Journal of the Korean Data & Information Science Society
2017, **28**(6), 1427–1435

http://dx.doi.org/10.7465/jkdi.2017.28.6.1427
한국데이터정보과학회지

스마트팜 데이터를 이용한 토마토 최적인자에 관한 연구[†]

나명환[1] · 박유하[2] · 조완현[3]

[123]전남대학교 통계학과

접수 2017년 10월 25일, 수정 2017년 11월 8일, 게재확정 2017년 11월 9일

요 약

최근 농업 분야에서는 빅데이터와 사물인터넷을 이용한 스마트팜의 확산이 이루어지고 있다. 스마트팜은 첨단 정보통신기술을 농업에 활용하여 농작물의 높은 생산성과 우수한 품질을 가져다줄 것으로 주목받고 있다. 스마트팜은 복합 환경제어시스템으로 온실 안에서 자라고 있는 농작물의 생육환경을 자동적으로 측정하여 실시간으로 환경 정보가 방대한 양의 데이터로 쌓이고 있다. 따라서 측정된 빅데이터를 활용한 농작물의 통계적 최적 생육환경설정 모형은 스마트팜에서 의사결정을 하는데 도움이 될 것으로 사료된다. 본 연구에서는 스마트팜 토마토 농가에서 실제로 수집된 자료를 이용하여 수확량과 환경변수의 연관성을 알아보고 이것을 토대로 수확량을 예측하기 위해 다중회귀분석을 실시하였다. 먼저 토마토 생육과정을 고려하여 환경인자에 대해서 적절한 변수 변환을 한 후 새롭게 생성된 변수들을 이용하여 모형을 적합시켰다. 그리고 적합된 통계적 모형을 이용하여 토마토의 수확량에 영향을 미치는 최적환경인자를 도출하였고, 이를 바탕으로 토마토 농가의 수확량을 예측할 수 있었다. 결론적으로 본 연구결과는 통계적 모형을 활용하여 토마토 생산성을 향상시킬 수 있는 최적의 생육환경을 조절할 수 있는 재배전략을 제시하는데 의미가 있을 것으로 기대된다.

주요용어: 수확량, 스마트팜, 시차변수, 토마토, 환경인자.

1. 서론

빅데이터와 사물인터넷 (IoT)의 활용이 활발히 이뤄지고 있다. 방대한 양의 유전자 정보를 관리하고 분석하는 생물정보학에서부터 뇌 과학, 기후기상학, 지리정보학, 지질학 등 다양한 학계의 연구뿐만 아니라 제조업, 금융, 의료, 마케팅 등 여러 산업 분야에서 활용되고 있다.

특히 농업 분야에서는 빅데이터와 사물인터넷 기술의 응용이 본격적으로 이뤄지고 있다. 최근 농업 선진국은 농업과 정보통신기술 (ICT)과의 융합을 통해 스마트팜 (smart farm)을 구축하여 빅데이터와 사물인터넷 기술을 활용한 농작물의 생산성과 품질 향상에 주력하고 있다. 사물인터넷 기술을 활용하여 센서에서 농작물의 생육·환경정보를 실시간으로 측정, 모니터링하고 최적생육관리시스템 구축을 통해 농작물을 자동적으로 관리하여 생산성과 품질이 비약적으로 증가하였다 (Ahn과 Lee, 2015).

일본은 식물공장에서 온도, 습도, 조도 등 센서를 통해 수집된 온실정보를 이용한 스케줄링 소프트웨어로 토마토의 생육환경을 제어하고, 인공광 등 최적 재배환경을 통해 농업의 공업화를 이루고 있다 (Ahn과 Lee, 2015). 네덜란드의 경우, 1945년 시설원예를 구축한 이래로 정밀하고 체계적인 환경조절 시스템과 자동화 기술의 발전으로 생산성 향상을 이루었다. 네덜란드의 채소 수출량은 2010년 42억 유

[†] 본 논문은 농촌진흥청 연구사업 (세부과제번호 : PJ011924082017)의 지원에 의해 이루어진 것임.
[1] (61186) 광주광역시 북구 용봉로 77, 전남대학교 통계학과, 교수.
[2] (61186) 광주광역시 북구 용봉로 77, 전남대학교 통계학과, 연구원.
[3] 교신저자: (61186) 광주광역시 북구 용봉로 77,전남대학교 통계학과, 교수. E-mail: whcho@jnu.ac.kr

로에 달하였지만 네덜란드의 원예작물 재배면적 중 시설채소 재배면적은 2011년 3.4% (5,041ha)에 불과하다 (Lee, 2015). 부족한 일조 등 열악한 기상환경과 부족한 노동력에도 불구하고, 토마토 생산량이 70kg/m^2에 달하는 유리온실 농가가 있으며, 2000년대 후반부터 세계 토마토 수출시장에서 2위를 차지하고 있다.

한국의 경우 정부 주도로 스마트팜의 보급과 확산이 이뤄지고 있다. 2015년 '경쟁력 제고 및 에너지 이용 구조개선을 위한 사업', '시설원예농업 분야 ICT 활성화 방안', 2016년 '시설원예 ICT 융복합확산사업'을 추진하였고, 2016년 3월 '스마트팜 확산 가속화 대책' 발표 이후 스마트팜의 확산을 통해 농업의 신성장 동력 마련을 위한 정책적 노력을 기울이고 있다 (Kim과 Chai, 2016).

스마트팜 대표작물 중 하나인 토마토는 2016년 재배면적 6,391ha, 연간 생산량 390,303톤으로 한국의 과채류 생산량 중 20.1%를 차지하지만, 토마토는 시설에서만 재배되고 있으며 노지에서는 재배되고 있지 않다. 이중 스마트팜 추정재배면적은 3.2% (202ha)에 불과하다. 현재는 스마트팜의 도입 초기단계이므로 스마트팜에서 센서를 통해 자동적으로 수집되는 대용량의 환경정보를 이용한 작물생육 최적 환경설정모형에 대한 연구는 미미한 실정이다 (Ahn과 Lee, 2015; Korea Rural Economic Institute, 2016; KOSIS, 2017).

기존의 기상기후와 관련한 농작물의 생장에 대한 연구는 표본에 대한 실험 방식과 RCP 기후 시나리오에 대한 모델링 방식 등으로 이루어지고 있다 (Choi와 Baek, 2016). 실험 방식의 특성상 온도, 습도, 광, 관수, 양액 등 여러 환경인자에 대한 적정한 기준 범위에 대한 결과로 한정되어 있고, RCP 기후 시나리오에 대한 모델링은 작물생육 최적 환경설정 모형의 도출로 귀결되지 않는다는 점에서 스마트 시설원예를 중점으로 한 생육환경 모형에 대한 연구가 요구되고 있다.

따라서 본 연구에서 스마트팜 농가에서 자동적으로 측정된 환경정보와 생산량 데이터를 활용하여 변수들 간의 연관성을 도출하는 통계적 기법을 통해 데이터를 분석하고, 스마트팜 최적 토마토 생육환경인자를 도출하고자 한다. 그러나 토마토와 같은 농작물의 환경인자를 횡적 데이터 분석 방법을 통해 데이터를 분석하는 데에는 한계가 있다. 왜냐하면 수확량과 환경정보가 시간차 없이 매칭되어 있어 적절한 시간차이를 고려하지 않고 데이터를 분석하면 토마토 생육에 영향을 미치는 환경인자의 주효과를 분석할 수 없기 때문이다. 환경인자에 따른 수확량 변화를 면밀히 분석하기 위해서는 환경설정의 효과가 발생할 시간차이를 고려해야한다.

본 연구에서는 이러한 데이터를 분석하기 위하여 토마토 생육과정을 고려하는 방법을 이용하였다. 환경설정에 따른 농작물의 생장에 대한 영향이 일정기간이 지난 후 발효됨을 고려하여 시차를 가지는 환경인자를 설명변수로 하여 토마토 수확량에 주요한 영향을 미치는 몇 개의 환경인자를 선택하기 위한 다중회귀분석을 수행하였다.

수확량을 파악하기 위해 외부온도 등 26개의 환경인자를 설명변수로 하여 다중회귀분석을 실시하고자 할 때 토마토와 같은 농작물은 꽃이 핀 이후 수확을 할 때까지 많은 시간이 요구된다. 토마토의 경우는 약 7주 정도 걸린다. 따라서 적절한 변수 변형이 필요하다. 그러나 불행하게도 아직까지 참고문헌에 환경인자를 이용한 수확량 예측 모형은 나와 있지 않다. 본 연구에서는 적절한 변수 변환의 형태를 생성하기 위해서 시차이동 (time-lag)을 실시한 후에 다중회귀분석을 실시하고 결과를 제시하였다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 분석 자료에 대해서 살펴본다. 3절에서는 토마토 생육과정을 고려한 시차를 가지는 환경 인자를 구하는 방법과 이를 이용한 데이터 분석 방법을 설명하고, 4절에서 분석 결과를 논의한다.

2. 분석 자료

분석에 사용된 자료는 스마트팜 한 농가로부터 수집된 토마토의 수확량과 환경 데이터를 사용하였다. 조사대상 농가는 연동 비닐온실의 형태로 40,000평 (13,200m^2)의 규모이며 대프니스 품종의 토마토를 재배하고 있다. 정식일자는 2014년 8월 31일로, 2014년 10월에 첫 수확을 시작하여 이듬해 7월까지 수확되었다.

일반적으로 농작물이 일주일간의 환경상태에 반응하는 것을 이용하여 변수들을 일주일 단위로 가공하여 분석하였다. 수확량 예측 모형의 반응변수로 사용된 토마토 수확량은 단위면적당 (m^2) 일주일 누적 평균수확량 (weekly cumulative yields)이다. 정식일자를 기준으로 일주일간 누적된 수확량을 단위 면적으로 환산하였다. 대상 스마트팜에서는 ICT 복합 환경제어시스템으로부터 매분마다 온실의 환경상태가 수집된다. 이렇게 측정된 환경정보를 연구 모형의 설명변수로 활용하여 수확량과의 연관성을 알아보고, 생육상태 변화의 파악이 가능하도록 일주일 단위로 누적·평균을 계산하였다.

환경정보는 크게 외기기상, 내부 환경, 관수, 양액 등으로 분류되며, 세부적으로는 다음의 Table 2.1과 같다. 평균, 최고, 최저 등 외기온도 관련 3가지 인자, 평균 누적 일사량 등 광 관련 1가지 인자, 평균, 최고, 최저 등 내부온도 관련 3가지 인자, 낮 평균, 밤 평균, 최고, 최저 등 습도 관련 4가지 인자, 잔존 이산화탄소농도 등 이산화탄소농도 관련 1가지 인자, 투입량, 배수량, 잔존량, 흡수량 등 관수 관련 10가지 인자, EC, pH 등 양액 관련 4가지 인자이다. 측정된 환경정보에 대한 자세한 설명은 Table 2.1과 같다.

본 연구의 대상 스마트팜 농가의 토마토 수확은 조사기간 동안 제1화방에서부터 제23화방에 걸쳐 이루어진다. 토마토는 일반적으로 환경조건이 좋으면 개화 후 6∼7주면 수확기에 달하는 반면, 불량한 환경조건에서는 개화 후 10 13주에 달해야 수확기가 된다. 대상 농가의 제1화방의 첫 수확은 정식 후 11주가 지나고 이루어졌다. 그 후 1∼2주 간격으로 제1화방에서부터 제2화방, 제3화방, $\cdots$, 제23화방 순으로 수확을 하였다. 따라서 일주일 단위로 생산량과 환경 인자를 이용하는 모형이 타당하다고 판단할 수 있다. 제2화방부터는 개화 후 수확을 할 때까지의 기간이 약 7주가 걸렸고, 본 연구에서는 이러한 기간을 환경인자의 수확량에 대한 반응시간 (response time)으로 보고 원래의 환경인자를 시차를 이동한 새로운 변수로 변환하였다. 반응시간을 고려한 환경인자의 변환은 다음 절에서 설명한다. Figure 2.1은 토마토의 생육단계를 나타내는 그림으로 위에서 설명한 바와 같다. 토마토의 제1화방부터 수확을 시작하여 화방 순으로 차례차례 수확하는데 이때 화방 간 수확기간의 차이는 1∼2주가 걸린다. 본 연구에서는 일주일 단위로 수확량을 계산하였으므로, 주차별 수확량은 화방별 수확량을 나타냄을 알 수 있다.

3. 통계적 자료 분석

2절에서 설명한 반응시간은 꽃이 핀 후 열매를 수확할 때까지 시간을 말하며, 보통 토마토는 7주 정도가 반응시간이다. 종속변수인 수확량 (y)에 가장 영향을 미치는 환경변수 (x_i)는 몇 주 전의 값이 가장 많이 영향을 미칠 것인가 하는 것은 알려져 있지 않다. 따라서 본 연구에서는 반응시간을 반영하기 위하여 시간차이 (time-lag)의 개념을 도입하여 통계적인 자료 분석을 위해 새로운 변수를 정의하였다, 예를 들어, Figure 3.1은 수확량 (y)에 환경변수 (x_1)의 3주전의 값이 가장 영향을 미치는 것을 도식화 한 것이다. 즉 Figure 3.1과 같이 환경변수 (x_1)의 시간차이가 3주이면, 새로운 $x_{1,-3}$을 Figure 3.1의 마지막 열처럼 변수 를 3주 이동 (shift)시켜 구할 수 있다. 이 새로운 변수 $x_{1,-3}$을 변수 x_1의 지연변수 (lagged variable)라고 하자.

Table 2.1 List of response variable and explanatory variables

		Explanation	Unit
Response variable			
Yield	y	weekly cumulative yields	kg/m^2
Explanatory variables			
Outside environment			
	$x1$	average outside temperature	℃
	$x2$	highest outside temperature	℃
	$x3$	lowest outside temperature	℃
	$x4$	weekly cumulative solar radiation	J
Inside environment			
	$x5$	average inside temperature	℃
	$x6$	highest inside temperature	℃
	$x7$	lowest inside temperature	℃
	$x8$	day average humidity	℃
	$x9$	night average humidity	%
	$x10$	maximum humidity	%
	$x11$	minimum humidity	%
	$x12$	average CO2 level (remain)	ppm
The amount of irrigation			
	$x13$	gift-driper water	
	$x14$	gift-no water	
	$x15$	gift water	
	$x16$	gift water per m^2	
	$x17$	drain water per slab	
	$x18$	drain water per m^2	
	$x19$	water per m^2	
	$x20$	water uptake per m^2	
	$x21$	drain water per	
	$x22$	ratio of drain/gift water	
Nutrient solution			
	$x23$	gift EC	dS/m
	$x24$	gift pH	pH
	$x25$	slab EC	dS/m
	$x26$	slab pH	pH

환경변수 (x_i)의 최적 지연변수를 구하기 위하여 수확량 (y)과 환경변수 (x_i)의 지연변수 $x_{i,0} = x_1, x_{i,-1}, x_{i,-2}, x_{i,-3}, \cdots, x_{i,-14}$를 만들었다. 예로, Table 3.1은 환경변수 x_1의 각 지연변수를 나타낸 것이다.

수확량 (y)과 환경변수 (x_i)의 지연변수 $x_{i,0}, x_{i,-1}, x_{i,-2}, x_{i,-3}, \cdots, x_{i,-10}$들과의 표본상관계수를 구하여 표본 상관계수가 가장 큰 것을 각 환경변수 (x_i)의 최적 지연변수로 선택하였다. Table 3.2는 수확량 (y)과 환경변수 (x_i)의 지연변수 $x_{i,0}, x_{i,-1}, x_{i,-2}, x_{i,-3}, \cdots, x_{i,-10}$들과의 표본상관계수의 일부를 나타낸 것이다.

Table 3.2를 보면 변수 x_4 (weekly cumulative solar radiation)는 7주전의 값이 토마토 수확량과 가장 상관관계가 높아 최적 지연변수는 $x_{4,-7}$이고, x_{12} (average CO2 level (remain))는 4주 전의 값이 수확량과 가장 상관관계가 높아 최적 지연변수는 임을 알 수 있다. 따라서 각 환경변수에 대한 지연변수를 고려한 토마토 수확량 회귀모형은 다음과 같이 표현할 수 있다.

$$y = \beta_0 + \beta_1 x_{1,-1} + \beta_2 x_{2,-2} + \cdots + \beta_{26} x_{26,-2}. \tag{3.1}$$