PA-74

Random forest의 중요도 지수와 상관분석을 통해 본 기상요소의 쌀수량 영향력 분석

김준환[1*], 상완규[1], 신평[1], 조현숙[1], 백재경[1], 서명철[1]

[1]전북 완주군 이서면 혁신로 181 농촌진흥청 국립식량과학원 작물재배생리과

[서론]
벼 작황예측은 국내 식량 수급 정책을 위해 중요한 문제이다. 작황예측에는 다양한 접근법이 있으나 일반적으로 기상과의 관계를 규명하여 예측하는 것이 가장 간단한 방법이라고 할 수 있다. 기상과 수량과의 관계식을 얻기 위해서는 벼가 생육하는 기간 동안 어떤 기상요소들이 어느 시점에서 중요하게 작용하는지를 파악할 필요가 있다. 본 연구는 Random forest의 중요도 지수(variable importance)로 기상요소를 선정하고 기상과 수량과의 상관을 구하여 수량에 큰 영향을 주는 기상요소를 선정하여 이들을 비교하여 가장 중요한 요소를 가려내고자 한다.

[재료 및 방법]
기상자료는 기상청의 종관기상 관측지점 중 일조시간이 관측 되고 있는 54개 지점에 한정하여 수집하였다. 수집된 기상요소는 일조시간과 평균온도, 최고온도, 최저온도, 강수량이었다. 수집된 모든 기상요소는 순별로 년도에 따라 정리하였고 54개 지점에 대한 평균값을 구하여 사용하였다. 벼의 수량은 통계청의 전국 평균수량을 가공하지 않고 사용하였다. 기상요소와 수량 모두 1981년부터 2015년까지 35년간의 자료만을 이용하였다. 시계열상의 비교이기 때문에 연도 역시 변수에 포함하였다. Random forest는 기계학습의 일종이며 예측력이 약한 많은 수의 tree regression을 구한 후 이를 평균하는 일종의 앙상블 방법이다. Random forest의 중요도지수(variable importance)는 통계프로그램 R version 2.15.에서 randomForest package를 이용하여 importance()함수를 통해 얻었다. 기상요소와 수량과의 단순 상관은 R version 2.15의 cor() 함수를 이용하여 구하였다.

[결과 및 고찰]
시계열상에서 연도 및 기상과 수량과의 단순상관 중 상관이 높은 상위 5개 변수는 년도>8월 하순 강수량>6월 초순 최저온도>9월 하순 평균온도>6월 초순 일조시간이었다. random forest의 중요도 지수(variable importance,%IncMSE)에서는 년도>8월 하순 강수>5월 중순 최저온도>6월 초순 최저온도>5월 초순 최저온도였다. 중요도 지수는 IncNodePurity로도 선정될 수 있는데 이 경우에는 년도>8월 하순 강수량>5월 중순 최저온도>6월 초순 최저온도>10월 중순 일조시간 순이었다. 모두 년도>8월 하순 강수량 순으로 공통적으로 나타났으며 6월 초순 최저온도는 순서는 달랐으나 모두 포함되는 변수였다.
연도는 시계열상에서 기술적 요소의 변동에 따라 수량이 증가하기 때문에 중요하게 나타난 것으로 생각된다. 8월 하순 강수량은 등숙 중후반기의 일조상황을 간접적으로 반영한 것으로 생각되나 만일 그렇다면 8월 하순 일조시간 역시 중요한 변수로 선정되어야만 한다. 따라서 8월 하순의 강수량은 단순한 일조시간 뿐 만 아니라 높은 습도에 따른 각종 병의 발생상황을 반영하는 것으로 추정된다. 6월 초순은 지역에 따라 이앙·활착기이거나 분얼기이다. 따라서 이때의 최저온도는 2가지 방향에서 중요할 수 있을 것으로 생각된다, 첫째 활착기간을 결정함으로서 출수시기를 결정할 수 있기 때문에 중요한 변수로 나타날 수 있다. 두 번째는 분얼기 때의 야간온도가 낮을수록 호흡손실이 적어 분얼수가 늘어날 수 있어 일차적으로 수량에 유리하게 작용할 수 있다. 따라서 설명이 가능한 중요 변수가 공통적으로 선정됨을 알 수 있었다.

[사사]
본 연구는 농촌진흥청 아젠다 사업 (과제번호: PJ012855012018) 의 지원에 의해 수행되었다.

*주저자: Tel. 063-238-5283, E-mail. sfumato@korea.kr