $k+1$개의 시점 자료를 기반으로 가중평균값을 산출하였다. 이 값은 $x^*_{p,t}$로 표현하고 식 (2.2)와 같다.

$$x^*_{p,t} = \sum_{i=0}^{k} w_{t-i} x_{p,t-i}, \; P = 1, \cdots, p \tag{2.2}$$

이러한 과정은 모든 t시점에 대해 반복적으로 수행되며, 본 연구에서는 각 t시점에 대해 총 20개의 가중 시차 환경 변수를 생성하였다. 여기서 p는 환경 변수의 개수를 나타낸다.

마지막으로, 변환된 가중 시차 환경 변수 $x^*_{p,t}$를 이용하여 식 (2.3)과 같은 선형 회귀 모형을 구성하였으며, 이 모형을 통해 t시점의 수확량 y_t을 예측할 수 있다. 여기서 y_t는 반응변수(수확량)이며, ϵ_t는 평균이 0이고 분산이 σ^2인 독립 오차항이다.

$$y_t = \beta_0 + \sum_{p=1}^{P} \beta_p x^*_{p,t} + \epsilon_t, E(\epsilon) = 0, Cov(\epsilon) = \sigma^2 I \tag{2.3}$$

이러한 접근을 통해 환경 요인의 시간적 지연 효과와 누적된 영향을 통합적으로 반영할 수 있으며, 수확량 예측의 정확성과 신뢰도를 높일 수 있다.

2.3 연구 방법

기존의 단일 회귀 모델은 수확량과 재배환경 요인들의 비선형적인 관계를 충분히 설명하기 어려운 한계가 있다. 이에 본 연구에서는 보다 심층적이고 신뢰성 있는 예측을 가능하게 하는 부분 최소 제곱 회귀(partial least square regression, PLSR)와 서포트 벡터 회귀(support vector regression, SVR)를 활용하여 다중 선형 회귀(multiple linear regression, MLR; Lee and Kim, 2022)의 결과와 비교 분석하고자 한다. 이를 통해 변수 간 상관성과 다중공선성 문제를 해결하고, 머신러닝 모델의 예측 정확성을 유지하면서도 각 요인 간의 관계를 쉽게 파악할 수 있다. 또한, 생육·환경 요인들과 수확량 간의 비선형 관계를 효과적으로 반영할 수 있다(Kim et al., 2024). 이러한 모델 선택은 스마트팜 환경에서 실시간 모니터링과 신속한 피드백 제공이 중요함을 고려한 것으로, 연구의 목적에 부합하는

최적의 접근법을 제시한다.

첫 번째, 부분 최소 제곱 회귀(Choi and Jun, 2020)는 다중회귀분석의 일종으로 예측 변수와 반응변수 간의 관계를 분석할 때 주로 사용된다. 이 방법은 예측 변수의 선형 결합을 통해 잠재 요인을 도출하여 반응변수와 예측 변수 간의 최대 공분산을 찾는다. 이를 통해 예측 변수와 반응변수 사이의 관계를 잘 설명할 수 있도록 최적의 축소된 예측 변수 집합을 생성한다.

먼저, 예측변수 X와 반응변수 Y를 잠재 요인들의 선형 결합으로 표현하기 위해 다음과 같은 행렬 분해를 사용한다. 여기서 Z와 U는 각각 예측 변수와 반응변수의 잠재 요인 행렬, P와 Q는 로딩 행렬이며, E와 F는 잔차 행렬이다. 단, 위첨자 T는 전치행렬(transpose)을 나타낸다.

$$X = ZP^T + E,\ Y = UQ^T + F$$

Z와 U 간의 최대 공분산을 고려하여, Z를 사용해 Y를 예측하는 다음의 회귀 모형식을 구성한다. 여기서 C는 회귀계수 행렬을 의미하나, 본 연구는 반응변수가 하나이므로 회귀계수 벡터로 간주된다.

$$Y = ZC + F$$

이 방법은 다중공선성 문제를 해결하므로 설명변수 간의 상관관계가 높을 때 유용하며, 많은 변수 중 중요한 변수만을 활용하여 모델의 복잡성을 줄인다. 또한, 데이터가 많고 다차원적인 문제를 다루거나 예측 변수가 반응변수보다 많은 경우에도 유용하다. 이러한 특성 때문에 이 방법은 본 연구에서 최적화된 생육 환경을 분석하는데 적합한 방법이다.

두 번째, 서포트 벡터 회귀(Drucker et al., 1996; Cherkassky and Ma, 2004; Dhiman et al., 2019)는 서포트 벡터 머신(support vector machine, SVM)을 회귀 분석에 확장하여 적용한 방법으로, 데이터의 패턴을 학습하여 연속형 반응변수를 예측한다. 이 방법은 데이터 포인트와 예측값 사이의 오차가 특정 범위(허용 오차, ϵ) 이내일 경우 해당 오차를 무시하고, 초과 오차에만 패널티를 부과하여 최적의 예측 평면(hyperplane)을 설정하는 것이 목표이다. SVR은 허용 오차 범위 안에서 최대한 많은 데이터 포인트를 포함하면서, 오차를 최소화하는 방향으로 최적 평면을 결정한다. 즉, 허용 오차 범위를 벗어나는 데이터 포인트에 대해 패널티를 최소화하면서 마진을 최대화하는 방식으로 작동한다. 이를 수학적으로 정리하면 다음과 같은 최적화 문제 형태로 나타낼 수 있으며, 여기서 w는 예측 평면의 기울기 벡터, C는 규제 파라미터로 오차 패널티에 대한 가중치를 조정하며, ξ_i와 ξ_i^*는 허용 오차 범위를 벗어난 초과 오차를 나타낸다.

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

제약조건은 다음과 같이 정의되며, 여기서 y_i는 실제 반응변수, x_i는 데이터 포인트를 의미한다.

$$y_i - (w \cdot x_i + b) \leq \epsilon + \xi_i,\ \ (w \cdot x_i + b) - y_i \leq \epsilon + \xi_i^*$$

비선형 데이터를 처리하기 위해 가우시안 RBF 커널, 다항 커널 등 여러 가지 커널 함수를 사용할 수 있으며, 이를 통해 데이터를 고차원 공간으로 매핑하여 선형적인 예측 평면을 도출한다. 이 방법은 고차원 데이터를 효율적으로 처리하고, 복잡한 패턴을 포착하여 예측 정확도를 높이는 데 유리하다. 또한, 허용 오차 범위를 설정함으로써 잡음이 포함된 데이터를 무시할 수 있어 잡음에 강한 예측 성능을 제공한다. 이러한 장점은 본 연구 자료의 비선형적인 패턴

을 효과적으로 처리하는 데 매우 적합하다.

세 번째, 동적 시간 워핑(Müller, 2007; Senin, 2008; Zhang et al., 2023)은 두 시계열 데이터(time series) 사이의 유사성을 측정하는 척도로서 수확량과 생육·환경 자료의 관계와 같이 패턴이 비슷하지만 시간 차이가 존재할 때 유용한 방법이다. 이 방법은 데이터의 변형을 허용하고 이동이나 왜곡의 영향을 최소로 하여 데이터 포인트가 발생하는 시점이나 순서가 다르더라도 유사한 모양을 감지할 수 있다.

이 알고리즘의 원리는 각각의 길이가 m, n인 두 개의 시계열 데이터 $Q=(q_1,q_2,\cdots,q_m)$, $C=(c_1,c_2,\cdots,c_n)$를 나열하여 $m\times n$행렬로 만든 후, 식 (2.4)의 유클리드 거리 방식을 이용하여 Q와 C사이를 맵핑하여 와핑 경로(warping path)를 만든다.

$$d(q_i,c_j)=(q_i-c_j)^2 \tag{2.4}$$

다음으로, 아래 3가지 조건을 충족하는 와핑 경로들 중에서 와핑 거리 w_k들의 총합이 최소로 되는 경로를 찾는다. 즉, 식 (2.7)과 같이 와핑 경로 비용(warping path cost)이 최소가 되는 최적의 와핑 경로 W를 찾는다. 최적 와핑 경로 W는 식 (2.5)와 같이 정의되며, 여기서 $w_k=(i_k,j_k)$는 Q의 i_k번째 요소와 C의 j_k번째 요소를 매칭하는 점으로 최적 경로 상의 매칭 요소를 나타낸다.

$$W=\{w_1,w_2,\cdots,w_K\},\ \max(m,n)\le K<m+n-1 \tag{2.5}$$

와핑 경로 W는 경계조건(boundary conditions), 연속성(continuity), 단조성(monotonicity)을 만족해야 하며, 이는 워핑 경로의 필수적인 속성이다. 즉, 시작점 $W_1=(1,1)$와 종점 $W_K=(m,n)$은 이어져야 하며, 경로 $w_k=(a,b)$는 대각선 요소 $w_k=(a',b')$를 포함한 인접한 셀로 제한되며, 음의 방향으로의 이동은 허용되지 않는다.

k번째 와핑 거리 w_k의 누적 와핑 거리 $D(i,j)$는 식 (2.6)와 같이 'i 번째 요소와 j 번째 요소의 거리'와 '(i,j)번째 요소까지의 누적 와핑 거리에서 인접 셀까지의 최솟값'의 합으로 계산된다. 최종적으로, 누적 비용 행렬에서 최소 비용 경로를 따라 계산된 비용 합이 $DTW(Q,C)$의 값이다.

$$D(i,j)=d(q_i,c_j)+\min\begin{pmatrix}D(i-1,j-1)\\D(i-1,j)\\D(i,j-1)\end{pmatrix},\quad i=0,\cdots,m,\ j=0,\cdots,n \tag{2.6}$$

$$DTW(Q,C)=\min\sum_{k=1}^{K}w_k \tag{2.7}$$

따라서, 누적 비용 행렬을 기반으로 한 최적의 정렬을 통해 두 시계열의 최적 워핑 경로를 따라 매칭된 점들의 대응관계가 형성되며, 이때의 최적 정렬이 완전히 대각선 방향이면 두 시계열 데이터의 패턴이 같거나 유사함을 나타낸다.

3. 실험 결과

본 연구에서는 오이의 수확량과 생육 요인 및 환경 요인 간의 관계를 분석하기 위해 상관분석, 동적 시간 워핑(DTW), 단계적 다중 선형 회귀(MLR), 부분 최소 제곱 회귀(PLSR), 서포트 벡터 회귀(SVR)를 활용하였다. 상관분석,

MLR, PLSR, SVR은 누적 수확과중을 반응변수로 설정하여 생육 요인 및 환경 요인과의 관계를 분석하였으며, DTW는 시간의 흐름에 따른 과중과 각 요인의 패턴 유사성을 분석하기 위해 주차별 과중값을 반응변수로 사용하였다.

3.1 수확량과 생육 및 환경 요인의 관계

본 연구에서는 오이의 수확량(누적 수확과중)과 생육 요인 및 환경 요인들 간의 관계를 파악하기 위해 상관분석을 실시하여 수량 및 품질에 중요한 요인들을 확인하였다. 먼저, 오이의 수확량과 상관성이 높은 생육 요인들을 분석한 결과, 줄기굵기(Stem)와 엽수(Leaf_num)가 각각 0.3 이상의 상관 계수를 보여 오이의 수량 및 품질을 결정하는 핵심 생육 요인으로 나타났다. 이외에도 착과수(Fruit), 엽장(Leaf_len), 엽폭(Leaf_w), 초장 증가량(Plant), 마디수 증가량(Node) 등이 수확량과 양의 상관관계를 보여 수확량에 긍정적인 영향을 미치는 요인으로 분석되었다. 반면, 암꽃수(P_f)와 평균 절간장 증가량(Internode)은 오이의 수확량과 음의 상관관계를 나타내어 수확량 감소에 기여하는 요인으로 작용하였다. 다음으로, 오이의 수확량과 환경요인들의 상관성을 분석한 결과, 20개의 환경 요인들 중에서 주간 내부온도(Temp_Day), 최소 상대습도(Hum_Min), 최대 내부온도(Temp_Max). 주간 상대습도(Hum_Day), 평균 내부온도(Temp_Avg), GDD(Temp_GDD), 일출전후 1시간 평균 CO_2(CO2_Sun1h) 등이 수확량과 0.5이상의 상관성을 보여 온도와 관련된 요인들이 대체로 오이의 수확량에 중요한 역할을 하는 것으로 나타났다. 구체적으로, 온도와 관련된 요소 중에서는 주간 평균값, 상대습도와 관련된 요소 중에서는 최솟값, CO_2 농도와 관련된 요소 중에서는 일출부터 정오까지 평균값, 토양온도와 관련된 요소 중에서는 최솟값이 상관성이 높은 편이었다. 온도, 토양온도, 상대습도는 수확량과 양의 상관관계를 나타냈으며, CO_2 농도와 최대 토양온도는 음의 상관관계를 보였다(Table 2).

Table 2. Correlation between Cucumber Yield, Growth Factors, and Environment Factors

Growth factors			Environmental factors					
Factor		Cor	Factor		Cor	Factor		Cor
1	Stem	0.33	1	Temp_Day	0.58	11	Temp_Min	0.38
2	Leaf_num	0.33	2	Hum_Min	0.57	12	Temp_Night	0.38
3	Fruit	0.28	3	Temp_Max	0.55	13	CO2_Max	-0.34
4	Leaf_len	0.19	4	Hum_Day	0.53	14	Stemp_Min	0.33
5	P_f	-0.18	5	Temp_Avg	0.52	15	CO2_Min	-0.31
6	Leaf_w	0.16	6	Temp_GDD	0.50	16	Hum_Night	0.20
7	Plant	0.14	7	CO2_Sunday	-0.50	17	Hum_Max	0.16
8	Internode	-0.13	8	CO2_Avg	-0.47	18	Stemp_Max	-0.09
9	Node	0.09	9	Temp_Sun1h	0.41	19	Stemp_Avg	0.06
-	-	-	10	Hum_Avg	0.39	20	Temp_DIF	0.03

다음으로, 수확량과 생육 요인 및 환경 요인간의 관계는 단순한 선형적인 관계만으로 설명하기 어려우며, 시간의 흐름에 따라 요인들이 영향도 고려할 필요가 있다. 이를 위해 동적 시간 워핑 기법을 사용하여 시간에 따른 수확량

(수확과중)과 생육 요인 및 환경 요인의 유사도를 측정하였다.

수확량과 유사한 패턴을 보이는 생육 요인으로는 마디수 증가량(Node)과 엽장(Leaf_len)이 각각 144.02, 159.89의 비교적 낮은 거리 값을 나타냈으며, 반면 평균 절간장 증가량(Internode)은 261.89로 가장 높은 거리 값을 보였으나 수확량과 음의 상관관계를 가지므로 수확량에 미치는 영향이 적다고 할 수 없다. 환경 요인 중에서는 최소 내부온도(Temp_Min), 주간 내부온도(Temp_Day), 최소 토양온도(Stemp_Min), 일출전후 1시간 평균 내부온도(Temp_Sun1h)가 상대적으로 짧은 거리 값을 나타냈으며, 특히 각 환경 요소에서 최소 내부온도(Temp_Min), 최소 토양온도(Stemp_Min), 주간 상대습도(Hum_Day), 일출부터 정오까지 평균 CO_2(CO2_Sunday)가 수확량과 유사한 패턴을 갖는 대표적인 환경요인으로 나타났다. 반면 최대 상대습도(Hum_Max)는 수확량과 양의 상관관계를 가지면서도 거리 값이 321.94로 가장 높아, 수확량에 미치는 영향이 상대적으로 적다고 할 수 있다(Table 3).

이와 같이 수확량과 생육 요인 및 환경 요인들의 관계에서 시간의 흐름을 고려하였을 때에 중요하게 도출된 요인들과 시간의 흐름을 고려하지 않고 선형적인 관계만을 고려하였을 때의 결과가 다르게 나타났다. 이를 통해 단순한 선형적인 관계만으로 수확량과 생육 요인 및 환경 요인들의 관계를 설명할 수 없으며, 시간의 흐름에 따른 패턴도 중요한 요소임을 알 수 있다.

Table 3. Results of Dynamic Time Warping for Yield, Growth Factors, and Environment Factors

Growth factors			Environmental factors					
Factor		Distance	Factor		Distance	Factor		Distance
1	Node	144.02	1	Temp_Min	43.50	11	Stemp_Max	190.09
2	Leaf_w	159.89	2	Temp_Day	75.22	12	CO2_Max	207.68
3	Leaf_len	204.14	3	Stemp_Min	83.50	13	Hum_Night	212.73
4	Leaf_num	218.17	4	Temp_Sun1h	87.69	14	Hum_Avg	232.24
5	Plant	234.08	5	Temp_Night	112.13	15	Temp_DIF	247.31
6	P_f	234.33	6	Hum_Day	120.82	16	Stemp_Avg	275.01
7	Fruit	246.75	7	Temp_Avg	130.12	17	Hum_Min	299.37
8	Stem	257.33	8	CO2_Sunday	167.52	18	CO2_Avg	303.53
9	Internode	261.89	9	Temp_Max	173.90	19	Temp_GDD	308.57
-	-	-	10	CO2_Min	189.36	20	Hum_Max	321.94

3.2 세 가지 예측 모델의 실험 결과

첫 번째, 수확량에 영향을 미치는 주요 생육 요인 및 환경 요인을 도출하기 위해 단계적 다중 선형 회귀를 적용하였다. Table 4는 누적 수확과중을 반응변수로 하고 생육 요인과 환경 요인을 설명변수로 설정하여 다중회귀모형을 적합한 결과를 보여준다. 이 모형은 수확량의 변동을 약 83% 설명하는 높은 설명력을 보였으며, 선정된 모든 설명변수가 유의미한 영향을 미치는 것으로 나타났다(F=203.3, p=<.001). 각 회귀계수의 검정 결과, 엽장(Leaf_len), 암꽃수(P_f), 최소 상대습도(Hum_Min), 최대 내부온도(Temp_Max), 최소 CO_2(CO2_Min), 최소 토양온도

(Stemp_Min)가 수확량에 유의미한 영향을 주는 변수로 확인되었다. 다른 조건들이 동일하다는 가정 하에 엽장, 최소 상대습도, 최대 내부온도, 최소 토양온도가 높을수록 수확량이 증가하며, 암꽃수와 최소 CO_2가 작을수록 수확량이 증가하는 경향을 보였다. 이러한 결과는 상관분석 결과와도 일치하여 선택된 변수들이 수확량에 유의미한 영향을 미친다는 결론을 더욱 뒷받침해준다. 또한, 표준화 계수를 기준으로 한 수확량에 대한 영향력은 최소 CO_2(β=-.535, p=<.001), 최소 상대습도(β=.508, p=<.001), 최소 토양온도(β=.259, p=<.001), 최대 내부온도(β=.230, p=<.001), 암꽃수(β=-.149, p=<.001), 엽수(β=.115, p=<.001) 순으로 나타났으며, 이를 통해 생육 요인보다는 환경 요인이 수확량에 더 큰 영향을 미치는 것을 알 수 있다. 특히, 낮 동안의 온도와 습도의 환경 관리가 수확량 향상에 중요한 역할을 한다는 것을 알 수 있다. 결론적으로, 이 자료를 기반으로 할 때 수확량을 증가시키기 위해서는 엽장, 최소 상대습도, 최대 내부온도, 최소 토양온도와 같은 변수를 높게 유지하는 것이 유리하며, 반대로 암꽃수와 최소 CO_2는 낮추는 것이 효과적임을 알 수 있다. 최종적으로 적합된 회귀모형은 식 (3.1)과 같다.

$$\begin{aligned} f(x) = & -8500.99 + 90.48 \times Leaf.len - 169.34 \times P.f + 92.85 \times Hum.Min \\ & + 298.97 \times Temp.Max - 22.78 \times CO_2.Min + 291.92 \times Stemp.Min \end{aligned} \tag{3.1}$$

Table 4. Results of Multiple Regression Analysis

Coefficient	B	Std.Error	Beta	t-value	Pr(>|t|)
(Intercept)	-8500.994	1680.869	-	-5.057	<.001***
Leaf_len	90.476	21.699	0.115	4.170	<.001***
P_f	-169.337	33.586	-0.149	-5.042	<.001***
Hum_Min	92.854	6.481	0.508	14.327	<.001***
Temp_Max	298.967	65.477	0.230	4.566	<.001***
CO2_Min	-22.773	1.847	-0.535	4.741	<.001***
Stemp_Min	291.915	61.572	0.259	4.741	<.001***
F(sig.)	203.3(<.001***)				
R^2(Adj-R^2)	0.84(0.83)				

두 번째, 9가지 생육 변수와 20가지 환경 변수를 모두 활용하여 다중공선성 문제를 해결하고 반응변수와의 관련성을 높이기 위해 부분최소제곱회귀(PLSR)를 고려하였다. PLSR을 적용한 결과는 Table 5, Table 6, Figure 2, Figure 3에 제시되어 있다.

Figure 2는 주성분 수에 따른 예측의 평균제곱근오차(root mean squared error, RMSE)의 변화를 보여준다. 주성분의 수가 증가함에 따라 RMSE는 급격히 감소하지만, 3개 주성분 이후에는 감소폭이 미미해지는 것으로 나타난다.

Table 5는 주성분 수에 따른 X와 Y 변수의 설명된 분산 비율을 보여준다. 주성분이 추가될수록 X와 Y에 대한 설명력이 증가하지만, 주성분의 개수가 많아질 경우 해석이 어려워지며 데이터 수가 충분하지 않다면 추정 결과의 신뢰성이 낮아질 수 있다. 이러한 이유로 인해 우리는 최적의 성분 개수로 3개의 부분최소제곱 주성분을 선택하였다.

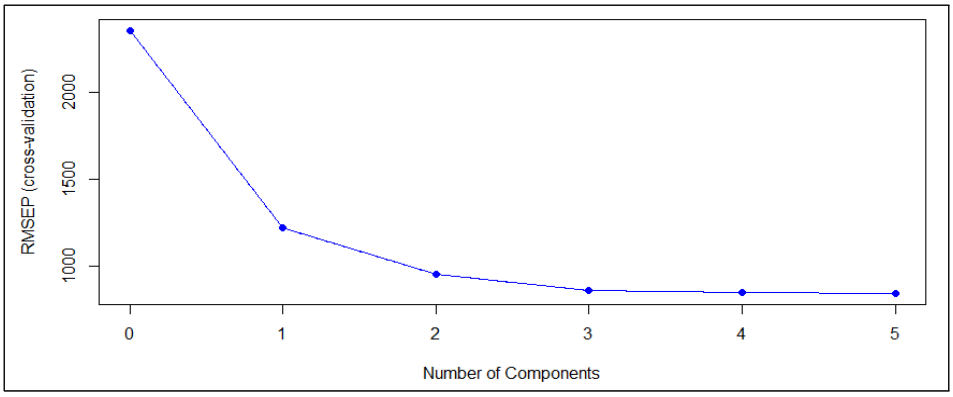

Figure 2. RMSE by Principal Component

Table 5. Proportion of Explained Variance of X and Y Variables by Number of Principal Components (%)

Components	1	2	3	4	5
X	30.71	49.01	52.01	60.58	68.92
Y	74.13	84.79	89.42	89.80	90.22

Table 6은 PLSR 회귀계수 검정 결과로, 반응변수(누적 수확과중)에 대한 주성분(PC1, PC2, PC3)의 회귀 계수 검정 결과를 나타내고 있다. 적합된 모형은 통계적으로 유의하며(F=476.8, p=<.001), 이 모형의 설명력은 86%로 나타났다. 세 개의 주성분이 모두 반응변수에 매우 유의한 양의 영향을 미치며, 표준화 회귀 계수를 통해 첫 번째 주성분(PC1)의 영향력이 가장 크고(β=1.104), 다음으로 PC2(β=0.668), 그리고 PC3(β=0.213)이 상대적으로 낮은 영향력을 가진다. 최종적으로 적합된 PLSR 모형은 식 (3.2)와 같다.

$$f(x) = 3004.32 + 804.67 \times PC1 + 616.34 \times PC2 + 355.58 \times PC3 \quad (3.2)$$

Table 6. Results of Partial Least Square Regression Coefficients

Coefficient	Estimate	Std.Error	Beta	t-value	Pr(>|t|)
(Intercept)	3004.32	57.48	-	52.269	<.001***
PC1	804.67	21.28	1.104	37.820	<.001***
PC2	616.34	30.17	0.668	20.427	<.001***
PC3	355.58	46.86	0.213	7.588	<.001***
F(sig.)	476.8(<.001***)				
R^2(Adj-R^2)	0.86(0.86)				

Figure 3은 부분최소제곱회귀의 각 주성분에 대한 변수 기여도 값을 나타낸 바 차트이며, 이를 통해 각 주성분이 수확량에 어떻게 영향을 미치는지를 알 수 있다. 첫 번째 주성분은 온도 관련 변수들이 높은 양의 기여도 값을 보이며, 두 번째 주성분은 습도 관련 변수들이 양의 기여도 값을, 토양온도와 CO_2 농도와 관련 변수들이 음의 기여도

값을 가지며, 특히 야간 상대습도, 평균 토양온도, 최소 CO_2의 절댓값이 크게 나타났다. 세 번째 주성분은 생육 특성 관련 변수들이 높은 음의 기여도 값을 가지는 것으로 나타났다. 이러한 결과는 첫 번째 주성분은 온도, 두 번째 주성분은 야간 상대습도, 평균 토양온도, 최소 CO_2, 세 번째 주성분은 생육 특성의 관계를 크게 반영하고 있다는 것을 알 수 있다.

이를 통해 생육 특성과 재배 환경이 주성분에 따라 서로 다른 방식으로 설명되고 있음을 확인할 수 있다. 또한, 여러 변수의 결합된 변동을 포착하고 변수들 간의 복잡한 관계를 하나의 모델에서 포괄적으로 이해할 수 있다.

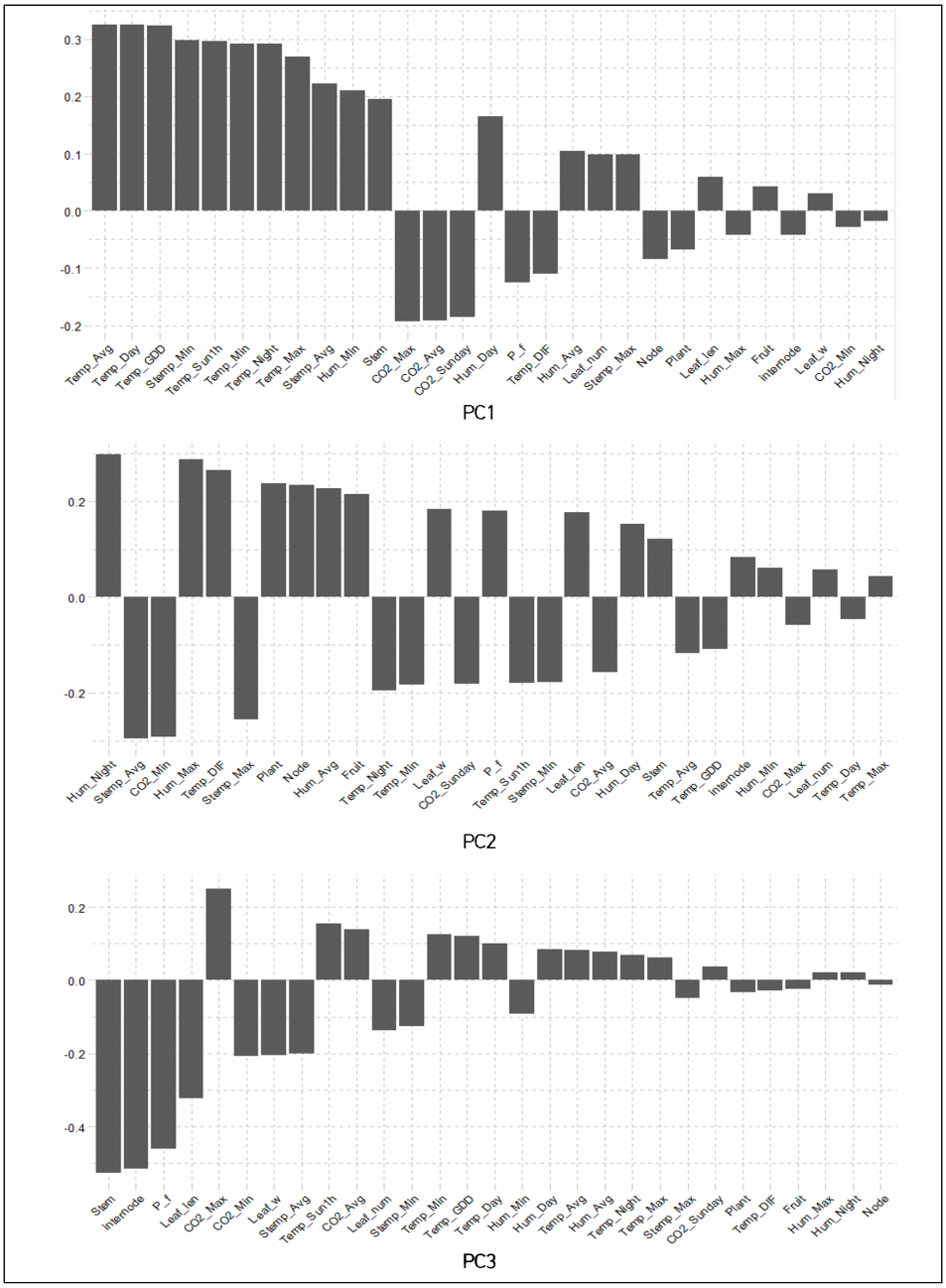

Figure 3. Contribution of Factors by Main Component

세 번째, 많은 설명변수와 복잡한 패턴에 유리하고 비선형 회귀모형에서 우수하다고 알려진 서포트벡터회귀(SVR)를 적용하여 분석을 실시하였다. 분석은 R 프로그램의 “e1071” 패키지를 이용하여 수행하였으며, 모형의 매개변수(parameter)는 커널 함수(kernel)를 ‘radial’, 코스트(cost)를 2, 감마(gamma)를 0.03448276, 엡실론(epsilon)을 0.1로 설정하였다. 다음의 Figure 4는 SVR 학습 모형에서의 설명변수별 기여도를 시각적으로 보여주고 있다. 초장 증가량과 평균 절간장 증가량은 모델이 예측을 수행할 때 중요하게 사용하는 변수로 다른 생육 요인들보다 훨씬 높은 중요도를 가진다. 그 외에도 주간 내부온도, 평균 내부온도 등 온도 관련 요소들과 최소 CO_2의 환경요인들이 상대적으로 높은 중요도를 가지며, 일교차, 줄기굵기, 최대 CO_2는 상대적으로 중요도가 낮게 나타났다.

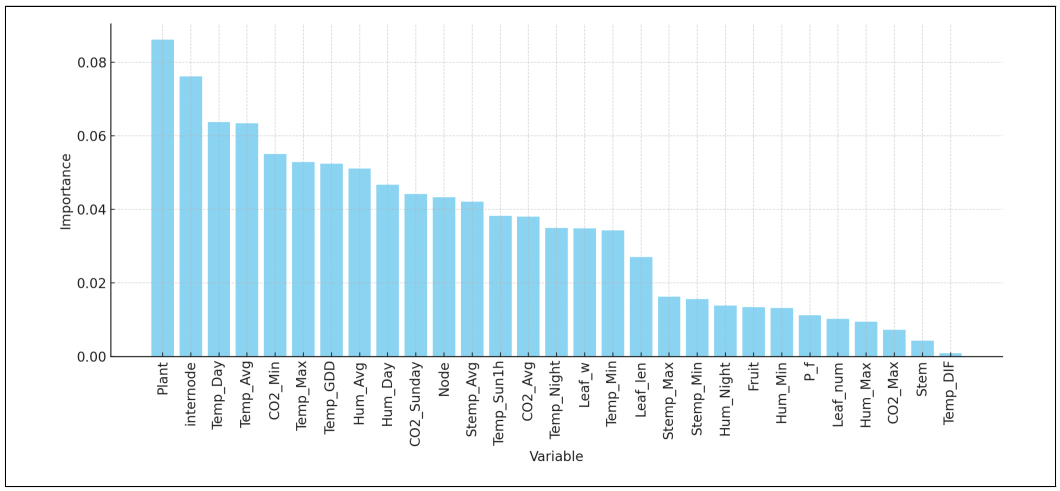

Figure 4. The Variable Importance in SVR Model

3.3 세 가지 예측 모델을 통한 핵심 요인 도출

수확량의 수량 및 품질에 중요한 생육 요인 및 환경 요인들을 알기 위해 세 가지 모델을 통해 핵심 요인들을 도출한 결과를 Table 7과 Table 8에 정리하였다. 먼저, 생육 요인은 분석 방법에 따라 핵심 변수가 모두 상이하였으며, 엽장, 암꽃수, 줄기굵기, 초장 증가량이 1회씩 가장 중요하게 나타난 것으로 확인되었다(Table 7). 다음으로, 환경요인은 모든 분석방법에서 내부온도, 상대습도, 토양온도, CO_2의 일부 요소가 중요한 환경 변수로 나타났으며, 분석방법에 따라 조금씩 다르게 나타났다. 각 환경 요소 중에서 내부온도는 평균값과 주간값, 상대습도는 최솟값, 토양온도는 최소값, CO_2는 최솟값이 2회 이상 공통적으로 중요하게 나타난 것으로 확인되었다(Table 8).

Table 7. Important Growth Factors for Yield through Analytical Methods

Method	Growth factors
MLR	Leaf_len, P_f
PLSR - 1 component	Stem
SVR	Plant

Table 8. Important Environmental Factors for Yield through Analytical Methods

Method	Temp	Hum	Stemp	CO_2
MLR	Max	Min	Min	Min
PLSR - 1 component	Avg, Day, GDD	Min	Min	Max
SVR	Day, Avg	Avg	Avg	Min

이러한 결과와 이전의 상관분석 및 동적 시간 워핑의 결과를 종합해보면, 각 변수들의 관계는 단일 변수의 독립적인 영향과 다변량 상호작용의 관점에서 다르게 해석될 수 있다. 온도와 습도는 대체로 오이 수확량과 양의 관계에 있으며, 수확량에 긍정적인 영향을 미치는 주요 환경 요인이다. CO_2는 특정 시점(일출부터 정오)에서는 양의 관계를 보이지만, 전체적으로는 부정적인 영향도 있을 수 있으며, 이는 CO_2 농도의 시점과 조건에 따라 수확량에 미치는 영향이 달라질 수 있음을 시사한다. 생육 특성 변수 중 줄기굵기와 엽수는 수확량에 긍정적인 영향을, 암꽃수와 평균 절간장 증가량은 수확량에 부정적인 영향을 미치는 변수로 작용한다.

3.4 세 가지 예측 모델의 성능 비교

Table 9와 Figure 5는 테스트 오이 자료를 바탕으로 각 예측 모델들을 통해 누적 수확과중을 추정한 결과를 보여주고 있다. 먼저, Table 9는 세 가지 예측 모델(MLR, PLSR, SVR)의 예측 성능을 비교한 결과로, 각 모델의 성능은 평균 제곱근 오차(RMSE)와 평균 절대 오차(MAE) 등의 두 가지 오차 기반 지표로 평가하였다. RMSE는 예측값과 실제값 간의 평균 제곱근 오차를 통해 예측 정확도를 평가하는 지표이다. MLR 모델은 720.63±202.81로 나타났으며, PLSR 모델은 827.34±186.53으로 다소 높은 RMSE 값을 보였으나, SVR 모델은 436.56±85.84로 세 모델 중 가장 낮게 나타났다. MAE는 예측값과 실제값 간의 평균 절대 오차로, MLR 모델은 573.80±124.63, PLSR 모델은 645.20±52.23으로 나타났으며, SVR 모델은 365.78±89.58로 가장 낮은 값을 보였다. 종합적으로 SVR 모델은 RMSE와 MAE 모두에서 가장 낮은 값을 기록하며, 큰 오차에 민감한 RMSE 측면에서뿐만 아니라 전체적으로 예측값과 실제값 간의 차이도 가장 적어 예측 성능 면에서 가장 우수한 모델로 평가된다. 따라서 오이 수확량 예측 모델로 SVR 모델이 가장 적합한 모델로 판단되며, MLR 모델과 PLSR 모델은 보조적인 비교 모델로 사용할 수 있다.

Table 9. Results of Predictive Performance Using MLR, PLSR, SVR

Method	RMSE (g/3.3m^2)	MAE (%)
MLR	720.63±202.81	573.80±124.63
PLSR	827.34±186.53	645.20±52.23
SVR	436.56±85.84	365.78±89.58

Figure 5는 실제 수확량과 세 가지 예측 모델(MLR, PLSR, SVR)의 예측 결과를 개체별로 시각화하여 보여주고 있다. SVR 모델이 전반적으로 수확량을 잘 예측하는 것을 확인할 수 있었으며, MLR 모델과 PLSR 모델은 실제 수확량의 추세는 어느 정도 잘 따라가나 일부 개체에서 과대 추정하거나 과소 추정하는 모습을 보였다. 특히, 마지막 시점의 수확량이 높은 상황(개체 1, 개체 2)에서는 모든 예측 변수들이 실제 수확량과 가까운 값을 예측하였다. 반면, 마지막 시점의 수확량이 낮은 상황(객체 3과 객체 4)에서 SVR 모델은 실제 수확량의 추세 변화를 비교적 잘 따라갔으나, 반면 MLR 모델과 PLSR 모델은 이러한 흐름을 반영하지 못하고 불안정한 예측을 하며 일부 변화 포착에 한계

가 있었다. 이는 SVR 모델이 비선형 패턴을 더 효과적으로 학습하여 급격한 증가나 감소를 더 정확히 예측하는 것으로 보인다. 따라서, 오이 수확량 예측에 있어 SVR 모델이 MLR 모델과 PLSR 모델에 비해 더 높은 예측 성능을 제공하며, 특히 비선형적이고 변동이 심한 데이터에서 유리한 모델임을 확인할 수 있었다. 결론적으로, 본 연구 자료에서는 SVR 모델이 예측 오차가 적고 실제 수확량의 변동성을 잘 반영하는 성능을 보여, 오이 수확량 예측에 가장 적합한 모델로 사료된다.

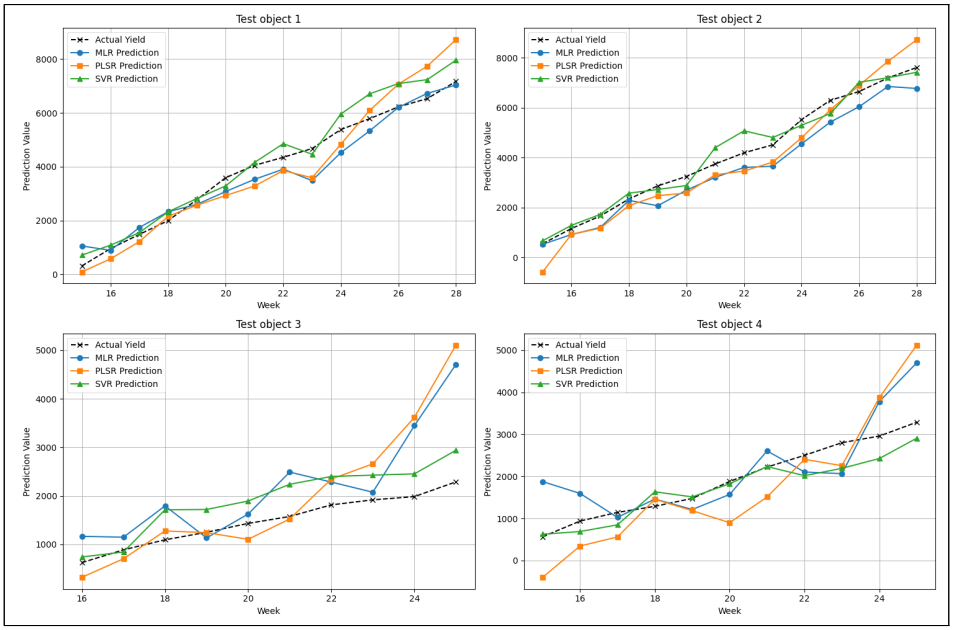

Figure 5. Comparison of Actual and Predicted Yield Values Using MLR, PLSR, SVR

4. 결 론

본 연구는 시설 작물의 수량·품질을 최대화할 수 있는 핵심 생육 요인과 환경 요인을 추출하고 작물의 수확량 예측 모델을 제시함으로써 작물의 수급 조절과 농가 생산량 향상에 유용한 정보를 제공하고자 하였다. 이를 위해 본 연구에서는 품종과 재배시기가 상이한 경기도의 스마트팜 오이 자료를 활용하여 수확량과 생육 요인 및 환경 요인 간의 관계를 살펴보고, 이러한 특성을 반영하여 오이의 수확량 예측에 적합한 모델을 제안하고 중요한 요인들을 도출하였다. 먼저, 상관분석을 통해 생육 요인 중 줄기굵기와 엽수가 상대적으로 수확량과 상관성이 높고 환경 요인 중 온도 관련 요소들이 오이의 수확량에 중요한 요소임을 확인하였다. 또한, 온도, 토양온도, 상대습도는 수확량과 양의 상관관계를 보였으며, CO_2 농도와 최대 토양온도는 수확량과 음의 상관관계를 가졌다. 동적시간워핑(DTW)을 통해 수확량과 유사한 패턴을 보이는 생육 요인은 마디수 증가량과 엽장이었으며, 수확량과 유사한 패턴을 보이는 환경 요인은 상관분석과 마찬가지로 온도 관련 요소들이었다. 특히 각 환경 요소에서 최소 내부온도, 최소 토양온도, 주간 상대습도, 일출부터 정오까지 평균 CO_2는 수확량과 유사한 패턴을 갖는 대표적인 환경요인이었다. 이를 통해 단순한 선형적인 관계만으로 수확량과 생육 요인 및 환경 요인들의 관계를 설명할 수 없으며, 시간의 흐름에 따른 패턴도 중요한 요소임을 알 수 있었다.

다음으로 단위면적당 누적 수확과중을 반응변수로 하고 생육 요인과 환경 요인을 설명변수로 하여 MLR, PLSR, SVR 등의 세 가지 일반 회귀모형을 이용하여 수확량을 예측하는 모델을 구축하였다. 첫 번째, MLR 모델에서는 엽장, 암꽃수, 최소 상대습도, 최대 내부온도, 최소 CO_2, 최소 토양온도가 수확량에 유의미한 영향을 주는 변수로 선정되었으며, 이러한 요인들로 수확량의 83%를 설명 가능하였다. 또한, 생육 요인보다는 환경 요인이 수확량에 더 큰 영향을 미치며, 특히 낮 동안의 온도와 습도의 환경 관리가 수확량 향상에 중요한 역할을 한다는 점을 알 수 있었다. 두 번째, 다중공선성 문제 해결과 반응변수와의 관련성을 높이기 위해 적용한 PLSR 모델은 3개의 주성분으로 수확량의 86%를 표현하였다. 첫 번째 주성분은 온도, 두 번째 주성분은 야간 상대습도, 평균 토양온도, 최소 CO_2, 세 번째 주성분은 생육 특성의 관계를 크게 반영하고 있었으며, 첫 번째 주성분의 영향력이 큰 편이었다. 세 번째, 많은 설명변수와 복잡한 비선형 패턴에 유리한 SVR 모델을 적용한 결과, 초장 증가량, 평균 절간장 증가량, 주간 내부온도, 평균 내부온도, 최소 CO_2 등이 상대적으로 높은 중요도를 가지는 것으로 나타났다.

세 가지 모델에서 엽장, 암꽃수, 줄기굵기, 초장 증가량이 1회씩 가장 중요하게 나타났으며, 각 환경 요소 중 내부온도는 평균값과 주간값, 상대습도는 최솟값, 토양온도는 최소값, CO_2는 최솟값이 2회 이상 공통적으로 중요하게 나타났다. 이러한 주요 요인들을 조절하는 방향으로 생육 품질을 관리하면, 품질 향상과 작물 관리의 효율성을 높이는 데에 도움이 될 것으로 기대된다.

마지막으로 세 가지 모델을 RMSE와 MAE의 지표를 통해 예측 성능을 평가하였으며, SVR 모델이 두 지표값이 가장 작게 나타났다. 또한, SVR 모델이 비선형 패턴을 효과적으로 학습하여 급격한 증가나 감소를 정확히 예측하였으나, MLR 모델과 PLSR 모델은 이러한 흐름을 반영하지 못하고 불안정하게 예측하여 일부 변화 포착에 한계가 있었다. 결과적으로 SVR 모델이 비선형적이고 변동이 심한 자료에서 유리하므로 오이 수확량 예측 모델로 가장 적합한 것으로 판단하였다.

향후 연구에서는 생육에 중요한 요소인 일사량, 양액 정보 등을 추가적으로 고려하여 보다 정밀한 예측이 가능한 모델을 개발할 계획이다. 특히, 작물 생육에 영향을 미치는 비선형적 요인과 환경 변화의 실시간 반응을 반영할 수 있는 동적 모델링 접근법을 탐구하고, 다양한 품종 및 지역별 자료를 활용하여 모델의 범용성을 확장하는 데 중점을 둘 것이다. 본 연구는 스마트팜 기술의 적용 가능성을 확대하고, 데이터 기반의 예측 모델링을 통해 작물 생육 최적화와 생산성 향상을 도모하는 데 기여할 것으로 기대된다. 또한, 데이터 분석을 통해 도출된 최적의 생육 조건을 스마트팜 운영에 적용함으로써 지속 가능한 농업을 실현하는 데도 중요한 기초 자료로 활용될 수 있다.

REFERENCES

Cherkassky, V., Ma, Y. 2004. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. Neural networks 17(1):113-126.

Choi, J. H., Jun, S. H. 2020. AI Technology Analysis using Partial Least Square Regression. Journal of the Korea Society of Computer and Information 25(3):109-115.

Dhiman, H. S., Deb, D., Guerrero, J. M. 2019. Hybrid Machine Intelligent SVR Variants for Wind Forecasting and Ramp Events. Renewable and Sustainable Energy Reviews 108:369-379.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., Vapnik, V. 1996. Support Vector Regression Machines. Advances in Neural Information Processing Systems 9:155-161.

Hong, S. E., Park, T. J., Bang, J. I., Kim, H. J. 2020. A Study on the Prediction Model for Tomato Production and Growth Using ConvLSTM. The Journal of Korean Institute of Information Technology 18(1):1-10.

Jang, I. S., and Park, G. M. 2023. A Study on the Production Forecasting of Deep Learning-Based Facility Cultivation.

Journal of Broadcast Engineering 28(4):448-456.

Jeon, M. H., Jeong, G. H., Ji, J. H., Park, I. T., Lee, H. R. 2019. Correlation Analysis between the Internode Length of Cucumber and the Amount of Production. Horticultural Society of Korea Conference Abstracts :75-75.

Kim, S. W., and Kim, Y. H. 2021. A Study on the Application of Machine Learning Algorithm toPredict Crop Production. Journal of the Korea Academia-Industrial cooperation Society 22(7):403-408.

Kim, Y. E., Song, H. J., Shin, W. S. 2024. Analysis of Machine Learning Research Patterns from a Quality Management Perspective. Journal of Korean Society for Quality Management 52(1):77-93.

Lee, K. K., Ko, K. K., Lee, J. W. 2012. Correlation Analysis between Meteorological Factors and Crop Products. Journal of the Environmental Sciences 21(4):461-470.

Lee, H. J., Lee, S. G., Kim, S. K., An, S. W., Lee, J. H., Lee, H. S., Chun, H., Choi, C. K. 2018. Changes in Yield of Cucumber as Affected by Combination with Soil Water Stress and High Temperature. Horticultural Society of Korea Conference Abstracts:81-81.

Lee, J. E., Kang, S. R., Ok, Y. J., Jeon, M. H., Na, M. H. 2019. A Study on the Optimal Environmental Factors Affecting the Growth of Facility Cucumbers. Journal of The Korean Data Analysis Society 21(6):2913-2920.

Lee, S. H., Kim, Y. S. 2022. A Pre-processing Process Using TadGAN-based Time-series Anomaly Detection. Journal of Korean Society for Quality Management 50(3):459-471.

Lee, W. B., Lee, J. S., Kim, M. S. 2024. A Study on AI-Based Real Estate Rate of Return Decision Models be 5 Sectors for 5 Global Cities: Seoul, New York, London, Paris and Tokyo. Journal of Korean Society for Quality Management 52(3):429-457.

Müller, M. 2007. Dynamic Time Warping. Information Retrieval for Music and Motion:69-84.

Na, M. H., Park, Y. H., Cho, W. H. 2017. A Study on Optimal Environmental Factors of Tomato using Smart Farm Data. Journal of The Korean Data and Information Science Society 28(6):1427-1435.

Senin, P. 2008. Dynamic Time Warping Algorithm Review. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 855:1-23.

Zhang, Y., Cucuringu, M., Shestopaloff, A. Y., Zohren, S. 2023. Dynamic Time Warping for Lead-Lag Relationships in Lagged Multi-Factor Models. arXiv preprint arXiv:2309.08800.

저자소개

강소라 전남대학교 통계학과를 졸업하고 수학/통계학과에서 통계학 석사 학위를 취득하였다. 현재는 전남대학교 수학/통계학과 박사과정에 재학 중이다. 주요 연구 관심 분야는 통계적 품질관리와 모델링을 비롯해 머신러닝 및 딥러닝 기반의 모델링에도 큰 관심을 두고 있으며, 이를 다양한 응용 분야에 적용하는 연구를 진행하고 있다.

이혜림 동국대학교 통계학과에서 학부를 졸업하고 석사학위를 취득하였다. 현재는 농촌진흥청 스마트농업팀 농업연구관으로 재직하고 있으며, 스마트온실 빅데이터 수집, 활용 연구등의 업무를 수행하고 있다. 주요관심분야는 스마트팜, 빅데이터 등이다.

나명환 서울대학교 수학교육학과를 졸업하고 통계학과에서 석사와 박사학위를 취득하였다. 현재 전남대학교 통계학과에 재직하고 있으며, 전남대학교 통계연구소 소장, 농업빅데이터 연구회 회장, 스마트팜빅데이터연구실 지도교수, 인공지능 의학연구회 학술이사, 광주전남과총 기초과학분과 위원장, 한국품질경영학회 부회장・광주전남제주지회장, 한국신뢰성학회 운영이사, 한국통계학회 호남제주지회장, 한국표준협회 창의융합개발센터 전문위원으로 활동하고 있으며, 주요 관심 분야는 스마트팜, 스마트팩토리, 그린뉴딜 관련 빅데이터분석이다.