creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

경제학박사 학위논문

쌀 단수 예측 모형의 개선을 위한 기계 학습 기법의 적용

2020년 2월

서울대학교 대학원

농경제사회학부 지역정보전공

장 익 훈

쌀 단수 예측 모형의 개선을 위한 기계 학습 기법의 적용

지도교수 최 영 찬

이 논문을 경제학박사 학위논문으로 제출함
2020년 2월

서울대학교 대학원
농경제사회학부 지역정보전공
장 익 훈

장익훈의 박사 학위논문을 인준함
2020년 2월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

국문초록

주곡인 쌀의 수급 안정화를 위해 정부는 다양한 정책을 통해 쌀 생산 체계를 유지해오고 있으며 장기적인 식량 안보의 관점에서도 쌀 자급률을 꾸준히 관리하고 있다. 그러나 최근 WTO 개발도상국 지위의 포기로 인해 쌀 생산 농가의 불안감이 확대되고 있고, 전세계적 기후변화에 의해 기상 조건의 변동성이 커지는 등 쌀 생산에서의 불확실성은 점차 커지는 추세이다. 또한 우리나라는 기후변화협약국으로 농업부문의 기후변화 적응을 위해 벼의 재배 작기 이동도 고려해야 하는 중요한 전환점을 맞이하고 있다.

안정적인 쌀 수급을 위해서는 미래 시점의 생산량 관측이 선행되어야 하며 이를 위해서는 신뢰할 수 있는 쌀 단수 예측 모형의 개발과 운영이 필요하다. 국내에서는 농촌경제연구원이 운영하는 농업관측본부에서 쌀 단수 예측 모형을 개발하여 매년 9월 말에 예측치를 발표하고 있고, 쌀 단수 예측과 관련한 연구들은 타 작물에 비해 비교적 활발하게 연구되고 있다. 그러나 최근 예측 분야에서 많이 사용되고 있는 기계 학습 방법은 아직까지 국내의 쌀 단수 예측 분야에서는 활발하게 검토되고 있지는 않고 있다. 따라서 본 연구는 기계 학습 방법을 이용하여 쌀 단수 예측 모형의 개선에 기여할 수 있는 연구 주제를 탐색하고 도출된 연구 질문을 해결하기 위한 새로운 연구 방법을 다음의 실증 연구를 통해 제시한다.

첫 째, 쌀 단수 예측 모형의 변수 선별 방법으로 기존의 연구에서 사용되지 않은 기계 학습 방법인 베이지안 모형 평균화 기법을 이용하여 변수를 선별한 예측 모형이 기존의 예측 모형과 비교했을 때 예측력이 개선되는지 살펴보았다. 베이지안 모형 평균화에 의해 선별된 기상 변수들은 회귀분석 결과 대부분 유의한 것으로 나타났고, 교차검증 방법에의한 예측 성능 평가에서는 베이지안 모형 평균화에 의한 예측 모형의 성능이 기존의 선행연구 기반의 모형보다 우수한 것이 확인되었다. 기존의 통계 기반의 예측모형에서 사용되는 선형회귀모형(OLS)에 의한 예측보다는 서포트벡터회귀(SVR)와 같은 기계 학습그러나 2012년 이전의 과거 데이터로 학습한 예측 모형으로 최근 7년간 예측 성능을 평가했을 때에는 베이지안 모형 평균화에 의한 예측 모형의 성능이 기존 선행연구와

비교하여 오차 기반의 지표에서는 다소 앞서지만 유의미한 차이를 보이지 못하였고 설명력 기반의 지표에서는 오히려 상대적으로 단순한 선행 연구의 예측 모형보다 성능이 떨어지기도 하였다. 설명변수가 많은 복잡한 예측 모형은 불확실성이 큰 미래의 예측에서는 오히려 과적합 문제로 예측 성능이 떨어지는 것을 확인하였다.

둘 째, 종속변수의 관측치를 가지는 레이블 데이터는 부족하고 레이블이 없는 관측 데이터는 풍부한 상황에서 레이블이 없는 데이터를 예측 모형 학습에 이용하여 예측 성능을 개선할 수 있는 준지도 회귀 방법을 쌀 단수 예측 모형에 적용하였다. 결과는 쌀 단수의 예측 상황에서는 준지도 회귀 방법을 사용하지 않을 경우와 비교했을 때 오차 기반 지표에서는 평균적으로 4.6%, 설명력 기반 지표에서는 5.8%의 예측 성능 개선이 확인되었다. 2,000개 이상의 레이블이 있는 시·군 데이터를 이용한 기준모형과의 비교에서는 준지도 회귀 기법을 이용하여 200개의 레이블이 있는 도 데이터로 학습한 모형이 기준 모형의 예측 성능 대비 오차 기반 지표에서는 평균적으로 12%, 설명력 기반 지표에서는 8.9% 가량 낮은 예측 성능을 보였는데, 학습용 데이터 수를 고려하면 의미 있는 결과로 볼 수 있다. 따라서 준지도 회귀 방법은 레이블 없는 데이터를 활용하여 예측 모델의 성능을 개선할 수 있는 좋은 대안이 될 수 있다.

셋 째, 작물 단수의 조기 예측에 사용된 여러 방법들과 기존 연구에서 사용되지 않았던 기계 학습 기법을 적용한 방법을 비교했을 때 기계 학습을 사용하는 경우 예측 성능이 개선되는 것이 확인되었다. 조기 예측의 리드 타임별 예측 성능의 차이 비교에서는 4 가지 리드타임 모형 모두에서 미래 기상 변수를 사용하지 않는 모형보다 미래 기상 변수를 사용하는 모형의 예측 성능이 우월한 것으로 나타났다. 미래의 기상변수는 과거 시점의 평균 또는 중위수를 쓰더라도 미래 시점의 기상 변수를 사용하지 않는 예측 모형보다 예측 성능이 우수한 것이 확인되었다. 특히 조기 예측의 리드타임이 길어질수록 이와 같은 효과는 더 극명하게 나타났다. 또한 미래의 기상 변수를 사용하는 모형에서는 예측 시점이 일정 수준의 리드타임을 가지는 경우(본 연구에서는 수확 2~4개월 전)에서는 미래 기상 변수에 예측 모형에 의한 예측치를 적용하는 것이 예측 성능 개선에 도움이 될 수 있음을 확인하였다. 하지만 일정 수준 이상의 긴 리드타임(쌀의 경우 수확 5개월 전)에서의 단수 예측 모형에서는 기상

예측 모형에 의한 예측치 사용이 큰 이득을 주지 않는 것으로 나타났다.

실증 연구 결과들을 종합하면 작물 단수 예측 모형에서의 베이지안 모형 평균화 기법의 적용은 데이터에 기반한 변수 탐색과 예측력 개선이 가능하기 때문에 농업 분야 인공지능의 기초 기술로 발전할 수 있을 것으로 기대되며, 후속 연구의 수행을 통한 지속적인 검증이 필요하다. 준지도 회귀 방법은 시·군 단위 단수 데이터가 집계되지 않는 쌀 외의 다른 작물에 대해서도 특정 시·군 지역의 과거의 기상 정보를 이용하여 작물 재배가 적합한 지역인지에 대한 판단 근거로 단수 예측값을 제공할 수 있다는 점에서 활용이 가능하다. 조기 예측 모형의 실증 분석 결과에 따르면 예측 시점 기준에서 측정되지 않는 미래 시점의 기상 변수에 대해서는 예측 모형을 이용한 예측치를 사용하는 것이 권장된다. 이러한 맥락에서 작물별 작황 조기 예측을 위해서는 농업용 기상 예보에 초점을 맞춘 연구를 확대할 필요가 있다. 최근 기후변화로 인한 농작물 생산 체계의 불확실성이 커지면서 작황 조기예측의 중요성이 점차 증대될 것으로 예상된다. 따라서 기후변화로 인한 생산 체계의 변동성이 커지기 전에 작물별 품종 및 작부체계를 고려한 기상 요인의 영향에 대한 분석과 데이터 축적이 선행되어야 한다. 또한 다양한 품목과 지역별 분석까지 고려했을 때 현재의 국내 농업 분야의 분석 인력이 부족한데, 이를 해결하기 위해서는 농업 데이터 분석에 특화되어 있는 인공지능 기반 기술도 함께 연구하고 발전시켜야 할 것이다.

주요어 : 쌀 단수, 예측 모형, 기계 학습, 베이지안 모형 평균화, 준지도 회귀

학 번 : 2012-30990

목 차

제 1 장 서 론 ····· 1

제 1 절 연구 배경 및 목적 ····· 1

제 2 절 연구 범위 및 구성 ····· 2

제 2 장 리뷰: 통계 기반 작물 단수 예측 모형 ····· 7

제 1 절 서론 ····· 7

제 2 절 통계 모형을 이용한 작물 단수 연구 조사 ····· 10

제 3 절 주요 이슈와 개선 방향 ····· 20

제 4 절 연구주제 도출 ····· 23

제 3 장 쌀 단수 예측 모형 설명변수 선별을 위한 베이지안 모형 평균화 기법의 적용 ····· 26

제 1 절 서론 ····· 26

제 2 절 선행연구 검토 ····· 29

제 3 절 연구방법 ····· 34

제 4 절 분석결과 ····· 47

제 5 절 소결 ····· 68

제 4 장 부분지도학습 방법을 이용한 쌀 단수 예측 성능 개선 ····· 71

제 1 절 서론 ····· 71

제 2 절 선행연구 검토 ····· 73

제 3 절 연구방법 ····· 77

제 4 절 분석결과 ····· 92

제 5 절 소결 ····· 95

제 5 장 쌀 단수의 조기 예측을 위한 두 접근법에서의 기계 학습 방법 적용 ····· 99
제 1 절 서론 ····· 99
제 2 절 선행연구 검토 ····· 101
제 3 절 연구방법 ····· 102
제 4 절 분석결과 ····· 111
제 5 절 소결 ····· 118

제 6 장 결론 ····· 122
제 1 절 연구결과 요약 및 시사점 ····· 122
제 2 절 연구의 한계 및 향후 연구 방향 ····· 127

참고문헌 ····· 130

부록 ····· 143

Abstract ····· 173

표 목 차

<표 2-1> 조사 대상 논문의 게재 연도 ····· 10
<표 2-2> 조사 대상 논문의 게재 학술지 현황 ····· 12
<표 2-3> 조사 대상 논문의 항목별 내용 및 범위 요약 ····· 13
<표 2-4> 조사 대상 논문의 작물 및 국가 현황 ····· 14
<표 2-5> 조사 대상 논문의 국가별 데이터 수집 기간 ····· 15
<표 2-6> 조사 대상 논문의 작물별 데이터 수집 기간 ····· 16
<표 2-7> 조사 대상 논문의 작물별 종속 변수 ····· 17
<표 2-8> 조사 대상 논문의 작물별 설명변수 ····· 18
<표 3-1> 쌀 생산성 예측에 대한 기존 연구와 선행 연구 토대의 비교 모형의 사용 변수 ····· 41
<표 3-2> 쌀 단수 예측 모형 사용 변수 ····· 44
<표 3-3> 연도별 광역도 단수 (1997년~2007년) ····· 47
<표 3-4> 연도별 광역도 단수 (2008년~2018년) ····· 48
<표 3-5> 연도별 월 평균 기온(1997년~2007년) ····· 49
<표 3-6> 연도별 월 평균 기온(2008년~2018년) ····· 50
<표 3-7> 연도별 월 평균 강수량(1997년~2007년) ····· 51
<표 3-8> 연도별 월 평균 강수량(2008년~2018년) ····· 52
<표 3-9> 비교 모형의 선형회귀모형 계수 추정 결과 ····· 53
<표 3-10> 공선성 변수를 제외한 비교 모형의 선형회귀모형 계수 추정 결과 ····· 54
<표 3-11> 월 단위 BMA 모형의 선형회귀모형 계수 추정 결과 ····· 58
<표 3-12> 15일 단위 BMA 모형의 선형회귀모형 계수 추정 결과 ····· 59
<표 3-13> 10일 단위 BMA 모형의 선형회귀모형 계수 추정 결과 ····· 60
<표 3-14> 선행 연구 모형과 BMA 모형의 10-fold 교차검증에 의한 예측 성능 평가 ····· 62
<표 3-15> 선행 연구 모형과 BMA 모형의 리브-원-아웃(Leave-one-out) 교차검증에 의한 예측 성능 평가 ····· 64

<표 3-16> 선행 연구 모형과 BMA 모형의 최근 7년간에 대한 예측 성능 평가 ····· 67
<표 4-1> 준지도 회귀 쌀 단수 예측 모형 사용 변수 ····· 87
<표 4-2> 준지도 회귀 모형과 기준 모형 및 비교 모형의 예측 성능 비교 94
<표 5-1> 조기 예측 연구를 위한 쌀 단수 예측 모형 사용 변수 ····· 106
<표 5-2> 기상 요인 예측 모형 사용 변수 ····· 107
<표 5-3> 예측 시점별 변수 소거 모형과 BMA 모형의 10-fold 교차검증에 의한 예측 성능 평가 결과 ····· 111
<표 5-4> 예측 시점별 변수 소거 모형과 BMA 모형의 최근 7년에 대한 예측 성능 평가 결과 ····· 113
<표 5-5> 예측 시점별 평균, 중위수, 기상 예측모형 예측치 사용에 따른 최근 7년에 대한 예측 성능 평가 결과 ····· 115
<표 5-6> 두 가지 접근법에서의 4 가지 비교 모형의 최근 7년 예측 성능을 비교 ····· 117
<부표 1> 월 단위 변수의 베이지안 모형 평균화 분석 결과 ····· 142
<부표 2> 15일 단위 변수의 베이지안 모형 평균화 분석 결과 ····· 144
<부표 3> 10일 단위 변수의 베이지안 모형 평균화 분석 결과 ····· 148
<부표 4> 작물별 준지도 회귀(SSR) 방법과 OLS, SVR의 예측 성능 비교 ·· 155
<부표 5> BMA를 이용한 기상 변수 예측 모형의 변수 유형별 R^2 ···· 157

그 림 목 차

<그림 1-1> 연구의 구성 ····· 6
<그림 2-1> 연구 주제 도출 과정 ····· 24
<그림 4-1> Co-training 준지도 학습 방법 개념도 ····· 77
<그림 4-2> 회귀를 위한 CoBC 알고리즘(Hady et al., 2009) ····· 79
<그림 4-3> CoBCreg에서 연관 예제 선택을 위한 알고리즘(Hady et al., 2009) ····· 80
<그림 4-4> 준지도 회귀 모형 검증을 위한 데이터 할당 방법 ····· 89
<그림 5-1> 단수 조기 예측 방법 검증을 위한 4가지 유형의 연구 모형 개요도 ····· 101
<그림 부-1> 쌀 단수와 월 평균 기온과의 상관계수 및 분포 특성 ····· 161
<그림 부-2> 쌀 단수와 월 최고 기온과의 상관계수 및 분포 특성 ····· 162
<그림 부-3> 쌀 단수와 월 최저 기온과의 상관계수 및 분포 특성 ····· 163
<그림 부-4> 쌀 단수와 월 누적 강수량과의 상관계수 및 분포 특성 ·· 164
<그림 부-5> 1, 2월의 누적 강수량의 상관계수 및 분포 특성 ····· 165

제 1 장 서 론

제 1 절 연구 배경 및 목적

정부의 농업분야 주요 정책 중 농산물 수급 및 가격 안정 정책은 농산물 생산의 주체인 농업인은 안정적으로 농업 소득을 얻을 수 있게 하고 농산물 소비의 주체인 가구 소비자는 적절한 가격에 농산물을 구매하여 가구 소득에서 식료품 소비로 지출되는 비용에 대한 부담을 줄일 수 있도록 하는 중요한 역할을 한다. 이를 위해 작물별로 지역별 주산지협의체를 구성하여 생산자와 지자체의 자발적인 수급조절을 독려하고 사전적 재배면적 조절 기능을 강화하기 위한 육묘, 종자 구입 단계에서 부터 관련 정보를 수집하는가 하면 농업 관측을 고도화하기 위해 농산물 유통종합정보시스템을 구축하여 2019년 7월부터 시범운영을 하고 있다.

농산물 수급 안정화를 위한 사전적 조치를 위해서는 작물이 수확되기 이전에 생산량을 정확히 예측할 수 있는 조기 예측 시스템의 기능이 중요하다. EU에서는 전 가입국을 위해 MARS-Crop Yield Forecasting System(MCYFS)이 운영되고 있고(Velde & Nisini, 2019), 그 외 다양한 국가에서도 지역별 작물 수확량 예측 시스템을 운영하고 있다(Cantelaube & Terres, 2005; Capa-Morocho et al., 2016; Challinor et al., 2005; Iizumi et al., 2013; Sivakumar et al., 2000). 국내에서는 농촌경제연구원이 운영하고 있는 KREI-KASMO 모형 중 하위 시스템으로 작물별 단수 및 재배면적 예측을 통한 생산량 예측 기능이 포함되어 있으나 기상 요인을 고려하는 작물 단수 모형이 개발된 작물은 쌀, 콩, 배추, 무, 오이, 호박, 토마토, 사과 품목에 국한되어 있다(농촌경제연구원, 2005a, 2005b, 2011, 2013).

쌀은 우리나라 식량자급률의 대부분의 비중을 차지하고 있는 주곡으로 정부의 중요한 관리 품목이다. 그러나 벼의 재배면적은 1987년 126만

2324ha를 정점으로 줄어들기 시작해 2001년 108만3125ha, 2011년 85만3823ha, 2019년 72만9814ha로 꾸준히 감소하여 2019년의 재배면적은 2001년 재배면적의 67.4% 수준으로 감소하였다. 쌀 소비량의 경우 식생활의 변화로 1인당 연간 쌀 소비량이 2001년 88.9kg, 2011년 71.2kg, 2018년 61kg으로 꾸준히 줄어들어 2018년에는 2001년 소비량의 68.6% 수준으로 감소하였다. 이와 같은 쌀의 생산과 소비 감소 추세와 함께 2019년 10월에는 우리나라가 WTO(세계무역기구) 개발도상국 지위를 포기하면서 기존의 쌀 직불제 정책을 공익형 직불제 중심으로 개편하는 것이 불가피해졌다. 이로 인해 정부는 안정적인 소득 달성을 위한 제도적 장치로써의 쌀 변동직불제 폐지를 결정하였고 국내 쌀 생산 체계의 불확실성이 높아지고 있는 국면이다.

지구 온난화에 의한 기후변화는 벼의 생육에 영향을 미치는 기상 요인의 변화를 야기하기 때문에 우리나라의 쌀 생산 체계도 기후변화의 영향을 받게 된다. 벼 생육모형(ORYZA2000)을 바탕으로 할 경우, 기온과 이산화탄소 농도변화로 인해 쌀 잠재 생산량은 1971～2000년 대비 2010～2040년에는 6.4%, 2041～2070년에는 11.8%, 2071～2100년에는 15.0% 감소하지만 재배기간을 변경하였을 경우, 잠재생산량은 오히려 각각 8.1%, 12.3%, 14.9%씩 증가하는 것으로 조사되었다(한국농촌경제연구원, 2018). 농촌경제연구원의 KREI-KASMO의 단수 전망치를 베이스라인 모형으로 두고 현재 상황에서의 단수 전망치와 기후변화 적응 시나리오의 생산성 변화 가정(논물관리 감축 기술 적용 시 단수 2.8% 증가(시나리오2, 4), 작기 이동 시 단수 6.2% 증가(시나리오3, 4))을 반영하여서 2030년도 쌀 생산량을 산정 하고 소비량 예측치인 3,487천 톤에 대비한 쌀 자급률을 계산한 결과, 베이스라인은 89.7%, 작기 이동을 적용하기 전 시나리오1, 2는 83.1%, 88.6%, 작기 이동을 가정하는 시나리오3, 4는 88.2%, 94.0%인 것으로 나타났다(농촌경제연구원, 2018). 이는 기후변화 적응 방안인 논물관리 감축기술과 작기 이동이 모두 적용되는 가장 적극적인 정책 수행을 가정한 경우에도 쌀 자급률이 94.0%에 그치는 것으로 2018년의 97.3% 수준보다 낮아질 것으로 전망되고 있다. 즉 안정적인 쌀

자급률을 유지하기 위해서는 기후변화에 대한 적응 정책의 적용이 반드시 필요하며, 특히 벼의 재배 작기 이동은 향후 벼 생산 농가들에게 적용될 중요한 정책 과제가 될 것으로 예상된다. 따라서 벼의 재배 작기 이동 정책이 성공적으로 수행되기 위해서는 각 지역별로 작기가 변화할 경우 벼 생육에 영향을 미치는 기상요인이 어떻게 달라질 수 있는지에 대한 과학적인 검증이 필요하고, 이를 통해 신뢰할 수 있는 쌀 단수의 예측이 이루어져야 한다.

국내의 쌀 단수 예측은 현재 농촌경제연구원의 농업관측본부가 정책적으로 주도하고 있으며, 9월 말에 발간하는 쌀 관측월보 10월호에서 다루고 있다. 예측치는 9월 중순까지의 기상 관측 정보와 이후 시점의 기상청 예보 정보를 이용하여 KREI-KASMO 쌀 단수 예측 모형을 토대로 산출되고 있으며 벼 수매 시점의 1~2개월 전에 공표된다. 2012년 이전에는 8월 중순까지의 기상 관측 정보를 이용하여 벼 수매 시점의 2~3개월 전인 8월 말에 예측치를 전망하였으나 2008, 2009년에는 대풍으로 예측치가 실제 단수보다 6.3, 7.2% 낮았고, 2010, 2012년에는 8월 말에 발생한 태풍 곤파스와 볼라벤의 영향을 반영하지 못해 예측치가 실제 단수보다 4.5, 6.8% 높았다. 이후 2013년부터는 농촌경제연구원의 쌀 단수 예측은 벼 수매 시점의 1~2개월 전인 9월 말에 공표되고 있다. 쌀 수급 상황을 조기에 예측하는 것이 정책 당국의 중요한 과업인 것을 고려하면 수매 시점 1~2개월 전으로 예측 시기가 후퇴한 것은 아쉬운 점이다.

본 연구는 쌀 변동직불제 폐지로 인한 쌀 가격 안정화의 중요성이 고조되고 기후변화 적응을 위한 작기 이동이 정책적으로 검토되어야 하는 상황에서 쌀 단수 예측 모형의 개선에 기여할 수 있는 연구 주제를 탐색하고 도출된 연구 질문을 해결하기 위한 새로운 연구 방법을 제시하는 것을 목적으로 한다. 이를 위해 작물 단수 예측 모형에 관한 선행 연구를 리뷰하여 관련 분야에서 이슈가 되고 있는 연구 주제를 발굴하고 기존에 연구에서 많이 다루어지지 않고 있는 기계 학습 기법의 적용이 연구 질문의 해결에 긍정적인 영향을 줄 수 있는지를 검증하고자 한다.

선행연구 리뷰를 통해 도출된 연구 주제들의 연구 목적을 각각 정리

하면 첫 째, 쌀 단수 예측 모형의 변수 선별 방법으로 기존의 연구에서 사용되지 않은 기계 학습 방법인 베이지안 모형 평균화 기법을 이용하여 변수를 선별한 예측 모형이 기존의 예측 모형과 비교했을 때 예측력이 개선되는지 평가한다. 둘 째, 종속변수의 관측치를 가지는 레이블 데이터는 부족하고 레이블이 없는 관측 데이터는 풍부한 상황에서 레이블이 없는 데이터를 예측 모형 학습에 이용하여 예측 성능을 개선할 수 있는 방법을 쌀 단수 예측 모형에 적용하고 예측력의 관점에서 적용하는 것이 타당한지 평가한다. 셋 째, 작물 단수의 조기 예측에 사용된 여러 방법들과 기존 연구에서 사용되지 않았던 기계 학습 기법을 적용한 방법을 비교했을 때 예측 성능이 개선되는지 평가하고, 조기 예측의 리드 타임별로 예측 성능의 차이를 비교한다.

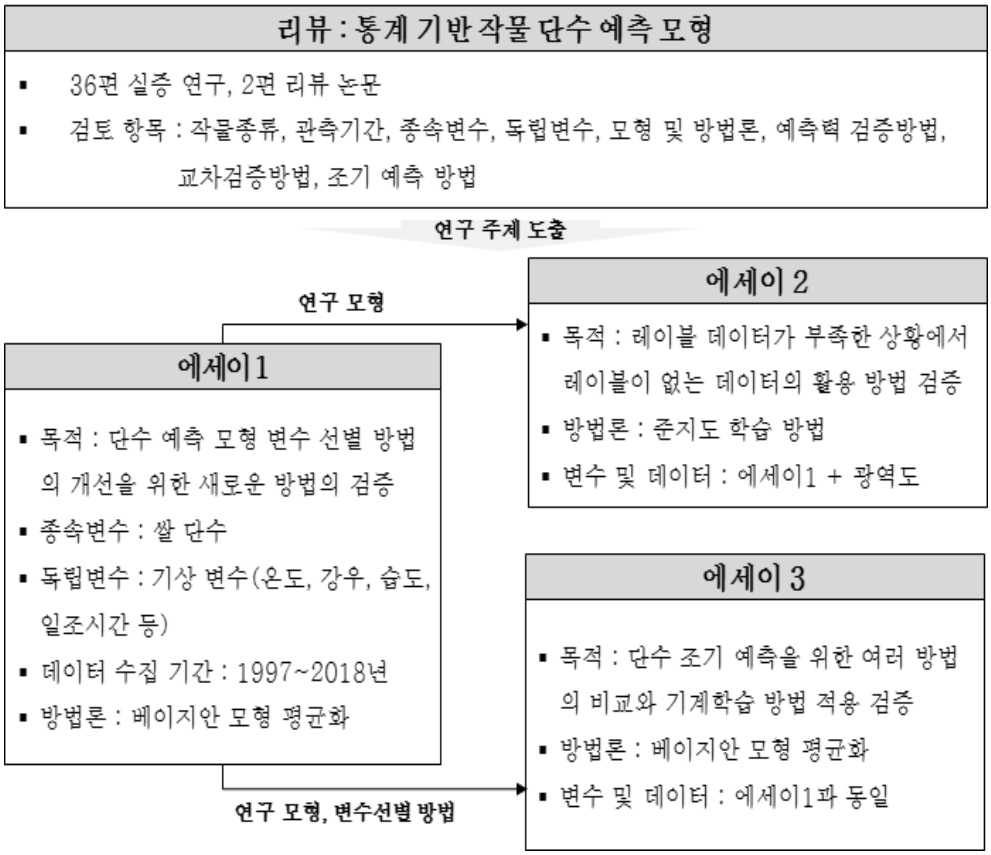
제 2 절 연구 범위 및 구성

본 연구의 대상 작물은 국내의 작물 단수 자료 중 가장 오랜 기간 많은 지역에서 수집되었고 광역도 단위와 시·군 단위에서 모두 수집이 되어 풍부한 데이터가 구축되어 있는 쌀을 대상으로 하였다. 쌀 외의 작물의 경우 광역도 단위로만 단수 데이터가 집계되어 실증분석을 할 수 있을 만큼의 충분한 데이터를 확보하는 것이 어려운 문제로 연구 대상으로 고려하지 않았다. 한편 단수 예측을 위한 연구 방법의 탐색은 쌀 외의 모든 작물로 확대하여 범위에 제약을 두지 않았다. 이는 본 연구에서 사용하는 쌀 단수 예측 모형이 통계 기반의 작물 예측 모형으로 많은 매개변수를 필요로 하고 작물 모형의 이해를 위해서는 생육과 관련한 전문적인 지식을 필요로 하는 작물 시뮬레이션 모형과 달리 통계 기반 모형은 설명변수에 대한 사전지식을 필요로 하지 않는 유연한 방법론이기 때문에 모든 작물로 범위를 확대하였다.

본 연구의 구성은 통계 기반의 작물 단수 예측 모형에 관한 선행연구 리뷰를 통한 연구 주제 도출에서부터 시작되며, 리뷰에서 도출된 연구 주제로 세 편의 에세이를 구성하였다. 첫 번째 에세이는 쌀 단수 예측 모형 설명변수 선별을 위해 베이지안 모형 평균화 기법을 적용하는 것으로 국내에서 수행된 쌀 단수 예측에 관한 선행연구들의 예측 모형과 베이지안 모형 평균화를 적용한 예측 모형의 예측력을 비교한다. 이 때 베이지안 모형 평균화를 이용한 변수 선별의 장점을 활용하기 위해 기존의 월단위 집계 변수가 아닌 반기(15일), 순기(10일) 단위의 기상 변수로 구성된 예측 모형도 함께 비교하였다. 두 번째 에세이는 종속변수의 관측치를 가지는 레이블 데이터는 부족하고 레이블이 없는 관측 데이터는 풍부한 상황에서 유용하게 활용될 수 있는 부분지도학습 방법이 작물의 단수와 기상 요인 간의 모형에 적용될 수 있는지 검증하기 위해 쌀 단수 예측 모형에 부분지도회귀 방법을 적용하여 예측력 개선을 확인한다. 세 번째 에세이는 쌀 단수의 조기 예측을 위해 기존 작물 단수 예측 연구에서 사용된 두 접근범의 방법들에서 검토되지 않았던 기계 학습 방법을

적용했을 때 예측력이 개선되는지 확인하고, 조기 예측의 리드타임이 증가함에 따른 예측 성능의 변화를 비교 분석한다.

<그림 1-1>은 본 연구의 구성에 대한 요약도로 에세이1에서의 연구 결과에서 도출되는 연구 모형은 에세이2와 에세이3에서도 활용된다. 에세이1에서의 검증되는 베이지안 모형 평균화 기법에 의한 변수 선별 방법은 에세이3에서 단수 예측 모형과 기상 요인 예측 모형의 변수 선별에 활용된다.



<그림 1-1> 연구의 구성

제 2 장 리뷰: 통계 기반 작물 단수 예측 모형

제 1 절 서론

작물 단수 예측을 위한 방법론은 작물 시뮬레이션 모형(Crop Simulation Model)과 통계 모형(Statistical Model)의 두 접근법으로 분류된다. 작물 시뮬레이션 모형은 프로세스 기반 모형((Process-based model)으로도 명명되며 탄소동화, 작물 성장, 토양수분이동, 비료효과 등 물리적, 생물학적, 화학적 과정을 나타내기 위한 기계론적 모형으로 설명된다(Whisler et al., 1986; Kaul et al., 2005). 작물 시뮬레이션 모형은 정해진 범위 내에서 변수를 조절할 수 있는 결정론적 하위 모형을 사용하여 작물 수확량과 관련된 생물학적 프로세스를 시뮬레이션한다. 작물 시뮬레이션 모형은 특정 환경 조건 하에서 식물이 어떻게 성장하는지 예측하고 작물의 환경 및 관리에 대한 반응 정보를 보여줄 수 있는 장점이 있어 제 3, 4 차 기후변화 평가 보고서에서 농경 시스템에 대한 기후 변화의 잠재적 영향을 평가하는데 주요하게 사용되었다(Gitay et al., 2001; IPCC, 2007). 그러나 밀, 옥수수, 콩, 쌀과 같은 주요 작물에 대한 모형은 연구가 많이 되었으나 나머지 작물 모형에 대한 연구는 부족하여 사용이 제한되는 문제가 있다. 통계 모형(Statistical Model)은 변수 간 상관 요인을 기반으로 하며 상대적으로 단순하고 수집된 영향 요인 범위 밖의 영역에서는 사용이 어렵지만 필요한 요인의 수는 상대적으로 적다(Kaul et al., 2005). 또한 설명 변수에 대한 제약이 적기 때문에 다양한 변수를 사용할 수 있어 주요 작물 외에 다양한 품목에 대한 적용이 가능하다.

작물 시뮬레이션 모형은 많은 연구를 통해 우수한 성과를 올리고 있으나 몇 가지 제약을 가지고 있다. 개별 하위 프로세스와 관련된 지식을 사용하지만 측정이 어렵거나 비용이 많이 드는 광범위한 기상 및 환경 변수를 포함하는 경우가 있는 경우에는 모형 개발과 사용이 복잡하고 많은 입력 변수를 필요로 한다(Basso et al., 2001; Kaul et al., 2005;

Wang et al., 2003). 입력 변수가 충실하게 수집되면 수확량에 대한 정확한 예측을 제공하지만 예상 범위 밖의 관측치를 이용한 추정에서는 불확실성이 커진다(Kaul et al., 2005). 즉, 작물 시뮬레이션 모형의 결과는 토양 조건, 날씨 및 관리 지수에 민감하게 반응하는 경향을 가진다(Shi et al., 2013). 한편 대부분의 작물 모형은 극한 기후 현상이 작물의 성장과 발달에 미치는 영향을 제대로 처리하지 못한다(Moriondo et al., 2011). 이러한 성능 저하는 특정 생육 프로세스에서 결과를 단순화하여 부정확한 결과를 초래할 수 있다(Feng et al., 2018). 또한 작물 모형은 지역 환경에서 훈련하고 보정하기 위해 수년 간의 실험 데이터를 필요로 하며 (Chen et al. 2010), 다른 지역에서 사용될 때 재보정이 수행되어야 한다. 이는 대부분의 경우 작물 시뮬레이션 접근 방식이 통계적 접근 방식보다 비용이 상당히 높을 수 있음을 시사한다(Lobell & Asseng, 2017). 작물 시뮬레이션 모형 개발 비용을 정확하게 측정하는 것은 어려우나 한 가지 증거로 한 작물에 대한 AgMIP 논문은 30 명 이상의 저자(예: Asseng et al., 2013)가 연구한 반면, 카사바, 기장 및 땅콩과 같은 작물을 포함한 여러 국가와 작물에 대해서는 통계적 연구에서는 1~2명의 저자(예: Schlenker & Lobell, 2010)가 연구하는 점에서 추측이 가능하다(Lobell & Asseng, 2017). 불확실성에 영향을 주는 다양한 요인이 있는 점도 작물 시뮬레이션 모형의 문제점 중 하나이다. 구조 설계 방법, 모형의 가설 설정, 데이터의 품질, 모형 보정 및 애플리케이션 등에서 불확실성이 발생하여 중첩되게 되면 불확실성이 어디에서 발생했는지 탐지가 어려게 된다(Zhang et al., 2015b). 또 다른 단점으로는 작물 시뮬레이션 모형에서는 포착할 수 없는 많은 단수 영향 요인이 있다는 점이다(Gornott & Wechsung, 2016). 경작 적응(관개, 작물품종, 농경기술) 또는 해충, 잡초 및 질병에 대한 기후 유발 효과가 그 예가 될 수 있다(Mueller et al., 2012).

지금까지 살펴본 작물 시뮬레이션 모형의 한계점을 보완하기 위한 방법으로 통계 모형은 좋은 대안이 될 수 있다. 밀 단수 연구에서는 시계열 데이터를 이용하여 생산량과 날씨 변수 간의 관계를 실증적으로 밝혀

통계 모형의 유용성을 확인하였다(Lobell and Burke, 2010; Lobell et al., 2011; Nicholls, 1997). 통계 모형은 데이터 중심적이며 토양 특성과 같은 보조 정보는 더 적게 필요한 반면 작물 시뮬레이션 모형에서와 같은 선행 정보가 거의 없이도 과거 데이터에만 의존하여 모형 개발이 가능하다(Mathieu & Aires, 2018). 지구 온난화로 인한 극단적인 기상 상황의 잦은 발생으로 작물 시뮬레이션 모형의 오류가 확대되고 있는 상황에서 통계 모형을 이용한 보완이 가능하다(Zhang et al., 2015b). 단수에 대한 기상 요인의 영향이 달라지더라도 통계 모형에서는 요구되는 데이터의 형식이 단순하여 모형의 보완이 쉬운 이점이 있고 모형의 불확실성을 R^2 등의 지표로 투명하게 평가할 수 있는 장점도 있다(Lobell & Burke, 2010). 또한 실측 데이터를 이용한 실증 연구에서 통계 모형의 예측 정확도가 높은 것이 지속적으로 보고되고 있다(Barnett et al., 1997; Qian et al., 2009; Bornn & Zidek, 2012).

본 리뷰는 작물 단수 예측의 주류 연구 분야인 작물 시뮬레이션 방법을 보완할 수 있는 방법으로 통계 모형을 이용한 실증 연구를 대상으로 한다. 리뷰는 최신의 연구 동향을 조사하고 통계 모형에서 사용되고 있는 수집 데이터의 특징, 종속 변수의 유형, 기상 변수의 범위, 변수 선별의 방법, 예측 모형의 학습 기법, 예측 성능의 검증 방법 등을 비교한다. 이를 통해 통계 모형 연구가 가지는 공통점과 차이점을 정리한다. 최종적으로는 통계 모형을 이용한 연구들에서 제기하고 있는 주요 이슈가 무엇인지를 살펴보고 작물 단수 예측에 관한 새로운 연구 주제를 탐색하고 본 논문의 세부 연구 주제를 도출하는 것을 목표로 한다.

제 2 절 통계 모형을 이용한 작물 단수 연구 조사

작물 단수와 기상 요인 간의 관계에 대한 통계 모형 연구 중 비교적 최근에 연구되었고 국제적으로 저명한 학술지에서 소개된 실증 연구 논문 36편을 선정하여 조사하였다. 다양한 후보 논문들 중 이들 논문을 선정한 기준은 첫 째, 선행연구가 충실히 수행되었는지. 둘 째, 통계 모형 분석에 적용한 방법론이 과학적이고 최신의 추세에 뒷떨어지지 않았는지. 셋 째, 분석 결과에 대한 충분한 논의가 포함되었는지의 3가지 기준으로 선정하였다. 36편의 실증 연구 외에도 작물 단수 예측 모형에 대한 다양한 지식과 이슈를 파악하기 위하여 2 편의 리뷰 논문도 조사를 하였으나 기초 통계에는 포함시키지 않았다. 선정된 연구 목록은 <부록 1>에 표로 정리하였다.

조사 대상 논문이 게재된 시점은 2005 ~ 2019년까지 다양하나 최근 5년 이내에 발표된 논문이 전체 36편 중 23편으로 다수를 차지하고 있다. 작물 단수 예측을 위한 통계 모형에서 다루고 있는 방법론에서 최신의 기법이 적용되고 최근 기후변화에 의한 기상 요인의 영향이 반영된 분석 결과를 조사하기 위해서 최근에 게재된 논문을 중심으로 조사를 수행하였다.

<표 2-1> 조사 대상 논문의 게재 연도

연도	빈도	연도	빈도
2005	3	2014	2
2008	1	2015	4
2009	3	2016	7
2010	1	2017	3
2011	1	2018	5
2012	1	2019	4
2013	1	총합계	36

조사 대상 논문이 게재된 학술지의 현황은 <표 2-2>와 같다. Agricultural and Forest Meteorology가 13편으로 가장 많아 작물 단수 예측과 관련한 주제가 가장 많이 다루어지고 있는 것으로 추측할 수 있다. 다음으로는 Agricultural Systems가 4편, Climatic Change, European Journal of Agronomy, Journal of Agricultural Science, Scientific Reports 2편, 나머지 학술지가 1편씩을 차지한다. 2018년 학술지 영향도지수(Journal Impact Factor)는 1.33 ~ 21.72의 범위에서 다양하였으며, 4점대 전후의 학술지가 가장 많은 비중을 차지하였다.

<표 2-2> 조사 대상 논문의 게재 학술지 현황

학술지명	빈도	IF(2018)
Agricultural and Forest Meteorology	13	4.19
Agricultural Systems	4	4.13
Climatic Change	2	4.17
European Journal of Agronomy	2	3.38
Journal of Agricultural Science	2	1.33
Scientific Reports	2	4.01
Agriculture Ecosystems & Environment	1	3.95
Agronomy for Sustainable Development	1	4.26
Field Crops Research	1	3.87
Geophysical Research Letters	1	4.58
Global Change Biology	1	8.88
International Journal of Biometeorology	1	2.38
Journal of Applied Meteorology and Climatology	1	2.36
Nature Climate Change	1	21.72
Plos One	1	2.78
PNAS*	1	9.58
Theoretical and Applied Climatology	1	2.72
총합계	36	

* Proceedings of the National Academy of Sciences of the United States of America(PNAS)

<표 2-3> 조사 대상 논문의 항목별 내용 및 범위 요약

항목	내용 및 범위
작물종류	밀 , 옥수수 , 쌀 , 대두 , 사탕수수 , 면화 , 수수 , 유채 , 기름야자 , 해바라기 , 감자 , 사탕무
관측기간	10~19 년 (3), 20~29 년 (14), 30~39 년 (12), 40~49 년 (4) , 50 년 이상 (4)
종속변수	단수 (23), 시계열 회귀선 기준 단수 편차 (7), 단수 차분 (6)
독립변수	기온 , 일교차 , 강수량 , 일사량 , 습도 , Killing(High) Degree Days(KDD), Freezing Degree Days(FDD), Growing Degree Days(GDD), Consecutive Wet Day , Consecutive Dry Day, 포화수증기압 (VPD),잠재증발산량 (PET), 성장기 가용수분 , 수분스트레스 , 토성 , 토양벌크밀도 , 진흙밀도 , 토양수분 , 수리전도도 , 토양유기탄소량 , 엘니뇨 등 기후지수 , 관개유무 , 위도 , 고도 외 해당 변수의 최대 , 최소 등 파생변수
변수선별 방법	임의 지정 (15), 상관계수 /회귀계수 /R 2 순위 (8), Stepwise 설명력 비교 (8), Random Forest 중요도 지표 (2), PCA/Lasso 등 변수축약 (2), PLS(1)
모델링 방법론	OLS 단독 (15), Random Forest(4), 시계열 /패널회귀 (4), 인공신경망 (3), Locally weighted Regression (3), Polynomial Regression(2), PLS/PCA(2), OLS 와 시뮬레이션모형 비교 (2), Multi-Level Model(1)
예측력 평가지표	실측치-예측치 산포도의 R^2 , RMSE , RRMSE, MAE, MPE, MAPE, NSE(Nash-Sutcliffe Efficiency) 등
교차검증 방법	36 개 실증연구 중 23 개 Leave-one-out(연단위), k-fold, bootstrap, 샘플 임의 분할 (과거 /최근)
사전예측 기간	36 개 실증연구 중 13 개에서 사전 예측력 검토 . 9 개는 Blind 변수 제외 , 4 개는 과거 기상 이용 .

36편의 조사 대상 논문에서 다루고 있는 작물 종류, 관측 기간, 종속변수, 독립변수, 변수 선별 방법, 모델링 방법론, 예측력 평가지표, 교차검증 방법, 사전 예측 기간에 대한 내용과 범위는 <표 2-3>에 요약하였

다. 각각의 항목에 대한 구체적인 설명은 단일 항목 또는 2개 항목의 교차분석표로 현황을 정리하고 각 항목에서의 특징이 있는지 살펴본다. 현황 정리에서 하나의 논문에 여러 작물의 분석 결과가 포함되어 있는 경우에는 작물을 분리한 후 하나의 논문 빈도가 중복되는 것을 허용하여 집계하였다.

먼저 조사 대상 논문의 작물 및 국가 현황을 살펴보면 국가별로는 미국, EU, 호주, 독일, 중국, 캐나다 사례가 많았고, 작물에서는 옥수수, 밀, 대두 순으로 나타났다<표 2-4>. 미국, EU, 호주와 같은 농업 선진국에서 연구가 가장 활발하였고 중국을 대상으로 한 연구들도 5건으로 적지

<표 2-4> 조사 대상 논문의 작물 및 국가 현황

국가	합계	옥수수	밀	대두	사탕수수	감자	유채	기타
미국	14	8	1	3		1		면화
EU	9	2	3			1	1	사탕무, 해바라기
호주	8		5		2			수수
독일	6	3	3					
글로벌	5	1	2	1				쌀
중국	5	5						
캐나다	4	1	2	1				
브라질	3	1	1	1				
프랑스	3	1	2					
러시아	2	1	1					
덴마크	1						1	
말레이시아	1							기름야자
멕시코	1	1						
브루키나파소	1	1						
아르헨티나	1		1					
남부아프리카	1	1						
인도	1							커피
작물별 합계	66	26	21	6	2	2	2	

않았다. 글로벌 범위의 연구는 대부분 기후변화에 의한 주요 식량 작물에 대한 생산량 변화를 전망하기 위한 연구에 초점이 맞추어져 있다. 작물별로는 옥수수는 미국과 중국에서 연구가 두드러지게 많았다. 밀은 EU, 호주에서 많이 연구되었고 대두는 미국이 3건, 캐나다, 브라질이 1건씩 다루었다. 사탕수수는 호주의 사례만 조사되었고, 감자는 미국, EU 1건씩, 유채는 EU, 덴마트에서 1건씩 조사되었다. 그 외에 1건만 조사된 작물은 면화, 사탕무, 해바라기, 수수, 쌀, 기름야자가, 커피가 있었다. 우리나라의 주곡인 쌀에 대한 연구는 작물 시뮬레이션 모형 연구에서는 빈번하게 발견되나 통계 모형에서는 아직까지 많이 다루어지지 않고 있는

<표 2-5> 조사 대상 논문의 국가별 데이터 수집 기간

행 레이블	합계	10~19 년	20~29 년	30~39 년	40~49 년	50 년 이상
미국	14		3	6		5
EU	9		7	2		
호주	8	1	1	3	1	2
독일	6		4	2		
글로벌	5		4			1
중국	5	1	2		2	
캐나다	4			4		
브라질	3			3		
프랑스	3		3			
러시아	2			2		
남부아프리카	1				1	
덴마크	1		1			
말레이시아	1		1			
멕시코	1			1		
브루키나파소	1			1		
아르헨티나	1	1				
인도	1			1		
총합계	66	3	26	25	4	8

것이 확인되었다.

조사 대상 논문의 국가별 데이터 수집기간은 <표 2-5>와 같이 집계되었다. 연구들에서 수집된 데이터의 기간은 20~39년이 전체 66건 중 51건으로 대부분을 차지하고 있다. 국가별로는 미국, 호주, 캐나다와 같은 농업 선진국에서 30년 이상의 데이터를 이용한 연구가 많았고 그 외의 국가 중에는 브라질, 러시아를 대상으로 한 연구에서 30년 이상의 데이터가 사용되었다.

조사 대상 논문의 작물별 데이터 수집기간은 <표 2-6>와 같다. 작물별로는 주요 식량 작물인 옥수수, 밀, 대두의 경우 장기간 데이터가 많이 사용되고 있는 것으로 나타난 반면 반면 빈도가 적은 작물에서는 데이터 수집 기간도 길지 않은 편으로 주요 식량 작물의 경우는 많은 연구가 수행되어 온 만큼 수집된 데이터도 오랜기간 축적되고 있는 것을 확인할

<표 2-6> 조사 대상 논문의 작물별 데이터 수집 기간

작물	작물별 합계	10~19 년	20~29 년	30~39 년	40~49 년	50 년 이상
옥수수	26	1	9	11	3	2
밀	21	1	7	9	1	3
대두	6		2	3		1
감자	2		1			1
사탕수수	2		1	1		
유채	2		2			
기름야자	1		1			
면화	1					1
사탕무	1		1			
수수	1	1				
쌀	1		1			
커피	1			1		
해바라기	1		1			
총합계	66	3	26	25	4	8

수 있다.

조사 대상 논문에서 사용된 종속변수를 작물별로 살펴보면 <표 2-7>와 같다. 작물별로 분리하여 집계한 총 58건 중 40건의 연구에서 종속변수로 단수를 그대로 사용하였고 단수 차분 변수는 12건, 트렌드를 제거한 단수 편차 변수는 6건에서 사용되었다. 작물에 따른 종속변수 사용의 차이는 특별히 관찰되지 않았다. 오히려 연구자 그룹 중 특정 그룹에서 단수 차분 또는 단수 편차 변수를 선호하는 것으로 나타났다. 단수의 절대값 대신 차분 또는 편차와 같이 단수의 변화를 쓰는 이유는 종속변수의 시계열 추세를 제거하고 외생 변수에 대한 체계적인 편향을 줄이기 위한 것으로 설명되고 있다(Lobell, 2013).

조사 대상 논문에서 사용된 설명변수를 작물별로 살펴보면 <표 2-8>와 같다. 기상 관측소에서 수집되는 기온, 강수량, 일사량의 직접 변수 외에

<표 2-7> 조사 대상 논문의 작물별 종속 변수

종속변수	단수	단수 차분	단수 편차	합계
옥수수	15	3	3	21
밀	10	6	2	18
대두	4	1		5
사탕수수	1	1		2
유채	2			2
감자	2			2
기름야자	1		1	2
면화	1			1
사탕무	1			1
수수	1			1
쌀		1		1
커피	1			1
해바라기	1			1
전체작물	40	12	6	58

<표 2-8> 조사 대상 논문의 작물별 설명변수

변수	합계	옥수수	밀	대두	유채	감자	사탕수수	기름야자	면화	사탕무	수수	커피	해바라기	쌀
강수량	46	19	12	4	2	2	2	1	1	1		1	1	
기온	27	13	5	3	1	1		1	1			1		1
일사량	16	6	5	1	1	1	1	1						
PET	11	3	6	1				1						
건조지수	8	2	2		1	1				1			1	
습윤지수	8	2	2		1	1				1			1	
연도	8	2	3	1			1		1					
토양수분	8	1	4	1							1			1
비료투입	5	3	2											
수분stress	5		3				1				1			
GDD	4	2	2											
VPD	4	3				1								
거시기후	4	1	1				1	1						
FDD	3	1	1	1										
KDD	3	1	1	1										
지리적위치	3	1	1			1								
토성	3	1	1		1									
온도	2	2												
일교차	2		1				1							
토양등급	2	1		1										

일부 연구들에서는 잠재증발산량(Potential Evapotranspiration: PET), 유효적산온도(Growing Degree Days: GDD), 냉해적산온도(Freezing Degree Days: FDD), 열손상적산온도(Killing Degree Days: KDD), 포화수증기압차(Vapor Pressure Deficit: VPD), 수분스트레스(Crop Water Strees Index: CWSI)와 같은 파생변수를 사용하였다. 이 외에 바이오매스(Biomass), 강우파생 변수, 경사도, 고온지수, 기술변화, 발아시점, 상대습도, 성장기저 온도, 잠재적 물사용량, 전작기 작물, 지역더미, 질소비료 비율, 토양 깊이, 토양 밀도, 파종일, 품종, 하지 기간, 해수온도 변수가 1회씩 사용되었다. 사용 빈도에서는 강수량 변수가 46회로 가장 많이 사용되었고 다음으로 기온 27회 일사량 16회 잠재증발산량(PET)이 11회 사용되었다. 강수량, 기온, 일사량 변수는 작물의 단수를 설명하는 중요한 변수로 모형의 중요한 변수로 사용되는 것을 알 수 있다. 강수량, 기온, 일사량 변수는 직접 측정된 관측치를 월단위 또는 생육기간 내에서 집계하여 사용하는 경우가 많았으나 일부 연구들에서는 온도의 경우 유효적산온도(Growing Degree Days: GDD), 냉해적산온도(Freezing Degree Days: FDD), 열손상적산온도(Killing Degree Days: KDD)와 같은 파생변수, 강수량의 경우 습윤지수, 수분스트레스, 폭우지수, 잠재적 물사용량 등의 파생변수에 각 기상 요소의 속성을 반영하여 사용하는 연구들도 있으며 이러한 새로운 변수가 작물 단수 예측 모형의 설명력과 예측 성능을 얼마나 개선하는지에 대한 연구도 수행되었다.

제 3 절 주요 이슈와 개선 방향

통계 모형 기반의 작물 단수 예측에 관한 연구들에서 제기하는 주요한 문제는 변수 선별 절차에 대한 어려움이다. 단계적(Stepwise) 회귀분석은 종속 변수를 예측하는 데 가장 효과적인 변수 집합을 찾는 방법으로 활용되어 왔으며(Hocking, 1976), 모형의 예측 효율성을 극대화하기 위해 설명변수 선택을 위한 반자동 절차가 수행되어 변수의 통계적 유의성을 파악하고 공선성 문제 등의 식별에 도움이 되어 왔다(Chatfield 1995). 그러나 동일한 데이터 세트가 모형을 공식화하고 적합성을 평가하는 데 사용되므로 생성된 결과가 편향 될 수 있다(Villamil et al., 2012). 또한 단계적 회귀분석 방법 중 Sequential Backward Selection(SBS) 변수 선택 기법은 과적합에 대해 계산적으로 효율적이고 강력하지만, 일단 제거된 다른 변수와 함께 변수 효과를 재평가 할 수 없다는 단점이 있다(Guyon & Elisseeff, 2003; De Silva & Leong, 2015; Shine et al.,2018). Whittingham et al.(2006)의 연구에서는 선택 절차와 선택 기준이 다를 때 결과가 달라져 Stepwise 회귀 방식으로 변수를 선별할 때 모형의 편향이 발생할 수 있는 가능성을 언급하였다. 또한 변수 선택은 추정 데이터 세트에 크게 의존하기 때문에 특정 지역과 특정 기간에 국한될 경우 결과가 제한될 수 있으며 실제로 강수와 같은 중요한 변수는 때때로 단계적 접근법에 의해 고려되지 않은 경우도 있다(Gornott & Wechsung, 2016). Hoffman et al.(2018)은 동일한 대상에 대해 성장계절 월평균 기온과 토양수분 데이터를 사용한 연구(Iizumi et al., 2013), 성장 계절 월평균 기온과 총강수량을 사용한 연구 (Lobell et al., 2008; Schlenker & Lobell, 2010)로 다양할 수 있음을 지적하였다. 통계 모형의 변수 선별은 여러 가지 방법으로 구성될 수 있으며, 사용 가능한 데이터에 의해 제한되고 체계적으로 선택되지만 변수의 선택은 여전히 주관적이다(Hoffman et al., 2018).

입력 변수의 선택은 모형의 예측 가능성에 상당한 영향을 미칠 수 있으며, 너무 많은 변수를 포함하면 설명력이 부족한 변수에 의해 모형 예

측 정확도가 떨어질 수 있고 반대로, 변수를 너무 적게 포함하면 정확한 예측을 할 수 있는 충분한 정보가 모형에 반영되지 않을 수 있다(Shine et al., 2018). 즉 통계 모형에서는 변수 선별이 어떤 방법과 절차를 통해 이루어지는지가 중요하다. 단순성, 효과성 및 객관성으로 인해 통계적 방법에 의한 기상 영향 평가는 오랜 전통을 가지고 있지만 이러한 평가를 최적으로 수행하는 표준화된 접근 방법은 없다. 높은 예측 기술을 갖춘 견고한 모형을 개발하려면 연구 과정에 대한 점진적 개선과 모형 확정을 위한 광범위한 공식화와 실증이 필요하다(Conradt et al., 2016).

통계 모형이 가지는 또 다른 문제는 모형 학습에 필요한 데이터의 양에 따른 제약이다. 통계 모형에서 유의한 변수 간 관계를 발견하기 위해서는 충분한 양의 데이터가 존재해야 모형 개발이 수월하다(Lobell & Asseng, 2017). 또한 충분한 수확량 및 기상 데이터가 있을 때 통계 모형의 설명력이 높아진다(Innes et al., 2015). 설명변수의 극단에서 데이터 포인트가 부족한 경우에는 한계효과를 과소평가하여 작물 단수 예측에 부정적 영향을 줄 수 있다. 데이터의 크기가 충분히 커지면 일부분 해결이 가능할 수 있으나 데이터 수집 전에 체계적인 수집 전략을 세울 필요가 있다(Jeong et al., 2016). 적절한 데이터 수집 기간에 대해서 Lobell & Burke(2009)는 통계적 방법의 회귀 모형을 구축하기 위해 작물 수확량 및 기상 관측에 대한 시계열 데이터를 20년 이상의 충분히 오랜 기간 확보하는 것을 권장하였다. Douak et al.(2013)은 기계 학습 방법이 모형을 적절하게 학습하기 위해서는 충분하고 품질 좋은 데이터의 중요성을 언급한 만큼 작물 단수 예측 모형에 기계 학습 방법을 적용하여 예측 성능을 높이기 위해서도 충분한 데이터를 확보하는 것은 중요하다.

대부분의 통계 모형에서는 월 단위 이상 기간을 집계한 변수를 사용하는데, 기상 요인의 일 단위 또는 주간 단위 패턴에 의한 생육 영향이 있을 수 있으나 월 단위 또는 성장기 계절 기상 요소 평균값을 사용하는 것은 통계 모형의 주요한 문제로 지적하였다(Semenov, 1997). 지구온난화와 기후변화로 극단적인 기상 현상이 잦아질 경우 통계 모형 오류의

문제가 확대될 수 있다(Zhang et al., 2015b). 실제로 극단적인 기상 조건(예: 급격한 온도 변화, 강풍, 우박, 범람이나 홍수로 이어지는 강수량)은 짧은 기간에 발생하지만 때로는 작물이 완전히 고사될 수 있다(예: Hawkins et al., 2013). 월 단위의 변수를 사용하면 이러한 이벤트가 부분적으로 월평균 또는 합계에 반영되어 포착이 어려워지는 문제가 발생한다(Ceglar et al., 2016). 따라서 단수에 영향을 주는 성장기의 극한 기상 현상에 대한 부정적인 영향을 줄이고 관련한 적응 조치를 적절히 계획하기 위해서는 평균화된 변수에서 탈피하여 극단적 기상에 관한 변수의 연구가 필요하다(Ceglar et al. ,2016). 한편 Chipanshi et al.(1998)의 연구에서 나타났던 극도로 건조한 해의 물부족에 대한 민감성과 습윤한 해의 해충 및 질병의 영향과 같이 작물 시뮬레이션 모형에서 포착하지 못하는 특징을 통계 모형에서는 변수로 추가하는 것이 가능하다(Hansen & Jones, 2000). 통계 모형이 작물의 생육 과정에 대한 과학적 해석을 제공하지 못한다는 Barnett(2004)의 비판에 대한 해결 방법으로 작물 시뮬레이션 모형에서 사용되는 변수의 활용이 필요하며(Qian et al., 2009), 이 들 변수를 통계 모형에 적용함으로써 기상에 의한 수확량 변동을 이해하고 예측하는 데 도움을 줄 수 있다(Liu et al., 2016). 지금까지 살펴본 것과 같이 작물 단수에 영향을 미치는 새로운 변수를 탐색하는 것은 중요한 연구 주제가 될 수 있다.

여러 연구에서 사용되고 있는 통계 모형인 선형 모형은 비선형 관계를 검출하거나 다중공선성을 갖는 인자를 식별하지 못하는 문제를 극복하기 위한 방법으로 설명변수와 종속변수 간의 비선형성 및 계층적 관계를 포착할 수 있는 기계 학습 기법을 사용할 수 있다(Shalev-Shwartz & Ben-David, 2014). 그리고 기계 학습 기법은 일반적으로 전통적인 선형 회귀 모형에 비해 예측의 관점에서 성능이 우수하기 때문에 통계 모형이 예측을 목적으로 할 경우에는 기계 학습의 도입에 의한 효과를 기대할 수 있다(Everingham et al., 2015; Feng et al. ,2018)

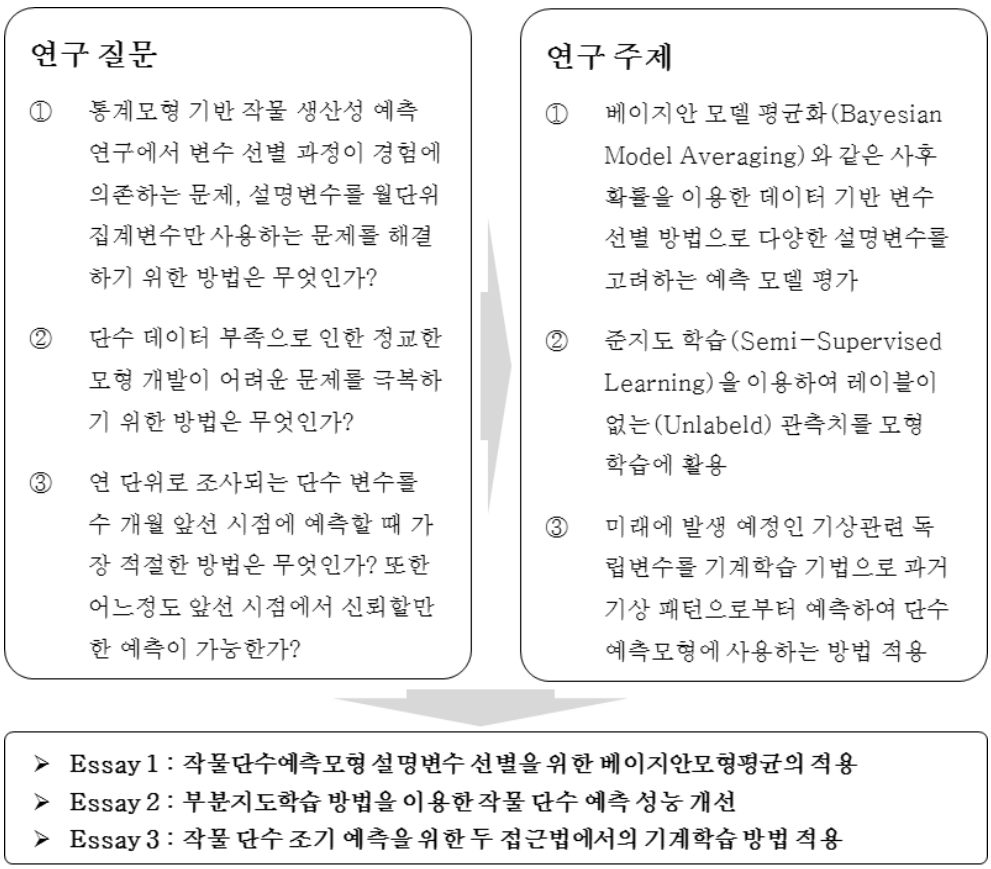
제 4 절 연구주제 도출

통계 모형 기반의 작물 단수 예측 연구에서의 주요 이슈와 개선 방향을 요약하면 첫째, 경험에 의존한 변수 선별, 단순 상관관계, 설명력에 근거한 선별이 많고, Stepwise 회귀 분석에 의한 선별은 변수가 추가되는 순서에 의해 결과가 달라지는 문제 있는 등 과학적인 방법에 기반한 다양한 변수 선별에 관한 연구가 필요하다. 또한 대부분의 통계 모형에서 설명변수로 월 집계 변수를 사용하는 한계에서 벗어나 더 짧은 기간에 대한 집계 변수에 대한 연구도 필요하다. 이러한 맥락에서 “통계모형 기반 작물 생산성 예측 연구에서 변수 선별 과정이 경험에 의존하는 문제, 설명변수를 월단위 집계변수만 사용하는 문제를 해결하기 위한 방법은 무엇인가?”라는 첫 번째 연구 질문을 수립하였다.

둘째, 작물 단수 예측 연구에서 사용되는 기상 데이터는 관측소별로 수집되고 보간법에 의해 더 조밀한 지역 단위에서도 생성이 가능하여 관측치의 수가 많으나 단수 데이터는 수집 단위가 행정구역 단위로 집계되기 때문에 관측치의 수가 많지 않은 문제가 있다. 이마저도 미국, 캐나다, 호주, 중국과 같은 넓은 농업 지역을 가지지 못한 국가에서는 수집되는 단수 데이터의 양이 절대적으로 부적하다. 이 경우 머신러닝 같은 데이터 중심 모형의 적용이 어렵고 모수적 방법에 의존해야 한다. 데이터 수가 부족한 경우는 모수 추정 과정에서 사용할 수 있는 변수의 수가 제한되기 때문에 다양한 설명변수 검토나 외부요인을 통제 변수로 사용하는 것이 어려운 문제가 있다. 이러한 점에서 “단수 데이터 부족으로 인한 정교한 모형 개발이 어려운 문제를 극복하기 위한 방법은 무엇인가?”라는 두 번째 연구 질문을 수립하였다.

셋째, 대부분의 작물 단수 예측을 위한 통계 모형 연구에서 초점을 맞추고 있는 것은 영향변수의 유의성을 검증하고 예측모형의 설명력이 개선되는지 여부이다. 새롭게 제시된 통계 모형을 이용하여 예측력을 평가하는 경우에도 통계 모형의 추정방법에 의한 예측 성능의 차이를 밝히거나 지역별로 예측치의 차이가 발생하는 편차에 대한 해석과 같이 예측

성능 평가가 실제 현장에서 필요로 하는 수확 시점 이전의 조기 예측에 대한 연구로는 이어지는 경우가 드물다. 연구가 많지 않은 상황에서 조기 예측 정확도를 개선하기 위한 다양한 방법론의 검토가 부족하다. 이러한 맥락에서 “연 단위로 조사되는 단수 변수를 수개월 앞선 시점에 예측할 때 가장 적절한 방법은 무엇인가? 또한 어느 정도 앞선 시점에서 신뢰할만한 예측이 가능한가?”라는 세 번째 연구 질문을 수립하였다.



<그림 2-1> 연구 주제 도출 과정

제시한 세 연구 질문에 대한 해답을 찾기 위한 구체적으로 방안으로 각각의 연구 질문의 연장선에서 연구 주제를 도출하였다. 첫 번째 연구 주제는 “베이지안 모형 평균화(Bayesian Model Averaging)와 같은 사후

확률을 이용한 데이터 기반 변수 선별 방법으로 다양한 설명변수를 고려하는 예측 모형 평가", 두 번째 연구 주제는 "준지도 학습 방법(Semi-Supervised Learning)을 이용하여 레이블이 없는(Unlabeld) 관측치를 모형 학습에 활용", 세 번째 연구 주제는 "미래에 발생 예정인 기상관련 독립변수를 기계 학습 기법으로 과거 기상 패턴으로부터 예측하여 단수예측모형에 사용하는 방법 적용"이다. 각각의 연구 주제는 본 논문을 구성하는 세 개의 에세이로 발전을 시켰고 각 3~5장에서 제시한 주제에 대한 해결을 위해 실측 데이터 이용한 실증 분석을 수행하였다.

제 3 장 쌀 단수 예측 모형 설명변수 선별을 위한 베이지안 모형 평균화 기법의 적용

제 1 절 서론

작물 단수 예측 분야에서 전통적으로 많이 사용되고 있는 작물 생육 과정 기반 모형(작물 모형 또는 시뮬레이션 모형으로도 명명됨)은 연구자가 조절할 수 있는 결정론적이고 확률적인 모형을 사용하여 작물 수확량과 관련한 생물학적 생육을 시뮬레이션한다. 작물 생육 시뮬레이션 모형은 개별 생육 단계별로 필요한 변수를 요구하기 때문에 종종 수집하기 어렵거나 비용이 많이 드는 외부 환경 변수를 포함하는 단점은 있으나 모든 생육 과정에서의 원인과 결과가 명확하게 설명 되는 장점을 가진다(Bornn & Zidek, 2012). 밀, 쌀, 옥수수, 콩과 같은 주요 작물에 대한 작물 시뮬레이션 모형은 오랜 기간 연구되고 개선되어 모형에 필요한 입력변수가 이미 결정되어 있고 모형의 구조에 변화를 줄 수 있는 여지는 많지 않다. 통계 기반 모형은 실증적 방법을 토대로 하기 때문에 예측의 정확도를 높일 수 있는 장점을 가진다(Barnett et al., 1997). 반면 일반적으로 작물 시뮬레이션 모형과 비교해 해석 가능성이 부족한 한계를 가진다(Barnett, 2004).

이와 같은 비판을 극복하기 위해 작물 시뮬레이션 모형에서 사용되는 변수를 도입하여 모형의 해석 가능성을 높이려는 연구들도 있다. 대표적으로 수자원 스트레스 지수를 포함시킨 연구들이 있다(Potgieter et al., 2005 , Potgieter et al., 2016 , Qian et al., 2009). 그러나 입력 변수가 정해져 있는 작물 시뮬레이션 모형과 달리 통계 기반 모형에서는 사용 가능한 변수를 이용해 실증분석을 통한 최적 모형을 찾는 과정이 필요하기 때문에 입력변수가 결정되어 있지 않다. 통계 기반 모형에서도 작물 시

뮬레이션 모형과 같이 입력 변수로 가능한 한 많은 정보를 사용하는 것이 설명 가능성을 높일 수는 있으나 수집 가능한 관측치의 수에 따라 모형의 복잡성이 제한될 수 있고, 지나치게 많은 변수를 사용할 경우 과적합의 문제가 발생할 수 있어 변수 선별을 통한 변수 수 제한이 필요하다(Mathieu & Airesa, 2018).

기존의 작물 단수 예측 관련 연구들의 변수 선별 방법에는 몇 가지 한계가 있다. 상당수의 연구는 기존 선행연구에서 사용된 독립변수를 그대로 사용하고 있다(Leng & Huang, 2017; Lobell & Burke , 2010; Peng et al. , 2018; Schauberger et al., 2018; Zhang, 2015a). 이 중 일부 연구는 예측 성능의 개선보다는 후보 변수의 영향력 평가에 초점이 맞추어져 있다(Gornott et al., 2016; Iizumi et al., 2013; Innes et al., 2015; Potgieter et al., 2016). 몇몇 연구들은 여러 변수들 중 공변량 계수 또는 회귀 모형의 설명력을 높이는 변수를 선별하는 방법을 사용하였다(Alvarez et al., 2009; Li et al., 2019; Matsumura et al., 2015; Oettli et al., 2018; Potgieter, 2005). 그러나 공변량 기준의 선별 방법은 다중공선성 문제로 인해 변수가 많을 경우 신뢰도가 떨어지는 문제가 발생할 수 있다. 또한 기존의 작물 단수 예측 연구에서는 월별 집계 변수를 주로 사용하고 있는데 15일 단위 또는 10일 단위로 집계된 변수를 사용한 연구가 거의 없는 것은 변수가 많아질 경우 효과적으로 변수를 선별하는 것이 어렵기 때문이다. 따라서 많은 변수들이 모형의 설명 변수로 고려되는 상황에서 방대한 변수 리스트로부터 유의한 변수를 선별하는 적합한 방법에 대한 연구가 필요하다.

작물 단수 예측 외의 타 분야에서는 변수 선별을 위해 기계 학습과 같은 다양한 방법들을 사용하고 있다. 농생명 분야에서는 유전자의 영향 변수 선별에서 기계 학습 방법이 가장 많이 사용되고 있다(예: Yeung et al., 2005). 생태학(예: Mutshinda et al., 2013), 수문학(예: Duan et al., 2007) 분야에서도 변수 선별에 대한 여러 연구가 있다.

본 연구에서는 기계 학습 방법 중 주어진 데이터를 이용한 사후확률 계산을 통해 데이터를 가장 잘 설명할 수 있는 모형에 대한 정보를 제공

하는 베이지안 모형 평균화 기법을 이용하여 변수를 선별하는 방법을 새로운 변수 선별 방법으로 제안한다. 이 새로운 방법의 검증을 위해 기존의 선행연구에서 제시하고 있는 영향 변수로 구성한 예측모형과 베이지안 모형 평균화 기법에 의한 예측 모형의 예측 성능을 다양한 방법을 통해 비교한다.

제 2 절 선행연구 검토

1. 통계 모형을 이용한 기존 연구의 변수 선별 방법

통계 모형 기반의 작물 단수 예측 연구에서 변수 선별에 가장 많이 사용되는 것은 공변량을 이용한 변수 선별 방법이다. 종속변수인 작물 단수 변수와 설명면수인 기상 변수 간의 상관계수가 높거나 회귀계수가 유의한 변수를 선택하는 단순한 방법을 사용한 연구(예: Matsumura et al., 2015)에서부터 회귀 모형에 설명 변수로 기상 및 토양 변수의 1차항과 2차항의 회귀계수의 유의성을 이용해 단수에 미치는 영향이 선형인지 곡선과 같은 비선형적인 관계인지를 규명한 연구도 있다(예: Alvarez et al., 2009). 공변량을 이용한 연구 중 가장 많은 유형은 설명 변수 추가에 따른 R^2의 증가 또는 평균절대오차(Mean Absolute Error: MAE), 제곱근평균오차(Root Mean Squared Error: RMSE)와 같은 모형의 예측오차의 감소를 비교하는 단계적(Stepwise) 회귀분석 방식으로 변수를 선별하는 연구이다. Stepwise 방식은 전방향(Forward)과 후방향(Backward) 방식으로 나누어진다. Forward 방식은 가장 공분산이 큰 입력변수를 1차로 선별하고 그 변수와 함께 입력 시 공분산이 가장 크게 증가하는 변수를 2차로 선별하여 R^2 가 유의하게 증가하지 않을 때까지 반복하는 방법으로 가장 중요한 입력 변수의 계층 구조를 얻기 위해 사용된다(Draper & Smith, 1998). 필요한 경우 의사 결정 을 돕기 위해 변수 간의 상관 행렬을 사용하여 다중공선성 문제 발생 시 제외할 수 있으며, 상관계수, RMSE, AIC를 비교하여 보다 엄격하게 변수를 선별할 수 있다(Mathieu & Airesa, 2018). Backward 방식은 반대로 후보 입력변수를 모두 포함하는 모형에서 출발하여 설명력이 유의하게 감소하지 않는 변수를 제외하는 방식으로 후보 변수가 많은 경우는 사용이 어렵다. 보다 발전된 방식으로 모든 후보변수를 이용해 조합이 가능한 모든 모형의 R^2 , 예측오차(RMSE)를 비교한 연구도 있다. Stepwise 방식이 단계적으로 변수의 넣고 빼는 과정으로 인해 많은 노력이 필요한 반면 모든 가능

한 모형에 대한 설명력과 예측오차를 구하는 방법은 자동화가 가능하나 변수가 많아질 경우 적용이 어려운 한계를 가진다. 실제로 Kaul et al.(2005)의 연구는 20개 변수로 120개 모형만 만들어서 인공신경망(Artificial Neural Network: ANN)을 이용한 모형은 RMSE를 비교하였고 선형회귀(OLS)를 이용한 모형은 R^2를 비교하여 변수를 선별하였다.

공변량 비교 또는 Stepwise 방법의 변수 선별에 대한 보완책으로 기계 학습 방법을 이용한 변수 선별에 관한 연구들도 있다. 랜덤포레스트(Random Forest) 기법은 의사결정나무(Decision Tree)의 엔트로피(entropy)를 기반으로 하는 설명변수가 종속변수에 미치는 상대적 중요도(Relative Importance)를 계산하여 제공하며, 앙상블(Ensemble) 학습에 의해 샘플링 편향(Bias)을 최소화하여 설명 변수의 중요도(Importance) 값을 계산한다. Folberth et al(2019)의 연구에서는 시기별 기상 변수를 이용한 파생변수를 생성하여 250개가 넘는 후보 설명변수 리스트에서 상대적 중요도 순위에 따라 입력변수를 선별하였다. 그러나 변수를 선별하는 기준이 되는 중요도 값을 결정하는 과정이 주관적이라는 한계를 보였다. Everingham et al.(2016)은 상대적 중요도에서 설명변수의 중요도 순위를 정하고 순위에 따라 Forward Stepwise 방식으로 랜덤포레스트(Random Forest)의 Out-Of-Bag(OOB) 예측오차를 순차적으로 비교하여 변수 선별하여 기존의 Stepwise 방식에 기계 학습 기법을 적용하였다. Jeong et al.(2016)은 유사한 방법으로 변수를 선별하였는데 학습 데이터셋의 오차인 OOB가 아닌 독립된 평가용 데이터셋의 제곱근평균오차(RMSE)를 비교하여 변수를 선별하였다는 점에서 차이가 있다.

이 외에도 다변량 방법을 이용한 차원 축소 및 정규화(Regularized) 회귀를 이용한 변수 선별 방법에 관한 연구도 있다. 전자는 다수의 입력변수가 존재할 때 변수를 줄이기 위한 방법으로 차원이 축소된 변수를 사용하는 방법이고 후자는 통계적 제약을 통해 유의성이 떨어지는 변수를 제외시키는 방법으로 볼 수 있다. 차원 축소 방법으로 주성분 분석(Principle Component Analysis: PCA)는 상관관계가 높은 변수들의 공변량으로 합성된 새로운 변수를 생성한다. Bornn & Zidek(2012)는 월별

집계 변수가 아닌 일별 측정치 변수로 사용하여 주성분분석으로 요인 추출하여 입력변수로 사용하였다. Sharif et al.(2017)는 주성분분석으로 추출된 변수와 라쏘(Lasso) 회귀, 능형(Ridge) 회귀로 모형에서 유의성이 떨어지는 설명 변수를 제외하는 방법을 사용하여 변수가 선별된 작물 단수 모형을 개발하였고 기존의 Stepwise 회귀분석을 이용하는 방식으로 변수를 선별한 모형과 예측 성능을 비교하였다. Ceglar et al.(2016)는 다변량 방법인 부분 최소제곱(Partial Least Square: PLS) 회귀를 이용하여 기상요소별 월별 변수를 하나의 잠재변수로 설정한 후 요인적재(Factor Loading)값으로 잠재변수 내에서의 설명변수의 영향력을 평가하였다.

본 연구에서는 변수 선별 방법으로 기존 연구에서 많이 사용되어 온 Stepwise 방법은 후보 설명변수가 많아질 경우 적용이 어려워 고려하지 않았다. 계수 추정 절차에서 벌점(Penalty)을 적용하여 영향력이 적은 변수를 제거하는 정규화 회귀(예: LASSO, Ridge, Elastic Net)로 변수 선별하는 방법은 예측오차를 평가하기 전 회귀계수 유의성을 기준으로 변수가 제외되는 점에서 고려하지 않았다. 차원 축소 방법인 주성분분석(Principle Component Analysis) 방법은 공선성 높은 변수들을 새로운 단일 변수로 축소시키기 때문에 개별 변수의 영향도를 판단하기 어렵고 모형의 설명 가능성이 줄어든다는 점 때문에 고려하지 않았다. 부분 최소제곱(Partial Least Square: PLS) 회귀 방법 역시 공선성 높은 변수를 잠재변수화하여 축약한다는 점에서 고려하지 않았다.

대신 본 연구에서는 기존의 작물 단수 예측 연구에서는 다루어지지 않은 베이지안 모형 평균화(Bayesian Model Averaging) 방법이 단수에 영향을 미치는 유의한 기상 변수를 선별하는 것에 적합한지를 검증하기 위하여 본 연구의 변수 선별 방법으로 선정하였다.

2. 베이지안 모형 평균화(Bayesian Model Averaging)를 이용한 변수 선별 연구

여러 분야의 학술지에서 각 분야의 방법론에 베이지안 모형 평균화(Bayesian Model Averaging) 기법의 적용을 위해 기존의 Stepwise 변수선별 방법과 비교를 통해 베이지안 모형 평균화 기법을 이용한 분석결과를 보고한 연구들이 있다. Wang et al.(2004)은 보건 데이터를 이용한 질병 예측에서 로지스틱 회귀를 이용한 Stepwise 방식의 변수별 방법과 예측 성능을 비교하기 위해 베이지안 모형 평균화를 적용하였다. 시뮬레이션 된 데이터 세트와 Framingham Heart Study 데이터를 사용하여 유의수준 0.05, 0.01의 로지스틱 회귀 Stepwise 변수선별에 의한 최적 모형보다 베이지안 모형 평균화에 의해 변서 선별된 모형의 예측 성능이 우수한 것을 확인하였다. Genell et al.(2010)는 베이지안 모형 평균화 기법이 의학분야 연구에 사용하는 것이 적합한지 판단하기 위해 의학 분야의 데이터 특성을 가지는 시뮬레이션 데이터를 이용해 기존 방식인 Stepwise 회귀분석 방법과 비교하여 성능이 베이지안 모형 평균화에 의한 변수 선별 방법이 우수한 조건에 대하여 보고하였다. 30 가지의 다른 효과 크기 값, 총 20 개의 설명 변수가 포함되었으며, 3개의 상관 관계 없는 변수 세트, 2 개의 상관 관계 있는 변수가 포함된 5개 데이터 세트로 실험을 하였으며, 중복변수를 선별하지 않을 확률, 올바른 모형을 선택하는 확률에서 베이지안 모형 평균화에 의한 모형의 성능이 우수하다느 것을 검증하였다. 농업분야에서는 Prost et al.(2008)가 밀의 생산성에 미치는 영향 요인에 대한 실험 연구에서 Stepwise 방식의 변수선별과 베이지안 모형 평균화에 의해 선별된 변수의 추정 계수를 비교하였다. 20개 실험 지역의 8년간의 데이터를 사용하여 모든 영향변수를 사용한 경우와 BIC·AIC 기준으로 변수선별을 한 경우, 그리고 베이지안 모형 평균화로 변수선별을 한 경우에 대하여 10,000회 Bootstraping에 의한 계수 추정결과와 변수 선택빈도를 비교하였다. 추정된 계수의 Bootstraping 분포에서 베이지안 모형 평균화가 가장 작은 표준편차로

추정 계수 값의 신뢰도가 높은 반면 BIC, AIC Stepwise 방식은 특정 변수에서는 변수 선별 없이 모든 변수를 사용한 경우보다 표준편차가 크고 불안정한 결과가 나타났다. 40개 샘플, 160개 샘플에 대한 변수선별빈도 비교에서는 40개 샘플에서 BIC, AIC Stepwise 변수 선별 방법의 경우 변수 선택 빈도가 낮아져서 과소추정되는 문제가 있는 반면 베이지안 모형 평균화 방법은 160개 샘플을 이용한 결과와 큰 차이가 없었다.

베이지안 모형 평균화를 변수선별에 직접 적용한 다른 분야의 실증 연구들로 Mutshinda et al.(2013)은 식물성 플랑크톤의 개체군에 영향을 미치는 환경 요인 연구에 베이지안 모형 평균화 기법을 적용하였다. 유사한 3가지 주요 개체군에 속하는 67가지 플랑크톤의 확산에 영향을 미치는 환경 요인을 베이지안 모형 평균화를 이용하여 각 개체별, 각 개체군에 대하여 영향도를 조사하였는데 유사한 특성을 공유하는 종속변수의 수가 많은 경우 개별 종속변수에 대한 독립변수의 영향도 차이를 반복적으로 확인해야 하는 작업에서 베이지안 모형 평균화 기법이 유용하게 활용되었다. Yeung et al.(2005)은 특정 발현과 관련한 마이크로어레이 형태의 유전자 데이터를 이용하여 영향 유전자를 선별하는 연구에 베이지안 모형 평균화를 적용하였다. 종속변수보다 독립변수의 수가 훨씬 많은 유전자 선별 연구의 특수한 상황에서 베이지안 모형 평균화에 의해 선별된 변수의 수는 기존 연구에서 사용되어 온 변수 리스트보다 줄어든 반면 예측 성능은 개선되는 것을 실증 분석을 통해 증명하였다. 사회과학 분야에서는 Jacobson & Karlsson(2004)이 스웨덴 경제지표들 중 인플레이션에 미치는 요인을 찾기 위한 방법으로 베이지안 모형 평균화 기법을 사용하였다. 86개의 분기 지표 변수를 이용한 베이지안 모형 평균화에 의한 연구 모형이 전문가 집단의 예측치 RMSE 0.99보다 개선된 0.63를 기록하였고 베이지안 모형 평균화에서 도출된 최상위 모형들에서 선택된 변수가 설명력이 높은 변수라는 점을 제시하며 기존 연구 결과를 보완할 수 있는 새로운 설명변수를 찾고 유연한 예측 모형을 개발하는 것에 도움을 줄 수 있음을 밝혔다.

제 3 절 연구방법

작물 단수 예측 모형에 관한 기존 연구들에서 모형을 정의하기 위한 변수 선별 과정에서 두 가지 한계점을 보이고 있다. 첫 째, 선행연구에서 사용된 변수들 중 설명력이 높은 변수를 선택하기 위한 방법이 단계적 (Stepwise) R^2 비교 방법에서 크게 벗어나지 않는 것이다. 둘 째, 기존의 선행연구에서 사용되어온 변수의 범위를 넘어서는 새로운 변수의 탐색과 검증에 대한 연구가 많지 않은 것이다. 이와 같은 문제를 해결하기 위한 방법으로 본 연구에서 제안하는 베이지안 모형 평균화 기법이 적합한지를 검증하기 위해서 두 가지 방향에서의 검증 절차를 수행하도록 연구를 설계하였다.

첫 번째 검증 절차는 기존의 작물 단수 예측 연구에서 사용하고 있는 변수들을 이용한 모형을 비교 모형으로 두고 선행연구에서는 검토되지 않았으나 고려 가능한 모든 변수를 대상으로 베이지안 모형 평균화 기법을 이용하여 최적의 변수를 선별한 모형의 예측 성능을 비교한다. 이 때 기존의 연구들은 유의한 설명 변수에 대한 검증을 목적으로 한 연구들이기 때문에 예측 성능을 비교하는 것이 적절치는 않으나 기존 연구들의 예측 모형에서 사용된 변수들만 고려되는 경우의 예측 성능이 얼마나 차이날 수 있는지 파악하기 위하여 기존 연구에서 사용되었던 변수만 사용하여 본 연구 설정에 맞게 재구성한 모형을 이용하여 예측 성능을 비교하였다.

두 번째 검증 절차는 기존의 연구들에서는 월 단위의 기상 변수만을 사용한 예측 모형을 연구하였다면 본 연구에서는 15일 단위, 10일 단위로 기상 변수의 집계 간격이 월 단위 대비 1/2, 1/3로 줄어든 변수를 예측 모형의 후보 변수로 고려하고 베이지안 모형 평균화 기법을 이용하여 최적의 예측 모형을 선별하였다. 이 모형은 선행연구에 기반한 비교 모형, 월 단위 변수를 이용하는 베이지안 모형 평균화 기법이 적용된 모형과 예측 성능을 비교한다. 전기, 후기로 구분되는 15일 단위 변수는 월 단위 변수에 비해 예측 모형의 후보 변수의 수가 2배가 더 많고, 상순,

중순, 하순으로 구분되는 10일 단위 변수는 월 단위 변수에 비해 예측 모형의 후보 변수의 수가 3배 더 많은 상황에서 예측 성능을 최대화할 수 있는 변수를 선별해야하기 때문에 기존 연구들에서 사용되는 단계적 (stepwise) 접근법에 의한 변수 선별은 적용이 어려운 상황으로 간주할 수 있다.

본 연구의 단수 예측 모형의 대상 작물은 주곡인 쌀로 선정하였다. 통계청에서 공표하고 있는 쌀 단수 통계는 타 작물 통계가 광역도 단위 또는 일부 주산지 단위로 집계되는 것에 반해 전국의 200여 시·군 단위로 집계되고 있어 활용 가능한 데이터의 수가 많다는 점이 선정의 첫 번째 이유이다. 두 번째 선정 이유는 쌀이 주곡이기 때문에 학계에서도 단수 예측과 관련한 연구가 가장 많이 수행되어 비교 가능한 선행 연구가 많다는 점이다.

1. 베이지안 모형 평균화(Bayesian Model Averaging) 기법

베이지안 모형 평균화 기법의 아이디어는 Leamer(1978)에 의해 시작되었으며 Raftery et al. (1997), Hoeting et al.(1999)에 의해 정립되었다. 베이지안 모형 평균화 기법은 단일 최상의 모형이 아니라 통계 모형의 전체 앙상블을 우선 고려하여 조건을 조정함으로써 연구 모형이 여러 경쟁 모형 중에서 선택될 때 발생할 수 있는 불확실성이 무시되는 문제를 극복한다(Raftery et al., 2005). 이 때 불확실성을 고려하는 방법으로 여러 모형의 사후 분포를 평균화하는 방법을 사용한다(Raftery et al., 1995; Hoeting et al.,1999).

베이지안 모형 평균화 기법에 대한 설명은 Wright(2008)의 논문을 참고하여 작성하였다. 후보 변수로 구성할 수 있는 모든 경우의 수의 모형을 $M_1,\cdots, M_n$의 n개의 모형 세트로 두고, i번째 모형의 파라메터 벡터는 Θ_i로 둔다. 이들 모형 중 하나의 모형이 참인 것을 알고 있으나 어떤 모형인지는 알지 못하는 경우, i번째 모형이 참일 가능성을 P(Mi)의 확률로 사전적 믿음으로 표현한다. 데이터 D가 관측되었을 때, 사전적 확신은 i번째 모형이 참일 가능성을 D를 이용해 사후적 확률로 계산하고 사전적 믿음을 업데이트 한다. 이 때 사후적 확률은 식 (3-1)과 같다.

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{j=1}^{n} P(D|M_i)P(M_j)} \qquad (3\text{-}1)$$

i번째 모형의 주변 우도(Marginal Likelihood)는 식 (3-2)와 같이 쓸 수 있다.

$$P(D|M_i) = \int P(D|\theta_i, M_i)\, P(\theta_i|M_i)\, d\theta_i \qquad (3\text{-}2)$$

i번째 모형의 파라메터 벡터의 사전분포는 P(Θi|Mi)이고, P(D|Θi,Mi)은 i번째 모형의 가능성(likelihood)이다. 각각의 모형은 예측을 내포한다.

모형의 불확실성의 표현에서 전반적인 예측치는 각 모형의 사후확률에 의한 예측치에 가중치가 적용되어 계산된다. 이는 최소 평균제곱오차(MSE) 예측치를 제공한다. 이제 특정 변수를 가지는 모형과 모형의 사전 확률 P(Mi), 그리고 사전 파라메터 확률 P(Θi|Mi)을 확정하는 것이 필요하다.

베이지안 모형 평균화에 사용되는 모형들은 반드시 선형 회귀 모형일 필요는 없지만 이 후 과정에서 선형 모형을 가정하면 식(3-3)으로 가정하였다.

$$y = X_i\beta_i + \varepsilon \qquad (3\text{-}3)$$

y를 예측하려는 변수의 관측치 벡터 T×1라고 두고, X_i는 T×p_i의 설명분수 매트릭스로 둔다. β_i는 p_i×1의 파라메터 벡터이고, ε은 평균이 0이고 분산이 σ^2인 오차항의 분포 벡터이다. 이 때 $\Theta_i=(\beta'_i, \sigma^2)$가 된다. 모형의 우도(likelihood) 계산을 위한 닫힌 형태의 식을 얻기 위해 설명변수들은 모두 외생적이라는 엄격한 가정을 전제로 한다.

파라메터 사전확률을 구하기 위해 σ^2이 $N(0, \sigma^2(X'_iX_i)^{-1})$이라는 조건에서의 β_i의 켤레 g사전분포를 사용한다. σ^2의 경우 사전확률이 $1/\sigma^2$에 비례하는 것을 가정하였다. 이는 요구되는 모형의 우도을 아래와 같이 분석적으로 계산할 수 있도록 해준다(Zellner, 1971).

$$p(D|M_i) = \frac{1}{2}\frac{\Gamma(T/2)}{\pi^{T/2}}(1+\varphi)^{-p_i/2}S_i^{-T} \qquad (3\text{-}4)$$

$$S_i^2 = Y'Y - Y'X_i(X_i'X_i)^{-1}X_i'Y\frac{\varphi}{1+\varphi} \qquad (3\text{-}5)$$

$$\widetilde{\beta}_i = E(\beta_i|D, M_i) = \frac{\varphi}{1+\varphi}(X_i'X_i)^{-1}X_i'Y \qquad (3\text{-}6)$$

β_i의 사전확률은 적절한 사전확률이다. 베이지안 모형 평균화에서는 모형 의존적인 파라메터의 부적절한 사전확률을 사용할 수 없다. 왜냐하면 부적절한 사전확률은 오직 임의의 유일한 상수 곱에 의해서 만들어지기 때문이며, 이것의 사용은 모형의 사후확률의 불확정성을 이끌게 된다(Kass and Raftery, 1995). 또한 β_i의 정보적 사전확률 얻기 위해서는 β_i의 사전확률분포의 중심이 0이되고 이로 인해 각 모형이 예측가능성을 결여시키는 방향으로 축소되어야 한다. 이 축소의 정도는 ϕ에 의해 조정된다. ϕ의 값이 작아지면 보다 더 많이 축소되는 것이고 이는 사전확률이 더 많은 정보를 가지도록 한다. 간혹 사전확률을 무정보적으로 두는 것도 가능하나 이는 표본외(out-of-sample) 예측의 예측성능을 떨어뜨리게 된다. ϕ의 또다른 역할은 데이터와 다른 모형들의 사후확률을 계산하여 얻어지는 사전확률 믿음의 상대적 가중치를 조정하는 것이다. 만약 ϕ=0이면 모든 모형들의 P(D|Mi)가 동일함을 의미하고 참값을 가지는 각 모형의 사후확률이 사전확률과 동일함을 의미한다. ϕ값이 커지면 커질수록 관측된 데이터에 반응하여 모형의 사전확률로부터 멀어지게 된다. 사전확률의 σ^2은 역-감마(0,0)의 사전확률분포를 가지고 이는 적절하지 않으며, 적절하지 않은 사전확률은 σ^2에 대하여 사용될 수 있다. 왜냐하면 이 파라메터는 모든 모형에 대하여 일반적이고 사전확률에서 임의의 상수곱이 다른 모형들의 사후확률에 영향을 주지 않기 때문이다.

본 연구에서는 베이지안 모형 평균화를 수행하기 위해 R 프로그래밍 언어와 R CRAN에서 제공하는 BMS 패키지(Zeugner & Feldkircher, 2015)를 이용하였다.

2. 국내 쌀 단수 예측 연구 비교

기존의 쌀 단수 예측과 관련한 선행 연구 토대의 비교 모형을 선정하기 위해 비교적 최근에 수행된 쌀 단수 예측 또는 쌀 생산성 평가에 관한 연구들 중 통계 기반 모형이 사용된 연구들을 검토하였다.

국내에서 통계 기반 단수 예측 모형이 대표적으로 사용되고 있는 곳은 농촌경제연구원의 농업부문 전망 모형인 KREI-KASMO이다. KREI-KASMO는 국내외 경제여건을 고려하여 농업부문 생산액, 부가가치, 소득, 농가인구 등의 거시지표와 농축산물의 품목별 수급 및 가격을 중장기 전망하는 것을 주요 목적으로 하고 있고 이 중 하위 영역에서 작물별 단수 예측 모형이 사용되고 있다. KREI(2011)의 쌀 단수 예측 모형은 시, 군, 도 지역별 데이터를 이용한 다년간의 패널 형태의 자료를 이용하여 고정효과 모형을 통한 패널분석으로 단수를 종속변수로 하는 기상변수의 계수를 추정하고 다양한 기상시나리오를 접목시킨 확률적 기법을 이용한 단수 예측 모형으로 개발되었다. KREI(2011)의 연구는 2001년부터 2010년까지의 10년간 단수를 종속변수로 8월 하순~9월 하순의 평균기온, 5월 상순~8월 중순의 적산온도, 8월 하순~9월 하순의 일교차, 8월 하순~9월 하순의 일조시간을 설명변수로 하는 모형을 제시하였다. 한석호(2016)는 KREI-KASMO의 쌀 단수 예측 모형이 개발된 2010년 이후의 2011~2014년 4년 간의 데이터를 이용하여 예측 성능 평가를 하였다.

권오상, 김창길(2008)은 기후변화가 쌀 단수변화에 미치는 영향을 분석하기 위해 월별 기상 데이터와 지역 더미변수를 이용하여 비모수적 기법인 커널회귀분석과 준모수적 기법인 부분선형모형을 적용한 예측모형을 연구하였다. 기상 변수로는 4~10월 6개월 간의 평균기온, 5월, 9월 평균기온, 4~10월 6개월 간의 누적강수량, 광역도 단위의 지역 더미변수를 이용하였다. 이병훈, 문한필(2012)은 CERES-Rice, ORYZA2000과 같은 작물 생육 과정을 고려한 시뮬레이션 모형의 한계를 보완하기 위한 기상 변수를 이용한 통계 기반의 단수 예측 모형을 연구하였고, 지역적

인 상관성을 고려한 공간자기회귀모형(Spatial Autoregressive Model: SAR)과 공간오차모형(Spatial Error Model: SEM)을 고려한 단수 예측 모형을 제안하였다. 기상 변수는 7~9월 누적강수량, 7~8월 평균온도, 7~9월 최고온도 및 최저온도, 7~9월 상대습도, 7~9월 풍속, 7~9월 누적일조시간, 7~9월 유효적산온도 변수를 사용하였다.

남영식 외(2012)는 RCP 8.5 기후변화 시나리오에 의한 생산성 변화를 예측하기 위한 예측 모형을 연구하였고, KREI-KASMO 쌀 단수 예측 모형과 유사한 고정효과 패널 모형을 이용하였으나 패널의 횡단면 지역 범위는 광역도 단위를 사용하였다. 예측 모형은 8~9월 평균기온, 8월 냉해(17℃이하), 8~9월 누적일교차, 7~9월 호우집중도, 7~9월 습도, 7~9월 풍속 기상 변수를 사용하였다. 이 외에 기술변화를 고려하기 위해 추세 변수의 자연로그값과 습관적/반복적 영농의 의미를 가지는 전기(T-1) 쌀 단수 변수도 독립변수로 고려하였다.

조현경 외(2013)는 기후변화에 의한 쌀 생산성 변화 예측을 위해 준모수기법인 벌칙스플라인 회귀분석 모형을 사용하여 기상 변수의 비선형성을 탐색하고 이상기후 지수를 고안하였다. 예측 모형은 5월, 9월 평균기온, 7월, 8월 누적일조시간, 폭우 더미를 기상 변수로 사용하였고, 연도 변수를 기술변화의 영향으로 사용하였다.

국내 쌀 단수 예측 연구의 모형에서 사용된 변수를 기상 요인와 기타 요인별로 정리하면 <표 3-1>와 같다.

<표 3-1> 쌀 생산성 예측에 대한 기존 연구와 선행 연구 토대의 비교 모형의 사용 변수

변수	KREI(2011)	권오상 (2008)	이병훈 (2012)	남영식 (2012)	조현경 (2013)	비교 모형
유효적산온도 (GDD)	5 월 상순 ~ 8 월 중순 적산값		7~9 월			5 월상순 ~ 8 월중순
평균기온	8 월 하순 ~ 9 월하순	4 월 ~10 월 , 5 월 , 9 월	7~8 월	8~9 월 (이차항)	5 월 , 9 월	5~10 월 각 월별 변수
최고기온			7~9 월 (이차항)			7~9 월 각 월별 변수
최저기온			7~9 월	8 월 냉해 (17℃ 이하)		7~9 월 각 월별 변수
일교차	8 월 하순 ~ 9 월하순			8~9 월 누적		8 월 하순 ~ 9 월하순
누적일조시간	8 월 하순 ~ 9 월하순		7~9 월		7 월 , 8 월 ,	5~10 월 각 월별 변수
강수량		4~10 월	7~9 월	7~9 월 호우집중도	폭우더미	5~10 월 각 월별 변수
상대습도				7~9 월		
풍속				7~9 월		
더미		도별 더미				
기술변화				O	O	
전기 단수				O		O
추정 방법	시군단위 패널 고정효과	비모수 , 준모수 방법	공간자귀 회귀모형	도단위 패널 고정효과	도단위패널 , 준모수	OLS, 비모수 , SVR,kNN

3. 베이지안모형평균화 기법과의 비교 모형

베이지안모형평균화 기법을 이용한 변수 선별 모형의 예측 성능을 평가하기 위한 비교 모형은 앞서 조사된 쌀 단수 예측 모형 선행연구에서 사용한 변수들을 종합적으로 포괄하도록 식(3-7)과 같이 모형을 구성하였다. 선행연구 중 일부는 설명변수의 이차항을 포함시켜 비선형 관계를 고려하였으나 본 연구에서는 일차항만으로 구성된 모형을 설정하였다. 대신 예측 모형의 학습 기법으로 선형회귀모형 외에 비선형성을 포착할 수 있는 일반화가법모형(Generalized Additive Model: GAM), 서포트벡터회귀(Support Vector Regression: SVR), k-최근접이웃(k-nearest neighbors)과 같은 머신러닝 기법을 이용하여 모형에 이차항이 포함되지 않더라도 비선형적 관계를 모형에 반영할 수 있도록 하였다.

$$\begin{aligned} Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 GDD5a\text{-}8b_t + \beta_3 tm_diff8c\text{-}9c_t + \beta_4 tm_avg5_t + \\ \beta_5 tm_avg6_t + \beta_6 tm_avg7_t + \beta_7 tm_avg8_t + \beta_8 tm_avg9_t + \\ \beta_9 tm_avg10_t + \beta_{10} tm_max7_t + \beta_{11} tm_max8_t + \beta_{12} tm_max9_t + \\ \beta_{13} tm_min7_t + \beta_{14} tm_min8_t + \beta_{15} tm_min9_t + \beta_{16} pr_sum5_t + \\ \beta_{17} pr_sum6_t + \beta_{18} pr_sum7_t + \beta_{19} pr_sum8_t + \beta_{20} pr_sum9_t + \\ \beta_{21} pr_sum10_t + \beta_{22} ss_dura7_t + \beta_{23} ss_dura8_t + \beta_{24} ss_dura9_t + \varepsilon \end{aligned} \quad (3\text{-}7)$$

비교 모형에서 사용된 변수의 약어로 GDD는 Growing Degree Days의 약어로 유효적산온도를 의미한다. tm_diff는 일교차, tm_avg는 평균기온, tm_min은, 최저기온, tm_max는 최고기온을 의미한다. pr_sum은 월 총강수량이고, ss_dura는 월 총일조시간을 의미한다. 변수명 끝의 숫자의 의미는 해당 월을 의미하고, a·b·c는 : 월의 상순·중순·하순을 의미한다. 비교모형에 포함된 변수 중 기상 변수는 예측하려는 단수와 같은 연도에 관측된 변수들만 고려하여서 종속변수와 동일한 t시점의 관측치이며, 전기 단수 변수는 t-1시점의 종속변수이기 때문에 Y_{t-1}로 표시하였

다. 비교 모형의 식은 선형회귀모형 추정을 고려하여 변수의 추정 계수를 포함하도록 표현하였으나 SVR, k-NN과 같은 머신러닝 기법의 경우 설명변수별로 계수를 추정하는 것이 아니기 때문에 해당 변수를 입력변수로 사용하는 모형이라는 의미로 식(3-7)을 해석해야 한다.

4. 분석 자료

본 연구에서는 사용하는 쌀 단수 자료와 기상 관측 자료는 각각 통계청과 기상청에서 제공하는 공식자료를 사용하였다. 쌀 단수 데이터는 통계청 농작물 생산조사 자료에 1965년부터 최근까지 시·군 단위와 광역도 단위로 집계되고 있으나 1977년부터 1995년까지의 19년간은 시·군 단위 단수는 집계되지 않고 광역도 단위 단수만 집계되었다. 따라서 본 연구에서는 단수 데이터의 연속성과 과거 1980년대 이전의 쌀 품종과 1990년대 이후의 쌀 품종의 차이로 인한 기상 요인의 영향 차이를 고려하여 1997년부터 2018년까지의 연도별로 집계된 157개 시·군의 단수 데이터만 사용하였다.

기상 관측 자료는 기상청의 기상자료개방포털에서 1973년 이후의 기상관측소 관측 데이터를 제공하고 있다. 제공되는 기상 요인 중 본 연구에서는 기온, 습도, 강수량, 일사량에 대한 일별 관측치 데이터를 사용하여 월 단위 최고온도, 최저온도와 같은 파생 변수를 생성하였다. 기상 데이터의 경우 쌀 단수 데이터의 시작 시점인 1997년에는 68개의 관측소가 운영되어쓰나 최근까지 꾸준히 증가하여 2018년에는 전국에 95개의 관측소가 운영되고 있다. 따라서 연도별로 관측소 수가 다른 문제와 68~95개의 관측소 관측치를 단수 데이터가 있는 전국 157개 지역을 대표하는 기상 정보가 될 수 있도록 매핑이 필요한 문제가 있다. 본 연구에서는 매년 157개 시·군 지역의 GIS 중심 좌표에 대하여 인근의 최근접한 기상 관측소 3개소의 GIS 좌표와의 거리 역수의 가중평균으로 계산된 기상 관측치를 각 지역의 기상 정보로 간주하였다. 기상 변수의 고려 범위

는 일반적인 경우 쌀의 이앙기가 5월이고 수확기가 10월~11월인 점을 고려했을 때 기상이 쌀의 생육에 직접적인 영향을 줄 수 있는 5월~10월의 기상 변수만 고려될 수 있는 점과 기존 연구들에서도 대부분 5월~10월 기상 변수만 사용되고 있다는 점에서 선행연구에 기반한 비교 모형은 5월~10월 기상 변수만 사용하여 구성하였다. 반면 베이지안 모형 평균화 기법에 의한 모형의 경우 선행연구의 고려 범위를 넘어서 1월~4월의 기상 변수도 후보 변수로 선정하였다. 이는 1월~4월 기상 조건에 의해 병해충과 관련한 요인이 영향을 받거나 1월~4월 기상 조건이 5월~10월 기상과 상관관계를 가질 수 있는 점을 고려하여 이앙전의 기상 요인도 예측 모형에 사용될 수 있도록 하였다<표 3-2>.

<표 3-2> 쌀 단수 예측 모형 사용 변수

변수	변수 설명	변수 범위
쌀 단수	단위면적(ha)당 쌀 생산량(kg)	
전기 쌀 단수	1년 전 시점의 쌀 단수	
유효적산온도(GDD)	(최고온도+최저온도)/2 - 기준온도의 누적값, 기준온도보다 낮은 경우는 0의 값을 가짐	5월~9월
평균기온	해당월의 일평균 기온의 평균값	1월~10월
최고기온	해당월의 일최고 기온의 평균값	1월~10월
최저기온	해당월의 일최저 기온의 평균값	1월~10월
일교차	해당월의 일최고-최저기온 차이의 평균값	1월~10월
누적일조시간	해당월의 일별 일조시간의 합계	1월~10월
강수량	해당월의 일별 강수량의 합계	1월~10월
상대습도	해당월의 일별 상대습도의 평균값	1월~10월

5. 예측 모형 학습과 예측 성능 평가 방법

예측모형의 학습에 고려된 기법은 선형회귀모형으로는 최소제곱법(Ordinary Least Squares: OLS)을 이용하였고, 머신러닝 기법으로는 일반화가법모형(Generalized Additive Model: GAM), 서포트벡터회귀(Support Vector Regression: SVR), k-최근접이웃(k-nearest neighbors) 기법을 사용하였다. OLS의 경우 모수적 방법론으로 최적화가 필요한 하이퍼파라메터(hyperparameter) 조정 과정이 필요 없기 때문에 학습용(Training) 데이터와 검증용(Validation) 데이터를 분리하지 않고 테스트용 데이터를 제외한 나머지 데이터 전체를 학습용 데이터로 사용하였다. 반면 나머지 GAM, SVR, k-NN의 경우 테스트용 데이터를 제외한 데이터를 다시 80:20으로 학습용 데이터와 검증용 데이터로 나누어 80%의 데이터로 예측 모형을 학습한 후 20%의 데이터로 하이퍼파라메터(hyperparameter)를 최적화하였다. 이 때 샘플링 편향(Sampling bias)를 최소화하기 위해 5-fold 교차검증(Cross Validation)을 수행하였다.

예측모형의 학습에 사용되지 않은 테스트용 데이터를 이용한 예측모형의 학습기법별 예측 성능 비교는 세 가지 방법으로 평가하였다. 첫 번째 방법은 10-fold 교차검증 방법으로 전체 기간 1997~2018년의 데이터를 먼저 랜덤 샘플링으로 10개 sub-sample로 분리한다. 10개 sub-sample은 다시 90:10의 학습·검증용:평가용 데이터로 구성되는 10개의 데이터 세트를 만든다. 10개의 데이터 세트에 대하여 90%의 데이터로 예측 모형을 최적으로 학습시킨 후 10개 세트의 10%의 테스트 데이터로 계산된 예측 성능 평가지표 평균값으로 각 학습 기법의 성능을 비교하는 방법이다.

두 번째 방법은 리브-원-아웃(Leave-one-out) 교차검증 방법으로 전체 기간 1997~2018년의 22년치의 데이터 중 1년치의 데이터는 평가용 데이터로 제외하고 나머지 21년치의 데이터로 모형의 학습과 검증을 수행하는 방법이다. 이 방법은 특정한 시점의 전체 데이터가 평가용으로 제외되기 때문에 시간적으로 독립된 데이터 데이터셋을 구성할 수 있고

10-fold 교차검증 방법에서 발생할 수 있는 관측치의 공간적 상관성에 의한 영향을 제거할 수 있는 장점을 가진다. 학습용과 검증용 데이터는 80:20으로 나누어 5-fold 교차검증을 통해 하이퍼파라메터를 최적화하였다.

세 번째 방법은 과거 시점의 데이터로 예측모형의 학습 및 검증을 한 후 최근 시점의 데이터를 테스트용으로 하여 예측 성능 평가지표의 평균값을 각 학습 기법별로 비교하는 방법이다. 학습 및 검증용 데이터는 1997~2011년 데이터를 사용하였고 2012~2018년 데이터는 평가용 데이터로 사용하였다. 2012년을 기준으로 한 것은 한석호(2016)의 연구에서 2012년의 경우 태풍의 영향으로 예측 오차가 커진 것으로 나타났는데, 본 연구에서도 예측력이 떨어질 수 있는 조건을 포함시키기 위하여 2012년이 테스트 데이터에 포함되도록 최근 7년 간의 쌀 단수 예측에 대한 예측 성능을 평가하도록 설정하였다. 이와 같이 과거 데이터로 학습한 후 최근 데이터로 평가하는 방법은 일반적으로 현업에서 모형을 학습은 과거 데이터로 하고 실제 예측은 미래 시점을 대상으로 하는 시계열 데이터 특성을 반영할 수 있다.

모형의 예측 성능 평가를 위해 사용된 지표는 평균제곱근오차(Root Mean Square Error: RMSE), 평균절대백분율오차(Mean Absolute Percentage Error: MAPE), 실측치와 예측치의 R^2(실측치(x축)와 예측치(y축) 산포도에 대한 R^2값) 3가지 지표를 사용하였다.

제 4 절 분석결과

1. 기술 통계

통계청에서 시·군 단위로 측정한 지열별, 연도별 쌀 단수 데이터를 광역도를 기준으로 집계한 평균과 표준편차는 <표 3-3>, <표 3-4>와 같다. 전체 기간에서는 충남, 전북, 경북, 충북, 경남, 전남, 경기, 강원, 제주 순으로 단수가 높았다.

<표 3-3> 연도별 광역도 단수 (1997년~2007년)

연도	강원	경기	경남	경북	전남	전북	제주	충남	충북
1997	460.4	484.1	485.9	498.3	532	523.1	435	556.6	506.3
	(31.2)	(35.5)	(14.4)	(23.3)	(18.2)	(26.6)	-	(17.8)	(23.7)
1998	432.5	468	456.3	441.1	490.9	506.2	411	510.5	475
	(39.2)	(41)	(19.8)	(24.3)	(14.5)	(12.4)	-	(29.3)	(15.3)
1999	460.6	497	449.6	471.1	471.7	524.8	417	552.2	510.3
	(27.1)	(33.5)	(19.8)	(29.5)	(23.1)	(13.1)	-	(27.4)	(22.6)
2000	467.4	464.5	453.9	478.3	490.7	515.4	424	547.5	512.4
	(21.1)	(27.2)	(20.3)	(21.7)	(24.9)	(12.1)	-	(20.6)	(20.4)
2001	488.4	490.2	504	504.7	498.7	534	453	550.7	525.4
	(26.8)	(23.1)	(18.4)	(23.6)	(22.5)	(23.1)	-	(18.3)	(16.3)
2002	418.2	469.4	431.5	469.3	435.6	483	380	508	487.6
	(80.2)	(34.9)	(24.5)	(26.4)	(32.1)	(15.1)	-	(27.7)	(18.5)
2003	419.5	428	404	405.5	438.6	439.1	380	473.6	429.2
	(27)	(28.6)	(18.2)	(32.4)	(16.6)	(32.5)	-	(31.5)	(33.1)
2004	459.3	476.2	490.7	494.9	487.1	528.9	442	531.1	496.5
	(18.7)	(27.8)	(18.7)	(18.6)	(15.5)	(22.8)	-	(30.3)	(19.4)
2005	469	475	473.1	487	472.6	493.2	409	517.1	484.3
	(22.8)	(27.7)	(20.3)	(23.4)	(19.3)	(24.6)	-	(26.8)	(12.8)
2006	442.2	466.7	480.6	487.3	462.6	508	421.5	525.4	480.7
	(28.1)	(31.4)	(21.2)	(22.6)	(17.9)	(25.3)	-	(23.1)	(27.2)
2007	437.1	451.2	461.9	469.4	427.1	474.4	381.7	495.2	461.6
	(20.9)	(19.9)	(17.7)	(20.3)	(18.2)	(27.8)	-	(21.8)	(12.5)

1997~2007년 11년에 비해 2008~2018년 11년에 단수가 증가한 지역은 증가량을 기준으로 경북, 경남, 강원, 전북, 충북, 충남, 전남 순이었고 제주는 유일하게 감소 추세를 보였다. 연도별로는 전국 평균 단수가 가장 높았던 해는 2015년 527.9kg/ha, 2016년 521.5kg/ha, 2009년 519.7kg/ha 순이었고, 가장 낮았던 해는 2003년 424.2kg/ha, 2007년 451.1kg/ha, 2002년 453.6kg/ha 순으로 태풍의 영향이 컸던 해이다. 가장 높은 해와 가장 낮은 해는 100kg/ha 가량의 차이를 보였다.

<표 3-4> 연도별 광역도 단수 (2008년~2018년)

연도	강원	경기	경남	경북	전남	전북	제주	충남	충북
2008	477.8	499.9	521.7	520.9	478.3	537.2	465.2	546.6	510.2
	(21.3	(24.6)	(19.8)	(25.9)	(31.8)	(21.9)	-	(20.8)	(10.9)
2009	502.5	492.8	518.4	543.2	491.1	568	471.6	563.9	525.7
	(25.6	(21.4)	(23.6)	(31.1)	(19.4)	(17.7)	-	(24.9)	(14.7)
2010	450.6	425.8	482.6	500	461.5	508.2	341.4	503.2	476.3
	(20.1	(19)	(12.5)	(15.7)	(26.4)	(22.7)	-	(32.5)	(17.9)
2011	469.9	428.2	493.3	509.6	475.7	513.4	415.3	521.4	487.6
	(21.5	(21.6)	(11.3)	(15.5)	(30.9)	(23.2)	-	(22.1)	(16.3)
2012	460.7	460.9	481.3	498.6	413.4	481.2	336.5	510.1	482.7
	(37.3	(12.6)	(14.6)	(15.1)	(51.7)	(29.1)	-	(27.6)	(20.4)
2013	463.3	448.6	504.4	520.9	484.8	529.4	374.2	535.3	507.5
	(28.9	(29.5)	(16.7)	(18.2)	(24.8)	(17.1)	-	(24.7)	(16.6)
2014	514.2	488.5	493.2	528.2	479	534.6	393.2	550.8	523.9
	(25.6	(13.3)	(17.5)	(25.2)	(27.1)	(24)	-	(32.7)	(16.4)
2015	529.7	505.8	525.1	553.8	512.6	564.4	464.6	560.9	533.8
	(19.4	(17.6)	(16.7)	(25)	(21.3)	(20.2)	-	(22.1)	(15.6)
2016	529.6	504.2	520.9	550.6	511.8	566.4	418.7	558.2	533.4
	(28.1	(15.7)	(19.7)	(28.5)	(21.3)	(16.4)	-	(20.6)	(18.6)
2017	493.1	484.1	528.3	547.6	512.4	549.9	395.7	533.6	506.7
	(29.9	(17.1)	(11)	(23.4)	(28.1)	(12.1)	-	(19.9)	(13.9)
2018	511.8	492.6	507.3	538.8	495	541.9	348.5	542.7	521.8
	(37.3	(18.6)	(14.5)	(25.6)	(21.9)	(19.2)	-	(19.6)	(23.1)

시·군별 월 평균 기온 변수를 연도별로 집계한 평균과 표준편차는 <표 3-5>, <표 3-6>와 같다. 1997~2007년 11년에 비해 2008~2018년 11년에 기온이 증가한 시기는 7월 0.9℃, 5월과 8월 0.5℃, 6월 0.4℃로 여름철의 기온이 증가하였고 기온이 감소한 시기는 1월 -0.8℃, 2월과 4월 -0.4℃로 여름은 더워지고 겨울은 추워지는 경향을 보였다.

<표 3-5> 연도별 월 평균 기온(1997년~2007년)

연도	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월
1997	-1.9	1.2	6.8	12.5	17.3	22.4	25.1	25.7	20.0	13.8
	(2.3)	(1.9)	(1.4)	(1.0)	(1.2)	(1.1)	(0.9)	(1.1)	(1.6)	(1.7)
1998	-0.3	3.5	7.4	15.1	18.4	20.7	24.6	25.3	22.3	16.6
	(2.0)	(1.8)	(1.2)	(1.0)	(0.9)	(1.2)	(1.4)	(1.3)	(1.3)	(1.4)
1999	0.2	1.6	6.6	13.1	17.3	22.2	24.4	25.0	22.6	14.2
	(1.9)	(1.7)	(1.4)	(1.1)	(1.0)	(0.9)	(1.2)	(1.2)	(1.2)	(1.5)
2000	-0.4	-0.6	6.3	11.8	17.4	22.2	26.0	25.7	19.9	14.7
	(2.0)	(1.8)	(1.2)	(1.1)	(1.1)	(1.1)	(0.9)	(1.2)	(1.3)	(1.6)
2001	-2.2	0.9	5.6	13.1	18.7	22.1	25.9	25.8	21.3	15.8
	(2.6)	(2.1)	(1.7)	(1.1)	(0.9)	(0.9)	(0.8)	(1.1)	(1.4)	(1.3)
2002	1.2	2.5	8.0	13.7	17.3	21.7	24.9	24.1	20.6	13.1
	(2.1)	(1.5)	(1.1)	(0.8)	(0.9)	(0.9)	(1.0)	(1.0)	(1.3)	(1.6)
2003	-1.9	2.5	6.2	12.8	18.1	21.2	22.7	23.9	21.1	13.7
	(2.0)	(1.6)	(1.3)	(1.0)	(1.2)	(0.9)	(1.2)	(1.0)	(1.4)	(1.5)
2004	-1.0	2.9	6.5	13.1	17.6	22.2	25.4	25.4)	21.1	14.5
	(1.7)	(1.6)	(1.2)	(0.9)	(0.9)	(1.0)	(1.2)	(1.3)	(1.2)	(1.4)
2005	-1.3	-0.7	4.7	13.4	17.4	22.7	25.2	25.2	21.9	14.5
	(1.7)	(1.7)	(1.3)	(1.1)	(1.1)	(1.0)	(0.9)	(1.0)	(1.6)	(1.5)
2006	0.5	1.0	5.9	11.4	17.7	21.7	23.4	26.8	19.9	17.2
	(1.9)	(1.6)	(1.3)	(1.1)	(1.1)	(1.2)	(1.3)	(1.1)	(1.4)	(1.4)
2007	0.9	4.2	6.9	11.7	18.0	22.0	23.8	26.3	21.4	15.2
	(1.9)	(1.6)	(1.5)	(1.1)	(1.0)	(1.1)	(1.1)	(1.0)	(1.1)	(1.5)

<표 3-6> 연도별 월 평균 기온(2008년~2018년)

연도	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월
2008	-0.3	-0.4	7.3	13.4	17.7	20.8	26.1	24.9	21.8	16.1
	(2.2)	(1.8)	(1.3)	(1.0)	(0.9)	(1.2)	(1.1)	(1.1)	(1.3)	(1.4)
2009	-1.1	4.0	6.9	12.7	18.5	22.1	23.8	24.8	21.2	15.6
	(1.9)	(1.7)	(1.3)	(1.1)	(0.8)	(0.9)	(0.9)	(1.1)	(1.2)	(1.4)
2010	-2.1	2.4	5.4	9.9	17.3	22.7	25.6	27.0	21.9	14.6
	(2.4)	(1.7)	(1.5)	(1.0)	(0.9)	(1.0)	(0.7)	(0.9)	(1.3)	(1.3)
2011	-5.0	1.8	4.5	11.2	17.5	22.1	25.2	25.3	21.2	13.9
	(2.2)	(1.5)	(1.2)	(1.1)	(1.0)	(1.0)	(1.1)	(0.8)	(1.3)	(1.4)
2012	-1.3	-0.9	5.6	12.7	18.7	22.5	25.6	26.6	20.3	14.5
	(2.0)	(1.7)	(1.3)	(1.0)	(1.1)	(1.3)	(0.8)	(1.1)	(1.1)	(1.4)
2013	-2.2	0.5	6.4	10.2	18.0	22.9	26.3	27.4	21.2	15.4
	(2.2)	(1.8)	(1.6)	(1.1)	(0.9)	(1.1)	(1.0)	(0.8)	(1.2)	(1.4)
2014	0.3	2.4	7.7	13.4	18.5	22.0	25.2	23.9	21.1	14.9
	(1.6)	(1.6)	(0.9)	(0.8)	(0.8)	(1.1)	(0.7)	(0.9)	(1.1)	(1.3)
2015	0.3	1.9	6.8	12.9	18.7	22.0	24.6	25.4	20.7	15.1
	(1.6)	(1.3)	(1.0)	(0.9)	(0.9)	(1.2)	(1.0)	(0.9)	(1.3)	(1.2)
2016	-1.2	1.6	7.2	14.0	18.8	22.5	25.6	26.8	21.8	15.8
	(1.8)	(1.6)	(1.0)	(0.8)	(0.7)	(0.9)	(1.0)	(1.1)	(1.2)	(1.2)
2017	-0.1	1.4	6.2	13.8	18.8	22.0	26.5	25.5	20.7	15.4
	(1.9)	(1.7)	(1.1)	(0.8)	(0.7)	(1.0)	(0.8)	(1.2)	(1.2)	(1.3)
2018	-2.3	-0.3	8.1	13.3	18.0	22.4	27.0	27.5	20.6	13.1
	(1.9)	(1.8)	(0.9)	(0.9)	(0.9)	(0.8)	(0.9)	(1.2)	(1.2)	(1.5)

시·군별 월 누적 강수량 변수를 연도별로 집계한 평균과 표준편차는 <표 3-7>, <표 3-8>와 같다. 강수량은 7, 8월에 집중되어 있고 다음으로 6, 9월이 비슷한 수준이다. 1997~2007년 11년에 비해 2008~2018년

11년에 강수량이 증가한 시기는 10월 22.9mm, 3월 11.2mm, 4월 5.1mm로 봄, 가을철에 소폭 증가에 그쳤으나 강수량이 감소한 시기는 8월 -87.8mm, 9월 -69.6mm, 6월 -51.4mm, 5월 -23.6mm, 7월 -17.9mm로 5~9월 강수량은 큰 폭으로 감소하였다.

<표 3-7> 연도별 월 평균 강수량(1997년~2007년)

연도	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월
1997	18.2	33.1	40.0	65.2	188.3	173.5	315.6	239.3	53.2	14.4
	(12.7)	(15.5)	(18)	(18.0)	(71.4)	(50.2)	(71.8)	(73.1)	(52.0)	(13.0)
1998	41.6	44.6	47.3	146.7	108.5	271.3	273.2	510.0	242.5	59.5
	(35.7)	(16.8)	(20.0)	(33.7)	(27.6)	(75.5)	(110.2)	(213.9)	(87.8)	(33.0)
1999	16.7	17.5	93.1	83.0	131.9	154.9	251.3	376.8	350.9	107.8
	(15.8)	(13.8)	(35.8)	(22.1)	(36.3)	(45.1)	(117.9)	(126.9)	(63.8)	(34.4)
2000	34.8	4.3	19.1	37.5	53.9	167.3	258.9	409.3	251.6	27.3
	(13.7)	(4.5)	(11.0)	(15.7)	(14.0)	(86.4)	(91.3)	(121.9)	(75.4)	(10.1)
2001	52.5	64.2	16.6	25.3	30.3	226.0	294.8	120.2	89.4	93.7
	(19.6)	(19.3)	(7.4)	(13.0)	(15.8)	(49.1)	(163.2)	(51.0)	(73.8)	(40.0)
2002	62.2	6.9	47.5	147.1	103.3	67.7	218.2	586.9	79.9	63.9
	(19.9)	(5.2)	(24.1)	(55.0)	(29.7)	(23.0)	(55.5)	(188.6)	(30.6)	(38.5)
2003	26.7	48.7	51.1	194.4	167.7	176.2	468.5	372.3	262.4	30.4
	(13.3)	(14.0)	(21.3)	(36.4)	(76.8)	(51.4)	(105.8)	(119.3)	(110.2)	(15.9)
2004	13.0	43.1	30.2	84.5	120.4	225.4	339.9	323.7	195.1	6.1
	(8.0)	(17.3)	(14.4)	(35.6)	(34.5)	(74.6)	(110.2)	(122.6)	(38.0)	(4.5)
2005	11.8	34.0	55.4	69.9	69.6	157.7	288.8	299.0	195.3	36.5
	(9.6)	19.3)	(30.2)	(23.7)	(25.3)	(34.6)	(82.3)	(72.7)	(130.4)	(33.8)
2006	27.0	(29	13.6	90.0	162.4	151.5	654.3	126.0	52.0	44.4
	(8.3)	(10.5)	(6.1)	(29.4)	(58.8)	(62.4)	(182.9)	(62.9)	(38.3)	(61.1)
2007	10.4	42.6	103.0	33.1	110.4	83.2	239.2	308.4	381.3	54.5
	(12.7)	(25.3)	(17.9)	(9.9)	(27.3)	(34.3)	(48.1)	(100.3)	(104.8)	(33.0)

<표 3-8> 연도별 월 평균 강수량(2008년~2018년)

연도	1월	2월	3월	4월	5월	6월	7월	8월	9월	10월
2008	39.4	8.6	47.3	49.3	103.1	195.8	261.7	185.8	65.6	30.5
	(23.3)	(4.6)	(16.2)	(12.2)	(34.6)	(71.1)	(170.1)	(61.0)	(27.8)	(11.3)
2009	14.7	38.6	54.3	51.0	111.6	129.2	533.6	176.4	48.0	43.5
	(9.7)	(18.0)	(12.5)	(27.4)	(19.4)	(40.4)	(118.8)	(73)	(15.8)	(18.8)
2010	32.3	84.7	89.9	88.2	123.8	65.7	249.1	416.1	283.8	38.7
	(8.4)	(31.0)	(25.8)	(43.5)	(39.6)	(43.4)	(88.0)	(125.4)	(154.4)	(16.8)
2011	6.4	57.7	22.0	112.8	119.9	285.3	546.1	280.9	66.9	50.9
	(5.1)	(23.0)	(5.9)	(24.5)	(36.3)	(99.1)	(265.0)	(99.0)	(57.7)	(27.5)
2012	14.9	9.9	82.2	121.0	34.4	75.1	314.8	408.4	244.7	59.8
	(9.1)	(14.1)	(35.5)	(40.2)	(14.2)	(17.8)	(85.4)	(87.6)	(77.4)	(22.8)
2013	26.2	50.3	58.4	71.9	121.8	102.8	342.4	164.9	121.4	43.6
	(10.6)	(16.0)	(18.2)	(13.1)	(47.0)	(49.0)	(190.8)	(59.3)	(37.8)	(37.0)
2014	11.0	25.5	68.4	76.0	56.5	77.4	170.3	352	112.1	108.5
	(9.1)	(35.2)	(40.9)	(34.2)	(30.8)	(24.8)	(54.7)	(142.2)	(27.9)	(29.1)
2015	25.6	23.2	36.9	118.8	51.0	98.2	171.5	109.1	50.2	69.2
	(11.2)	(6.5)	(20.2)	(45.1)	(38.4)	(30.6)	(39.4)	(54.0)	(38.0)	(22.0)
2016	22.8	46.2	55.1	147.7	103.4	71.8	294.6	73.5	181.9	133.2
	(20.6)	(14.4)	(24.2)	(61.2)	(36.7)	(33.0)	(101.9)	(54.4)	(127.1)	(52.3)
2017	14.5	29.2	21.3	64.6	28.8	48.3	349.9	246.8	82.9	56.7
	(8.4)	(12.5)	(9.0)	(22.7)	(10.0)	(19.5)	(184.5)	(73.3)	(41.5)	(32.2)
2018	20.9	32.4	104.0	131.8	132.5	139.4	171.8	292.4	130.7	156.0
	(13.0)	(8.2)	(41.9)	(13.4)	(45.4)	(56.6)	(49.8)	(80.8)	(50.5)	(41.2)

2. 비교 모형의 선형회귀 분석 결과

선행연구에 기초한 비교 모형의 선형회귀 분석 결과는 <표 3-9>와 같다.

<표 3-9> 비교 모형의 선형회귀모형 계수 추정 결과

변수	계수	표준오차	t-value	Pr(>|t|)		VIF
절편	225.93	15.61	14.47	0.000	***	
Yt-1	0.49	0.01	38.70	0.000	***	1.33
GDD5a8bt	0.04	0.01	4.49	0.000	***	5.91
tm_dif8c9ct	-0.12	1.05	-0.11	0.909		8.36
tm_avg5t	-5.65	0.78	-7.21	0.000	***	3.30
tm_avg6t	4.53	0.83	5.46	0.000	***	4.17
tm_avg7t	5.27	6.01	0.88	0.381		330.10
tm_avg8t	13.97	5.48	2.55	0.011	*	282.92
tm_avg9t	-0.45	4.52	-0.10	0.921		194.09
tm_avg10t	3.17	0.59	5.35	0.000	***	4.83
tm_max7t	1.76	2.90	0.61	0.543		118.26
tm_max8t	0.55	2.59	0.21	0.833		93.25
tm_max9t	-3.45	2.20	-1.57	0.117		46.22
tm_min7t	-3.67	3.40	-1.08	0.281		94.07
tm_min8t	-25.19	3.22	-7.83	0.000	***	98.62
tm_min9t	5.12	2.70	1.90	0.058	.	107.12
pr_sum5t	-0.02	0.01	-2.03	0.042	*	1.63
pr_sum6t	0.04	0.01	5.31	0.000	***	1.48
pr_sum7t	-0.02	0.00	-4.53	0.000	***	2.40
pr_sum8t	-0.04	0.00	-9.58	0.000	***	1.87
pr_sum9t	-0.04	0.01	-7.47	0.000	***	2.13
pr_sum10t	-0.03	0.01	-2.49	0.013	*	1.52
ss_dura7t	0.85	0.84	1.01	0.311		6.35
ss_dura8t	1.83	0.87	2.09	0.036	*	8.07
ss_dura9t	4.34	0.94	4.64	0.000	***	5.69
adjusted R^2	0.596					

예측성능 평가에 앞서 선행연구에서 사용된 변수들이 실제 관측 데이터를 이용한 변수 추정 결과에서 어느 정도 유의한지 살펴보기 위해 선형회귀 분석을 실시하였다. 추정 결과에서는 8월하순~9월하순 일교차, 7월평균기온, 9월 평균기온, 7월 최고기온, 8월 최고기온, 9월 최고기온, 7월 최저기온, 7월 누적일조시간 변수가 유의하지 않았다. 그러나 7, 8, 9월 기온 관련 변수들은 평균기온, 최고기온, 최저기온 간 상관관계가 높아 다중공선성에 의해 t-value가 과소추정된 것으로 판단된다. VIF 값을 참고하여 다중공선성이 의심되는 7, 8, 9월 온도관련 변수 중 상관계수가 높은 변수들을 제외한 모형에 대한 추정 결과는 <표 3-10>와 같다. 극단적으로 높은 VIF 값을 가진 변수가 제외된 후 대부분의 변수들이 유의해진 것을 확인할 수 있다. 그러나 조정된 R^2는 0.596에서 0.550으로 낮아져서 모형의 설명력은 떨어진 것을 확인할 수 있다.

<표 3-10> 공선성 변수를 제외한 비교 모형의 선형회귀모형 계수 추정 결과

변수	계수	표준오차	t-value	Pr(>|t|)		VIF
(Intercept)	216.04	15.57	13.87	0.000	***	
yield_l1	0.46	0.01	35.48	0.000	***	1.27
GDD5a8b	0.02	0.01	1.74	0.081	.	5.25
tm_dif8c9c	7.06	0.68	10.35	0.000	***	3.15
tm_avg5	-2.52	0.78	-3.24	0.001	**	2.93
tm_avg6	1.86	0.80	2.33	0.020	*	3.49
tm_avg10	-1.36	0.55	-2.45	0.014	*	3.79
tm_min7	-2.02	0.70	-2.86	0.004	**	3.61
tm_min9	2.25	0.58	3.90	0.000	***	4.40
pr_sum5	-0.09	0.01	-8.72	0.000	***	1.35
pr_sum6	0.04	0.01	6.22	0.000	***	1.44
pr_sum7	-0.03	0.00	-6.97	0.000	***	2.19
pr_sum8	-0.04	0.00	-11.02	0.000	***	1.77
pr_sum9	-0.05	0.01	-9.45	0.000	***	1.98
pr_sum10	-0.03	0.01	-2.74	0.006	**	1.46
ss_dura7	2.88	0.66	4.37	0.000	***	3.53
ss_dura8	4.93	0.56	8.86	0.000	***	2.95
ss_dura9	-1.95	0.60	-3.26	0.001	***	2.07
adjusted R^2	0.550					

본 연구에서 모형의 역할은 예측을 위한 것으로 <표 3-9>, <표 3-10>에서와 같이 변수의 유의성을 고려하기 보다는 모형의 예측 성능을 높이는 위하여 모형의 설명력을 높일 수 있는 변수는 모형에 포함시켜서 예측 모형을 구축하는 것을 기본으로 한다. 따라서 이후의 베이지안 모형 평균화 기법에 의한 최적 모형에서도 설명변수 간의 다중공선성 문제는 중요하게 다루지 않는다.

3. 베이지안 모형 평균화 기법에 의한 최적 모형

<표 3-2>에서 소개된 모형의 후보 변수를 이용하여 베이지안 모형 평균화 기법의 분석 수행 결과는 <부록 2>과 같다. PIP는 사후포함확률(posterior inclusion probabilities)를 의미하고, Post Mean은 베이지안 모형 평균화에서 사용된 모든 모형(이 때 변수가 포함되지 않은 모형에서는 해당 변수의 계수가 0이라고 봄)에서 추정된 계수값의 평균이다. Post SD는 Post Mean에서 평균값을 구한 계수의 표준편차를 의미한다. Cond.Pos.Sign은 변수가 모형에 포함되는 조건에서의 계수의 부호가 양수일 사후확률을 의미한다.

베이지안 모형 평균화의 분석 결과에는 사후모형확률(posterior model probability: PMP)에 의해 가장 예측 성능이 우수한 모형을 선별할 수 있다. 그러나 이 경우 분석에 사용되는 데이터에 의한 편향이 발생할 수 있기 때문에 본 연구에서는 사후포함확률(PIP)를 기준으로 50% 이상의 포함확률을 가지는 변수를 선정하여 모형을 구성하였다.

월 단위 기상변수를 후보변수로 사용하여 베이지안 모형 평균화를 수행한 결과 도출된 최적 예측 모형은 식(3-8)과 같다. 선행연구에 기초한 비교 모형에서는 사용되지 않았던 3, 4월 평균기온, 4, 5월 최고기온, 4, 5월 최저기온, 3, 6, 10월 상대습도, 5, 10월 누적일조시간, 1월 강수량 변수가 포함되었다. 또한 비교 모형에서는 포함되었던 5~7월 평균기온, 8, 9월 최고기온, 7, 9월 최저기온, 7~9월 누적일조시간, 5, 10월 강수량 변수는 포함되지 않는 차이를 보였다. 월 단위 변수에서 선정된 최적 모형에 대한 선형회귀분석 결과는 <표 3-11>와 같다. 모든 변수가 유의하였고 조정된 R^2는 0.6469로 비교 모형 대비 약 0.05 가량 높게 나타났다.

$$\begin{aligned} Yt = {} & Yt\text{-}1 + tm_avg4t + tm_avg8t + tm_avg9t + tm_avg10t + \\ & tm_max4 + tm_max5 + tm_max7 + tm_min4t + tm_min5t + \\ & tm_min8t + hm_avg3t + hm_avg6t + hm_avg10 + ss_dura5t + \\ & ss_dura10t + pr_sum1t + pr_sum6t + pr_sum7t + pr_sum8t + \\ & pr_sum9t \end{aligned} \quad (3\text{-}8)$$

15일 단위 기상변수를 후보변수로 사용하여 베이지안 모형 평균화를 수행한 결과 도출된 최적 예측 모형은 식(3-9)과 같다. 선정된 최적 모형에 대한 선형회귀분석 결과는 <표 3-12>와 같이 9월 하반기 누적일조시간을 제외한 모든 변수가 유의하였고 조정된 R^2는 0.6929로 비교 모형 대비 약 0.1, 월 단위 최적 모형 대비 약 0.5 가량 높아졌다.

$$
\begin{aligned}
Yt = {} & Yt\text{-}1 + tm_avg1at + tm_avg2bt + tm_avg3at + tm_avg4bt + \\
& tm_avg5at + tm_avg7at + tm_avg7bt + tm_avg10at + \\
& tm_max1at + tm_max2bt + tm_max3at + tm_max3bt + \\
& tm_max7bt + tm_max8at + tm_min1at + tm_min3at + \\
& tm_min3bt + tm_min5at + tm_min8at + tm_min8bt + \\
& tm_min9at + tm_min9bt + tm_min10bt + hm_avg2at + \\
& hm_avg6at + hm_avg10at + hm_avg10bt + pr_sum1bt + \\
& pr_sum3at + pr_sum4at + pr_sum6at + pr_sum7bt + \\
& pr_sum8at + pr_sum10at + ss_dura1at + ss_dura5at + \\
& ss_dura6bt + ss_dura8bt + ss_dura9bt \qquad (3\text{-}9)
\end{aligned}
$$

10일 단위 기상변수를 후보변수로 사용하여 베이지안 모형 평균화를 수행한 결과 도출된 최적 예측 모형은 식(3-10)과 같다. 선정된 최적 모형에 대한 선형회귀분석 결과는 <표 3-12>와 같이 6월 하순 최고온도를 제외한 모든 변수가 유의하였고 조정된 R^2는 0.7032로 15일 단위 최적 모형에서 크게 높아지지는 않았다.

$$
\begin{aligned}
Yt = {} & Yt\text{-}1 + tm_avg5bt + tm_avg6bt + tm_avg10bt + tm_avg10ct + \\
& tm_max2at + tm_max2ct + tm_max3bt + tm_max6ct + \\
& tm_max8bt + tm_min2bt + tm_min3at + tm_min5bt + \\
& tm_min8ct + tm_min9bt + tm_min10bt + hm_avg1bt + \\
& hm_avg1ct + hm_avg3bt + hm_avg6bt + hm_avg10bt + \\
& pr_sum1at + pr_sum1ct + pr_sum3ct + pr_sum4ct + \\
& pr_sum7at + pr_sum7bt + pr_sum8ct + ss_dura3ct +
\end{aligned}
$$

ss_dura4bt + ss_dura5bt + ss_dura6bt + ss_dura6ct + ss_dura9bt (3-10)

<표 3-11> 월 단위 BMA 모형의 선형회귀모형 계수 추정 결과

변수	계수	표준오차	t-value	Pr(>|t|)	VIF
절편	190.23	16.09	11.82	0.000 ***	
$Yield_{t-1}$	0.52	0.01	43.37	0.000 ***	1.35
tm_avg4	-28.37	2.98	-9.51	0.000 ***	109.73
tm_avg8	20.14	1.38	14.55	0.000 ***	20.69
tm_avg9	2.86	0.58	4.93	0.000 ***	3.65
tm_avg10	4.69	0.51	9.23	0.000 ***	4.06
tm_max4	15.02	1.67	8.97	0.000 ***	52.30
tm_max5	-9.90	0.59	-16.87	0.000 ***	4.38
tm_max7	6.94	0.50	13.97	0.000 ***	3.98
tm_min4	8.77	1.40	6.26	0.000 ***	36.07
tm_min5	6.64	0.81	8.25	0.000 ***	5.68
tm_min8	-30.14	1.52	-19.81	0.000 ***	25.21
hm_avg3	-0.51	0.09	-5.39	0.000 ***	2.08
hm_avg6	-0.58	0.13	-4.61	0.000 ***	2.69
hm_avg10	0.89	0.14	6.44	0.000 ***	3.33
ss_dura5	5.04	0.58	8.69	0.000 ***	2.67
ss_dura10	5.28	0.59	8.94	0.000 ***	1.98
pr_sum1	-0.27	0.03	-10.69	0.000 ***	1.39
pr_sum6	0.04	0.01	5.75	0.000 ***	1.81
pr_sum7	-0.03	0.00	-8.01	0.000 ***	2.03
pr_sum8	-0.03	0.00	-8.65	0.000 ***	1.92
pr_sum9	-0.03	0.00	-6.54	0.000 ***	1.63
adjusted R^2	0.6469				

<표 3-12> 15일 단위 BMA 모형의 선형회귀모형 계수 추정 결과

변수	계수	표준오차	t-value	Pr(>|t|)		VIF
(Intercept)	269.68	16.92	15.94	0.000	***	
$Yield_{t-1}$	0.55	0.01	45.54	0.000	***	1.58
tm_avg1a	28.37	3.30	8.61	0.000	***	420.82
tm_avg2b	7.34	0.89	8.29	0.000	***	30.41
tm_avg3a	-17.24	2.83	-6.10	0.000	***	154.44
tm_avg4b	-3.23	0.55	-5.83	0.000	***	4.79
tm_avg5a	-10.50	1.23	-8.51	0.000	***	14.82
tm_avg7a	5.72	0.60	9.59	0.000	***	5.97
tm_avg7b	-9.65	1.89	-5.10	0.000	***	47.89
tm_avg10a	7.90	0.67	11.85	0.000	***	9.58
tm_max1a	-10.75	1.58	-6.79	0.000	***	105.32
tm_max2b	-7.03	0.81	-8.64	0.000	***	29.94
tm_max3a	6.34	1.49	4.25	0.000	***	64.82
tm_max3b	5.44	0.54	10.16	0.000	***	9.49
tm_max7b	7.77	1.40	5.54	0.000	***	39.59
tm_max8a	3.14	0.72	4.35	0.000	***	9.84
tm_min1a	-16.76	1.79	-9.37	0.000	***	154.08
tm_min3a	9.01	1.43	6.28	0.000	***	45.54
tm_min3b	-4.65	0.58	-8.02	0.000	***	8.15
tm_min5a	9.40	1.24	7.58	0.000	***	16.64
tm_min8a	-9.06	0.89	-10.20	0.000	***	11.48
tm_min8b	-3.46	0.85	-4.07	0.000	***	10.12
tm_min9a	5.18	0.61	8.55	0.000	***	8.97
tm_min9b	1.00	0.45	2.23	0.026	*	4.35
tm_min10b	-7.08	0.62	-11.37	0.000	***	11.76
hm_avg2a	-0.18	0.09	-2.06	0.039	*	2.83
hm_avg6a	-1.22	0.12	-9.84	0.000	***	3.00
hm_avg10a	1.48	0.14	10.36	0.000	***	4.16
hm_avg10b	-0.66	0.15	-4.41	0.000	***	5.27
pr_sum1b	-0.23	0.05	-4.46	0.000	***	2.54
pr_sum3a	-0.25	0.03	-8.47	0.000	***	3.00
pr_sum4a	0.06	0.02	3.87	0.000	***	2.94
pr_sum6a	0.04	0.01	4.59	0.000	***	2.98
pr_sum7b	-0.05	0.01	-6.50	0.000	***	3.38
pr_sum8a	-0.02	0.01	-4.45	0.000	***	2.55
pr_sum10a	-0.18	0.02	-8.38	0.000	***	2.14
ss_dura1a	2.37	0.73	3.24	0.001	**	3.67
ss_dura5a	2.97	0.62	4.77	0.000	***	5.16
ss_dura6b	-3.07	0.49	-6.29	0.000	***	3.06
ss_dura8b	5.07	0.37	13.52	0.000	***	4.17
ss_dura9b	-0.27	0.48	-0.56	0.573		3.42
adjusted R^2	0.6929					

<표 3-13> 10일 단위 BMA 모형의 선형회귀모형 계수 추정 결과

변수	계수	표준오차	t-value	Pr(>|t|)		VIF
(Intercept)	222.52	17.17	12.96	0.000	***	
$Yield_{t-1}$	0.56	0.01	48.46	0.000	***	1.54
tm_avg5b	-15.42	1.10	-14.06	0.000	***	13.95
tm_avg6b	-2.61	0.60	-4.32	0.000	***	3.81
tm_avg10b	11.93	0.92	12.90	0.000	***	17.97
tm_avg10c	5.85	0.44	13.44	0.000	***	5.29
tm_max2a	-2.08	0.28	-7.36	0.000	***	4.13
tm_max2c	1.52	0.26	5.97	0.000	***	3.12
tm_max3b	3.03	0.31	9.86	0.000	***	3.66
tm_max6c	0.75	0.50	1.50	0.133		5.41
tm_max8b	5.43	0.34	16.19	0.000	***	4.22
tm_min2b	1.70	0.40	4.26	0.000	***	6.55
tm_min3a	-4.27	0.41	-10.51	0.000	***	4.70
tm_min5b	12.95	0.99	13.13	0.000	***	13.29
tm_min8c	-4.84	0.53	-9.17	0.000	***	5.23
tm_min9b	-5.56	0.37	-15.20	0.000	***	4.14
tm_min10b	-9.15	0.83	-11.01	0.000	***	22.94
hm_avg1b	-0.53	0.09	-5.76	0.000	***	4.18
hm_avg1c	0.43	0.09	4.82	0.000	***	3.72
hm_avg3b	0.96	0.09	10.17	0.000	***	3.11
hm_avg6b	-1.58	0.14	-11.07	0.000	***	6.64
hm_avg10b	0.65	0.12	5.37	0.000	***	3.39
pr_sum1a	-0.61	0.06	-10.04	0.000	***	2.47
pr_sum1c	-0.37	0.07	-5.49	0.000	***	1.93
pr_sum3c	-0.43	0.04	-11.55	0.000	***	2.74
pr_sum4c	0.10	0.02	5.27	0.000	***	2.74
pr_sum7a	-0.06	0.01	-8.35	0.000	***	2.04
pr_sum7b	-0.04	0.01	-7.95	0.000	***	2.04
pr_sum8c	-0.05	0.01	-8.80	0.000	***	2.15
ss_dura3c	-4.51	0.50	-8.95	0.000	***	3.73
ss_dura4b	6.19	0.51	12.19	0.000	***	3.30
ss_dura5b	5.48	0.59	9.36	0.000	***	6.36
ss_dura6b	-4.18	0.44	-9.61	0.000	***	4.58
ss_dura6c	3.33	0.46	7.20	0.000	***	4.79
ss_dura9b	-3.79	0.40	-9.56	0.000	***	2.84
adjusted R^2	0.7032					

4. 10-fold 교차검증에 의한 예측 성능 평가

10-fold 교차검증에 의한 예측 성능 평가는 선행 연구에서 조사된 5개의 모형과 이 모형들을 기초로 만들어진 비교모형, 그리고 베이지안 모형 평균화 기법에 의해 만들어진 월 단위, 15일 단위, 10일 단위 모형까지 총 9개 모형의 예측 성능을 비교하였다<표 3-14>.

학습 기법에서는 모형을 구성하는 변수 수가 많지 않은 기존의 선행 연구의 모형들에서는 k-최근접이웃(k-NN) 기법의 예측 성능이 전반적으로 높은 반면 변수의 수가 많은 비교 모형과 베이지안 모형 평균화에 의한 최적화 모형에서는 k-NN 보다 서포트 벡터 회귀(SVR) 기법에서 가장 우수한 성능을 보여주었다.

전체 모형 중 가장 예측 성능이 높은 경우는 제곱근평균오차(RMSE), 평균절대백분율오차(MAPE), 예측치와 실측치 간의 R^2 모든 평가 지표에서 SVR이 학습한 월 단위 변수의 베이지안 모형 평균화 최적 모형이었다. 그러나 선행연구에 기초한 비교 모형과 비교했을 때 예측성능의 개선이 뚜렷하지는 않았다. 베이지안 모형 평균화를 통해 만들어진 15일 단위, 10일 단위 모형은 월 단위 모형과 유사한 수준의 예측 성능을 보여주었다. 베이지안 모형 평균화를 통해 만들어진 3 가지 모형의 예측 성능은 SVR, k-NN과 같은 기계 학습 기법에서 보다는 OLS, GAM 기법에서 두드러지게 향상되었다. 이는 베이지안 모형 평균화에서 후보 모형의 변수들 간 선형 관계를 전제로하여 최적 모형을 도출하였기 때문에 선형성을 가지는 학습 기법에서 이점을 많이 보는 것으로 추측된다.

<표 3-14> 선행 연구 모형과 BMA 모형의 10-fold 교차검증에 의한 예측 성능 평가

지표	기법	KREI (2011)	권 (2008)	이 (2012)	남 (2012)	조 (2013)	비교모형	BMA 월변수	BMA 15 일 변수	BMA 10 일 변수
RMSE	OLS	31.95	31.76	30.24	29.31	30.38	27.54	25.70	**23.88**	23.89
	GAM	31.52	31.28	29.21	28.34	28.85	25.50	24.34	**23.02**	23.52
	SVR	28.00	28.68	26.15	25.79	23.31	21.07	**20.96**	21.05	20.99
	kNN	26.17	27.24	24.99	25.25	**23.32**	23.92	23.98	24.20	24.05
MAPE	OLS	5.0%	5.0%	4.8%	4.6%	4.8%	4.3%	4.0%	**3.7%**	**3.7%**
	GAM	5.0%	5.0%	4.6%	4.5%	4.6%	4.0%	3.8%	**3.6%**	3.7%
	SVR	4.3%	4.5%	4.0%	4.0%	3.6%	**3.2%**	**3.2%**	**3.2%**	**3.2%**
	kNN	4.0%	4.1%	3.8%	3.9%	**3.6%**	3.7%	3.7%	3.8%	3.8%
R^2	OLS	45.5%	46.1%	51.2%	54.1%	50.7%	59.5%	64.7%	**69.5%**	**69.5%**
	GAM	47.0%	47.7%	54.4%	57.1%	55.6%	65.5%	68.5%	**71.7%**	71.5%
	SVR	58.1%	56.0%	63.5%	64.4%	71.0%	76.3%	**76.6%**	76.4%	**76.6%**
	kNN	63.3%	60.3%	66.7%	65.9%	**71.0%**	69.5%	69.3%	68.9%	69.2%

5. 리브-원-아웃(Leave-one-out) 교차검증에 의한 예측 성능 평가

리브-원-아웃(Leave-one-out) 교차검증에 의한 예측 성능 평가 방법을 이용하여 10-fold 교차검증에서와 동일하게 총 9개 모형의 예측 성능을 비교한 결과는 <표 3-15>와 같다. 학습 기법별로는 모형별로 차이는 있으나 평균적으로 OLS와 SVR의 성능이 비슷한 수준으로 가장 좋은 편이었으며, 다음은 GAM, kNN 순이었다. GAM과 kNN은 남(2012)의 모형에서 예측성능이 높은 편이었으나 다른 모형들에서는 우위를 보이지 못하였다.

평가 지표별로 예측 성능 차이를 살펴보면 제곱근평균오차(RMSE) 비교에서는 OLS로 학습한 베이지안 모형 평균화로 만들어진 10일 단위 변수의 BMA 모형의 예측 성능이 가장 높았고 다음으로 OLS에 의한 15일 단위 BMA 모형, OLS에 의한 10일 단위 BMA 모형, SVR에 의한 월 단위 BMA 모형, SVR에 의한 15일 단위 BMA 모형 순으로 근소한 차이로 RMSE가 낮았다. 평균절대백분율오차(MAPE) 비교에서는 OLS에 의한 10일 단위 BMA, SVR에 의한 10일 단위 BMA, OLS, SVR에 의한 월 단위 BMA, 15일 단위 BMA 순으로 4.1%~4.5%의 MAPE 값을 보였고 나머지 모형에서는 5% 이상의 범위에서 MAPE 값이 형성되었다. R^2 비교에서도 10일 단위 BMA모형에서 SVR, OLS로 학습한 모형의 R^2값이 53.4%, 53.1%로 가장 높았고, 다음으로 OLS, SVR에 의한 15일 단위 BMA 모형이 51.4%, 51.3% 순으로 뒤를 이었다.

선행연구의 변수들을 조합한 비교모형의 경우 다른 선행연구 모형보다는 성능이 앞서는 편이었으나 남(2012) 모형보다는 열세를 보였고, BMA를 이용해 선별된 모형과 비교하였을 때에도 예측력이 낮은 것으로 분석되었다.

<표 3-15> 선행 연구 모형과 BMA 모형의 리브-원-아웃(Leave-one-out) 교차검증에 의한 예측 성능 평가

지표	기법	KREI (2011)	권 (2008)	이 (2012)	남 (2012)	조 (2013)	비교모형	BMA 월변수	BMA 15 일 변수	BMA 10 일 변수
RMSE	OLS	32.73	32.50	31.09	29.72	31.81	30.91	27.34	27.06	25.50
	GAM	33.68	33.35	32.10	29.98	33.22	34.87	34.77	33.40	33.81
	SVR	32.54	32.51	31.07	29.58	31.61	30.28	27.53	27.96	25.93
	kNN	37.50	37.29	34.04	31.90	36.75	36.07	37.64	40.92	40.01
MAPE	OLS	5.4%	5.4%	5.1%	4.9%	5.2%	5.0%	4.4%	4.4%	4.1%
	GAM	5.6%	5.5%	5.3%	4.9%	5.5%	5.9%	5.8%	5.8%	5.6%
	SVR	5.4%	5.4%	5.1%	4.8%	5.2%	4.9%	4.4%	4.5%	4.2%
	kNN	6.3%	6.2%	5.6%	5.3%	6.1%	6.0%	6.3%	6.9%	6.8%
R^2	OLS	45.3%	44.9%	45.9%	46.7%	45.4%	46.5%	49.1%	51.4%	53.1%
	GAM	41.6%	42.3%	40.5%	43.5%	41.9%	39.4%	40.2%	40.5%	41.3%
	SVR	47.0%	46.4%	47.2%	48.1%	47.3%	48.0%	49.7%	51.3%	53.4%
	kNN	26.9%	28.9%	31.5%	33.6%	30.0%	24.3%	23.9%	20.6%	17.4%

5. 과거 시점 학습에 의한 최근 7년간 예측 성능 평가

과거 시점의 데이터로 학습한 예측 모형에 의한 최근 7년 간의 예측 성능 평가 방법을 이용하여 총 9개 모형의 예측 성능을 비교한 결과는 <표 3-16>와 같다. 학습 기법별로는 모형별로 차이는 있으나 평균적으로 OLS와 SVR의 성능이 비슷한 수준으로 가장 좋은 편이었으며, 권(2008)의 모형의 GAM이 우수하였고, 남(2012)의 모형은 kNN으로 학습한 모형의 예측성능이 가장 높게 나타나 모형에 따라 상이한 결과를 보였다.

평가 지표별로 예측 성능 차이를 살펴보면 제곱근평균오차(RMSE) 비교에서는 OLS로 학습한 베이지안 모형 평균화로 만들어진 10일 단위 변수의 BMA 모형의 예측 성능이 가장 높았고 다음으로 OLS에 의한 15일 단위 BMA 모형, SVR에 의한 비교 모형, OLS에 의한 월 단위 BMA 모형, SVR에 의한 남(2012)의 모형 순으로 근소한 차이로 RMSE가 낮았다. 평균절대백분율오차(MAPE) 비교에서는 SVR에 의한 비교 모형과 OLS에 의한 10일 단위 BMA 모형이 3.8%로 가장 우수한 예측 성능을 보였고, 다음으로 월 단위, 15일 단위 BMA 모형, 남(2012)의 예측 모형이 동일하게 3.9%로 유사한 수준을 보였다. R^2 비교에서는 남(2012)의 모형이 OLS, SVR로 학습한 모형의 R^2값이 64.9%, 64.1%로 가장 높았고, 다음으로 OLS에 의한 10일 단위 BMA 모형이 63.4%, SVR, OLS에 의한 15일 단위 BMA 모형이 63.0%, 62.6%, SVR, OLS에 의한 월 단위 BMA 모형이 61.1%, 60.2%, 순으로 뒤를 이었다. 비교 모형은 SVR은 59.9% OLS는 57.6%로 남(2012)의 모형보다 4%point 이상 차이를 보였다.

최근 7년 간의 연도별로 예측 오차를 평가하기 위해 평균절대백분율오차(MAPE)의 예측 성능이 가장 높았던 학습 기법을 이용하여 각 연도별 MAPE를 비교하였다. 태풍의 영향으로 예측오차가 컸던 것으로 보고된 2012년의 경우 선행연구의 모형들과 그에 기초한 비교 모형, 베이지안 모형 평균화에 의한 월 단위 BMA 모형의 경우 MAPE가 5.5~6.0%

로 유사한 수준을 보이는 반면 베이지안 모형 평균화에 의한 15일 단위, 10일 단위 BMA 모형의 경우 MAPE가 4.7%로 0.8~1.3%point 개선된 성능을 보여주었다. 월 단위의 기상 변수를 이용할 때는 포착하지 못한 쌀 단수에 대한 영향 요인을 15일 단위, 10일 단위 변수를 사용한 예측 모형에서는 포착할 수 있었기 때문으로 추측된다. 15일 단위, 10일 단위 BMA 모형은 2017년을 제외한 나머지 연도에서도 비교 대상 모형들과 유사한 수준 이상의 예측 성능을 보여주고 있어 월 단위보다 짧은 시간 간격의 기상 변수를 이용한 예측 모형의 잠재력을 보여준다.

기존 선행연구의 모형 중에는 KREI(2011)의 모형은 2014년에 베이지안 모형 평균화에 의한 월 단위 BMA 모형과 함께 3.2%의 MAPE 최저치를 보였고, 권(2008), 이(2012)의 모형은 2017년에 선행연구에 기초한 비교 모형과가 함께 3.0%의 MAPE 최저치를 보였다. 2014년과 2017년의 쌀 단수에 영향을 미친 기상 요인이 각각의 모형에서 사용한 설명 변수에 의해 예측이 잘 된 것으로 볼 수 있다.

<표 3-16> 선행 연구 모형과 BMA 모형의 최근 7년간에 대한 예측 성능 평가

지표	기법	KREI (2011)	권 (2008)	이 (2012)	남 (2012)	조 (2013)	비교모형	BMA 월변수	BMA 15 일 변수	BMA 10 일 변수
RMSE	OLS	31.89	28.11	27.60	25.65	28.57	26.33	25.49	25.02	24.67
	GAM	32.65	29.16	30.47	39.81	39.81	39.81	39.81	38.50	39.12
	SVR	28.57	27.80	27.33	25.53	28.14	25.45	25.61	25.07	26.14
	kNN	31.90	31.39	33.32	30.75	31.93	35.10	36.11	38.02	38.66
MAPE	OLS	4.8%	4.1%	4.0%	3.9%	4.3%	4.0%	3.9%	3.9%	3.8%
	GAM	5.0%	4.3%	4.5%	6.3%	6.3%	6.3%	6.3%	6.3%	6.3%
	SVR	4.3%	4.0%	4.5%	3.9%	4.2%	3.8%	3.9%	3.9%	4.0%
	kNN	4.8%	4.7%	5.1%	4.7%	4.7%	5.3%	5.6%	6.0%	6.0%
R^2	OLS	38.6%	57.7%	55.0%	64.1%	52.2%	57.6%	60.2%	62.6%	63.4%
	GAM	34.6%	50.0%	42.9%	37.4%	37.4%	37.4%	37.4%	37.6%	37.5%
	SVR	50.6%	57.6%	56.3%	64.9%	53.2%	59.9%	61.1%	63.0%	60.4%
	kNN	38.0%	39.6%	36.0%	49.5%	39.3%	32.1%	27.4%	23.2%	26.4%
연도별 예측오차 (MAPE)	2012	6.0%	6.0%	6.0%	5.7%	5.7%	5.6%	5.5%	4.7%	4.7%
	2013	5.1%	4.4%	4.4%	4.6%	4.8%	4.3%	4.4%	5.4%	4.2%
	2014	3.2%	3.4%	3.4%	3.3%	3.6%	3.8%	3.2%	3.4%	3.3%
	2015	3.3%	2.9%	2.9%	2.9%	2.9%	3.6%	3.7%	2.9%	2.7%
	2016	4.0%	3.4%	3.4%	3.3%	3.7%	3.1%	3.2%	2.9%	3.2%
	2017	4.0%	3.0%	3.0%	3.4%	3.2%	3.0%	3.8%	4.1%	4.0%
	2018	4.5%	4.9%	4.6%	3.8%	5.3%	3.5%	3.4%	3.6%	4.7%
	7년평균	4.3%	4.0%	4.0%	3.9%	4.2%	3.8%	3.9%	3.9%	3.8%

제 5 절 소결

지금까지 베이지안 모형 평균화를 이용하여 기존의 쌀 단수 예측 모형 연구에서 변수 선별 방법이 가지는 한계를 극복할 수 있는지에 대하여 다각적인 방향에서 검증하였다. 지역은 일부 제외될 수 있으나 모든 시점의 데이터를 학습에 활용할 수 있는 10-fold 교차검증에 의한 예측 성능 평가에서는 베이지안 모형 평균화에 의해 선정된 최적 모형이 선행연구에 기초한 비교 모형과 비교했을 때 동등한 수준의 예측 성능을 보여주는 것을 확인하였다. 또한 선행연구에서는 다루고 있지 않은 새로운 변수들이 제시되었고 선형회귀분석 결과 제시된 새로운 변수들이 모두 통계적으로 유의하였으며 모형의 R^2 역시 높아지는 결과를 확인하였다. 한편 리브-원-아웃(Leave-one-out) 교차검증에 의한 예측 성능 평가에서는 베이지안 모형 평균화에 의해 만들어진 모형들이 선행연구의 모형과 선행연구를 종합한 비교모형보다도 우수한 예측 성능을 보여주고 있어 베이지안 모형 평균화에 의한 변수 선별이 예측의 관점에서도 성능이 우수한 예측 모형의 개발에 도움을 줄 수 있음을 확인하였다. 반면 최근 7년 간의 시점을 예측하기 위해 그 이전 시점의 데이터만으로 모형을 학습하여 예측 성능을 평가한 결과에서는 기존 선행 연구의 모형과 비교하였을 때 베이지안 모형 평균화에 의한 최적 모형의 예측 성능이 더 나아질 수 있다는 근거는 발견하지 못하였다. 불확실한 미래를 예측하는 상황에서는 오히려 단순한 모형이 더 나은 예측 성능을 보여주기도 하였다. 또한 베이지안 모형 평균화에 의한 최적 모형들은 SVR과 같은 변수 간의 비선형성을 잘 포착할 수 있는 기계 학습 기법 보다 OLS와 같은 모수 기반의 단순한 방법에 의해 학습한 모형에서 근소하게 예측 성능이 앞서는 것도 확인되었다. 이는 불확실성이 큰 예측 상황에서는 변수 간의 비선형성을 잘 포착하는 것만큼이나 설명력이 높은 변수가 추가되는 것에 의한 예측 성능 개선 효과가 크다는 것을 시사한다. 2012년과 같이 태풍에 의해 과거의 평균적인 기상 정보만으로는 예측이 힘든 상황에서는 베이지안 모형 평균화에 의해 15일 단위, 10일 단위의 짧은 시간 간

격에서의 기상 변수를 탐색하여 예측 모형을 구축하는 것이 월 단위 변수를 이용한 예측 모형보다 나은 예측 성능을 기대할 수 있다는 사실도 발견하였다. 이와 같이 본 연구는 베이지안 모형 평균화 기법은 쌀 단수 예측과 같이 긴 기간의 다양한 기상 요인에 의해 영향을 받을 수 있는 상황에서 예측 모형의 설명 변수를 탐색하기 위한 실용적인 목적에서는 유의미한 방법이 될 수 있다는 것을 확인하였다.

선행연구에 기초한 모형이 베이지안 모형 평균화에서 도출된 모형과 예측 성능의 관점에서 차이가 없다는 점은 베이지안 모형 평균화와 같은 기계 학습에 의해 제시되는 새로운 모형이 기존의 이론에 기반한 모형을 대체하는 것은 어렵겠지만 데이터에 기반한 모형이 기존의 이론 기반 모형의 보완재의 역할을 할 수 있다는 가능성을 보여주었다. 기존의 선행연구에서 제시한 모형의 변수로는 설명되지 않는 종속변수의 변량을 베이지안 모형 평균화와 같은 데이터에 기반한 모형을 통해서 찾아낼 수 있으며 새롭게 발견된 변수를 사용하여 기존의 예측 모형의 성능을 개선하는 것에 도움을 줄 수 있다.

그러나 선행연구에서 다루어지지 않은 새로운 변수에 대해서는 새로운 변수와 종속변수 간의 관계를 추가로 밝혀야 한다. 예를 들어 베이지안 모형 평균화에 의해 새롭게 발견된 변수인 1월의 최저기온, 최고기온, 강수량 변수는 회귀계수가 음수로 쌀 단수에 부정적인 영향을 미치는 것으로 나타났는데 이는 병해충에 의한 부정적 영향으로 추측해볼 수 있다. 농촌경제연구원(2012)의 연구에 따르면 벼의 해충인 애멸구, 벼물바구미는 겨울철 최저기온이 높을수록 병의 피해가 커지는 것으로 보고하고 있다. <부록 4>의 <그림 부-3>을 보면 2월의 최저기온은 쌀 단수와 음의 상관관계를 가지고 -0.14의 유의한 상관계수를 가진다. 벼의 생육기의 월별 기온 관련 변수 중 상관계수가 0.2정도의 값을 가지는 변수도 있는 것을 고려하면 유의미한 값으로 볼 수 있다. 한편 <표 3-11>에서 유의한 변수로 분석된 1월 강수량 변수는 <부록 4>의 <그림 부-4>에서도 쌀 단수와 -0.16의 유의한 상관계수를 가지는 것으로 나타났다. 겨울철의 강수량이 많을 경우 병해충과 같은 쌀 단수에 부정적인 영향을 줄

수 있는 요인이 있는 것으로 추정되지만 겨울철 강수량과 벼의 병해충 간의 관계에 대한 연구는 조사되지 않아서 근거를 찾지 못하였다. 다만 <부록 4>의 <그림 부-5>의 상관관계 표에서와 같이 1월 강수량은 1월 최저기온과 양의 상관관계를 가지는데 이는 1월 강수량이 많을 경우 겨울철 최저기온이 낮음으로 인한 병해충의 영향이 있을 것으로 추측해볼 수 있다. 쌀 단수의 영향 변수가 1월 최저기온이 아닌 1월 강수량이 유의한 변수로 추정된 것은 추가적인 변수들 간의 영향 관계를 살펴봐야 할 것이다. 1, 2월 기상 요인이 벼 생육에 미치는 직접적인 영향 외에 1, 2월 기상 요인이 다른 시기의 기상 요인과 가지는 상관관계로 인한 간접적인 영향도 고려해볼 수 있다. <부록 4>의 <그림 부-3>의 쌀 단수와 월별 최저기온 간의 관계를 보면 9월 최저기온이 쌀 단수와 -0.12의 상관관계를 가지고 있어 부정적인 영향을 미치고 있는데 9월의 최저기온은 1, 2월 최저기온과 0.58, 0.66의 높은 양의 상관관계를 가지는 것이 확인되었다. 즉 1, 2월의 최저기온이 높을수록 9월의 최저기온이 높고 이로 인한 쌀 단수에 부정적인 영향을 미칠 수 있는 것이다. 이와 같이 겨울철 기상 요인이 쌀 단수에 미칠 수 있는 직접적, 간접적 관계에 대한 여러가지 가능성은 고려해보았으나 통계적으로 유의한 관계를 밝히기 위해서는 추가적인 검증이 필요하다. 보다 엄밀한 연구를 위해서는 이론적인 타당성을 가지는 5~10월의 기상 요인을 이용한 예측 모형을 먼저 검증한 다음 1~10월의 기상 요인에 대한 추가적인 검토를 통해서 논의를 확장하는 절차가 고려되어야 했으나 그러지 못한 점은 본 연구의 한계점으로 향후 연구에서 검토가 필요하다.

10-fold 교차 검증의 경우 최근 7년 예측과 달리 전반적인 예측 오차가 낮고 선행연구의 단순한 모형보다 비교 모형과 베이지안 모형 평균화에 의한 보다 복잡한 모형이 좋은 예측 성능을 보였다. 이는 학습에 사용되는 데이터가 전체 기간 포함하고 있을 경우 학습에 사용되지 않는 테스트 데이터의 관측치 중 공간적 상관성이 높은 인근 지역의 관측치가 모형 학습 과정에 포함될 수 있기 때문이다. 이와 같은 문제를 보완할 수 있는 방법으로 리브-원-아웃(leave-one-out) 교차검증을 이용한 예측

성능을 평가하여 학습 데이터와 평가 데이터 간에 공간적 상관성이 없는 상태에서의 예측력을 비교하였다.

과거 시점의 데이터로 학습 후 최근 7년에 대한 예측 성능 평가에서는 예측 모형의 성능이 전반적으로 낮고 복잡한 모형과 단순한 모형의 예측 성능 차이가 크지 않은 것은 과거 시점의 15년 간 데이터로 학습된 모형이 최근 7년 간의 독립변수와 종속변수 간의 관계를 잘 설명하지 못하기 때문에 예측 성능이 떨어지는 것으로 볼 수 있다. 시계열 데이터를 이용해 과거 데이터로 학습된 모형으로 미래를 예측하는 상황에는 미래의 불확실성을 극복할 수 있는 근본적인 해결책은 없다. 그러나 2012년의 태풍으로 인한 이상 상황에서 15일, 10일 단위 변수를 사용하는 모형의 예측 성능이 높았던 것과 같이 미래의 불확실한 상황을 설명할 수 있는 변수를 지속적으로 발굴하여 예측 모형을 업데이트한다면 미래의 불확실성을 점차 줄일 수 있는 방안이 될 수 있을 것이다.

제 4 장 부분지도학습 기법을 이용한 쌀 단수 예측 성능 개선

제 1 절 서론

작물 단수 예측과 관련한 기존 연구 중 상당수는 최근까지도 회귀 모형과 같은 모수적 방법에 기반하고 있으며, 예측력의 개선보다는 설명변수의 영향력과 지역별 차이에 대한 설명에 더 초점을 맞추고 있다. 비모수적 방법 중 기계 학습 방법은 변수 간의 비선형적 관계를 포착하여 모델링할 수 있는 장점이 있고 관측된 데이터가 충분히 큰 상황에서는 예측 성능도 모수적 방법보다 우수하다. 그러나 모형의 학습이 적절히 수행되기 위해서는 충분한 양의 데이터가 필요하기 때문에 데이터가 부족한 상황에서는 기계 학습 방법을 사용하는 것이 적절하지 않다.

작물 단수 예측에 대한 연구 중 대부분의 연구들은 미국, 캐나다, 호주, 중국을 대상으로 하고 있는데 이는 농업이 이들 국가의 주요 산업이고 동시에 넓은 국토에서 수집될 수 있는 데이터의 양이 많기 때문이다. 반면 우리나라와 같이 국토의 면적이 넓지 않은 국가에서는 수집될 수 있는 데이터의 범위가 제한될 수 있다. 또한 농업선진국과는 달리 농업이 핵심 산업이 아닌 경우 전국을 대상으로 하는 통계 데이터 수집의 우선순위가 높지 않을 수 있다. 어떠한 경우이든 농업 선진국이 아닌 국가에서는 다양한 연구를 수행할 수 있는 충분한 데이터가 수집되지 않는 문제를 가지고 있다.

국내의 경우 주곡인 쌀은 시·군 단위로 단수 통계가 집계되고 있으나 나머지 작물의 경우 광역도 단위에서만 단수 통계가 집계되고 있기 때문에 쌀과 비교했을 때 데이터의 양이 10% 이하 수준에 그치고 있다.

머신러닝 분야에서는 독립변수는 관측되나 종속변수의 관측치가 없는 레이블 없는 데이터(Unlabeled Data)를 모형 학습에 활용하는 준지도 학습(Semi-Supervised Learning) 방법들이 연구되고 있다. 이와 같은 준지

도 학습 방법은 쌀 외의 광역도 단위에서만 단수 통계가 조사되고 있는 작물의 경우에 유용하게 활용될 수 있다. 광역도 단위로 집계되는 단수와 해당 도의 평균 기상 관측치를 레이블 데이터로 볼 수 있고 단수가 조사되지는 않지만 기상 데이터는 관측되는 시·군 지역은 레이블이 없는 데이터로 볼 수 있다. 따라서 광역도 단위의 레이블 데이터를 이용하여 모형을 학습하고 시·군 단위의 레이블 없는 데이터를 추가로 투입하면서 예측 성능을 높이는 준지도 학습 방법의 적용 상황과 일치한다.

본 연구는 광역도 단위 집계에 의해 관측치가 부족한 경우에 이 데이터를 이용한 예측 모형을 성능을 높일 수 있는 방법으로 준지도 학습 방법을 검증하는 것을 목적으로 한다. 이를 위해 쌀 단수 레이블이 있는 광역도 단위 집계 데이터만으로 지도학습으로 만들어진 예측 모형과 쌀 단수 레이블이 있는 광역도 단위 집계 데이터와 레이블이 없이 기상 정보만 가지고 있는 시·군 단위 데이터를 이용하는 준지도 학습 기법을 적용한 예측 모형의 성능을 비교한다. 준지도 학습 기법으로 학습된 모형의 예측 성능은 준지도 학습에 의해 기상 데이터만으로 예측된 시·군 단위 쌀 단수 예측치를 시·군 단위로 집계되어 레이블을 가지는 실제 쌀 단수 실측치와 비교하여 제곱근평균오차(RMSE), 평균절대백분율오차(MAPE), 예측치와 실측치 간의 R^2의 3 가지 지표를 계산하여 예측 성능을 측정하였다.

제 2 절 선행연구 검토

1. 준지도 학습 기법을 적용한 농업 분야의 연구

작물 단수 예측과 관련한 연구에서는 아직까지 준지도 학습(Semi-Supervised Learning) 기법이 사용된 연구는 없으며, 농업 분야에서는 이미지 판별 연구에서 준지도 학습 기법이 주로 사용되고 있다. 경작지나 위성에서 촬영한 이미지(다중스펙트럼 이미지 포함)에서 특정 대상을 판별하거나 특정 물질의 함량을 추정할 때 레이블이 없는 데이터는 많이 확보할 수 있으나 종속변수가 포함된 레이블이 있는 데이터는 별도의 측정이나 수작업에 의한 판별 과정이 필요하기 때문에 데이터의 수가 많지 않다. 이와 같이 레이블 데이터는 적고 레이블이 없는 데이터가 많은 상황이 준지도 학습 기법이 사용되기에 적합한 상황이다. 아직까지는 농업 분야의 준지도 학습 기법을 이용한 연구가 많지 않으나 이미지 판별과 관련한 일부 연구가 있다. 드론으로 촬영한 해바라기 경작지 이미지에서 잡초를 판별하는 분류 모형 연구에서 준지도 학습 기법을 적용할 경우 비지도 학습에 의한 분류 모형 대비 예측 오차가 50% 가량 개선되었고 지도 학습 기법들과 비교했을 때는 예측 성능이 20~40% 가량 개선되었다(Pérez-Ortiz et al., 2015). 경작용 로봇 촬영 이미지에서 잡초를 판별하는 연구에서는 준지도 학습에 의한 분류 모형이 지도 학습에 의한 분류 모형 대비 예측 성능이 2~4%point 개선되었다(Lottes & Stachniss, 2017). 과수원의 다중스펙트럼 이미지로부터 아몬드 과수의 잎, 열매, 줄기, 지면, 하늘을 판별하는 연구에서는 기존의 밴치마크 방법 대비 8%point 개선된 88%의 예측 정확도까지 성능이 개선되었다(Hung et al., 2013). 위성 이미지의 지표면 대상의 종류를 판별하는 연구들에서도 준지도 학습 기법으로 인해 학습용 데이터가 추가될수록 판별 정확도가 높아지는 것이 확인되었다(Romaszewski et al., 2016; Silva et al., 2017).

현재까지는 준지도 학습 기법은 대부분 명목형 종속변수에 대한 분류

연구에 사용되고 있는 반면 연속형 변수를 예측하기 위한 준지도 회귀(Semi-Supervised Regression) 기법은 분류 모형과는 달리 알고리즘이 복잡도가 높아 비교적 최근에 연구가 시작되어 상대적으로 관련 연구가 많지 않다. 토양의 유기탄소량 추정(Liu et al.,2017), 강의 수질 예측(Wang et al., 2010) 연구에서 Co-training을 이용하는 준지도 회귀 기법을 사용하여 예측 성능이 개선되는 것이 확인되었다.

2. 준지도 회귀(Semi-Supervised Regression) 기법 연구

준지도 학습 기법은 접근방식에 따라 Self-training, Generative Model, Co-training, Margin-based, Graph-based Method 등으로 알고리즘을 구분할 수 있다(Bagherzadeh & Asil, 2019). 준지도 학습에 관한 이전의 연구는 주로 분류에 초점을 두고 있는 반면 준지도 회귀 기법은 상대적으로 무시되어 왔는데(Zhou & Li, 2007), 최근에는 몇 가지 연속형 종속변수를 처리하는 준지도 학습 방법들이 제안되고 있다. 관련 연구로 Co-training 방식(Wang et al., 2011; Wang et al., 2010; Zhou & Li, 2007), Graph 기반 방식(Pozdnoukhov & Bengio, 2006), Kernel 기반 방식(Cortes & Mohri, 2007; Wang et al., 2006) 등이 있는데 여러 접근법 중 Co-training 방식은 Self-training 방식보다 성능이 우수하고 레이블 데이터가 적은 경우에도 좋은 성능을 발휘하는 것으로 알려져 있다(Bagherzadeh & Asil, 2019).

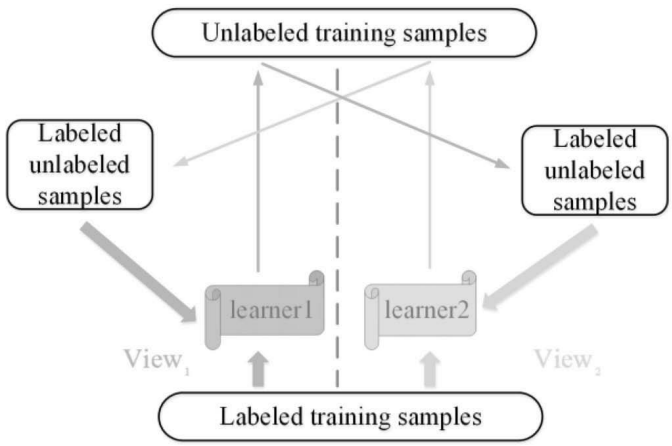
Co-training 방식은 Blum & Mitchell(1998)에 의해 처음 제안되었고, 두 개의 분류 자, 예를 들어 두 개의 독립적인 변수 집합을 가지는 모형에 대해 개별적으로 두 분류기(Classifier)를 훈련시키고, 레이블이 없는 관측치를 입력한 각 분류기의 예측치를 사용하여 다른 분류기의 훈련용 데이터 세트를 보강하는 개념이다. Zhou & Li(2005)는 레이블이 없는 데이터를 설명하기 위해 세 개의 분류기를 사용하는 Tri-training 개념을 제안했다. 이 과정에서 레이블이 없는 예제에 레이블이 지정되고 다른 두 분류기가 해당 레이블에 동의하는지 여부를 판별하는 데 사용된다. Li & Zhou(2007)은 렌덤포레스트(Random Forest) 앙상블 기법을 분류기에 적용하는 Co-Forest 개념을 제안하여 다수의 분류기를 훈련 과정에 통합함으로써 Co-training 아이디어를 더욱 확장시켰다.

그러나 준회귀 학습 기법에 관한 연구는 대부분 명목척도의 종속변수 가지는 예에 대한 분류의 문제에 초점이 맞추어져있는 반면 연속형 변수를 종속변수로 가지는 회귀(Regrssion)의 문제에서의 준지도 학습에 대한 연구는 많지 않다. Zhou et al.(2005)은 CoReg라는 Co-Training 스타

일의 준지도 회귀 알고리즘을 제안했다. 이 알고리즘은 두 개의 서로 다른 Minkowski 거리 순서 값을 사용하여 인스턴스화된 두 개의 서로 다른 k-최근접이웃(k-Nearest Neighbor: k-NN) 회귀분석기(Regressor)를 사용한다. 각각의 서로 다른 회귀분석기가 레이블이 없는 관측치에 레이블을 부여하고 이에 대한 신뢰도는 해당 관측치의 레이블이 있는 이웃에 대한 평균 제곱 오차 감소량을 통해 추정되는 방식으로 기존의 레이블이 지정된 훈련 세트와 일관성이 최대한 유지되도록 한다. Zhou & Li(2007)는 시뮬레이션 데이터 10종, 공개된 실증 데이터 5종에 대한 비교 분석 결과 레이블이 있는 데이터 수가 적은 경우일수록 성능 개선 효과가 좋았으며, 시뮬레이션 데이터에서는 10%, 실증 데이터에서는 3% 가량 성능이 개선되는 것을 확인하였다. 앞서 언급된 Liu et al.(2017), Wang et al.(2010)의 연구에서도 Co-training 방식의 준지도 회귀 기법이 적용되었으며 이를 통해 20% 가량 예측 성능이 개선되었다. 사회과학 분야에서도 주택가격 추정 연구에서 사용되어 기존 방식 대비 오차율 개선은 1%밖에 되지 않으나 예측치와 실측치 간의 R^2에서의 20~40%의 개선 효과를 확인하였다(Yang et al., 2016).

제 3 절 연구방법

1. 준지도 회귀(Semi-Supervised Regression) 기법



<그림 4-1> Co-training 준지도 학습 방법 개념도

준지도 회귀(Semi-Supervised Regression: SSR)의 기본 원리를 순차적으로 설명하면 우선 지정된 n개의 기본 회귀분석기로부터 준지도 회귀 알고리즘을 실행하는 커미티(committee)에 의한 Self-training 및 Co-training을 수행한다(회귀분석기 목록에 모형이 하나만 있으면 Self-training이 수행됨). 처음에는 레이블이 지정된 학습용 데이터 세트 L에 대해 독립적인 부트스트랩(bootstrap) 샘플로 학습된 n개의 기본 모형 세트가 커미티로 구성된다. 이 때 일부 샘플은 OOB(Out-of-Bag) 요소로 유효성 검사에 사용된다. 다음으로 레이블이 지정되지 않은 데이터 세트 U에서 가장 관련성이 높은 요소를 선택하여 각 기본 모형 b(base를 의미함)에 대한 학습용 데이터 세트가 보강된다. 각 기본 모형 b에 대해 가장 관련성이 높은 요소를 결정하기 위해 b를 제외한 다른 모형은 U로부터 u의 크기를 가지는 pool로 샘플링된 데이터 포인트 세트의 예측치 평균을 취하여 레이블값을 부여한다. 새로 레이블이 지정된 각 데이터 포인트에 대해 기본 모형 b는 현재 레이블이 지정된 학습용 데

이터와 신규 데이터 포인트를 추가하여 학습 알고리즘을 수행하고 OOB 유효성 검사를 통해 신규 데이터가 추가로 포함된 모형의 오차를 계산한다. 오차를 가장 많이 줄이는 최상위 gr개의 신규 데이터 포인트는 부여받은 레이블과 함께 b에 할당된 학습 세트에 포함되어 b의 성능을 보강하고 U에서는 제외시킨다.

본 연구에서 사용하는 준지도 회귀 방법은 앞서 설명한 커미티에 의한 Co-training에 기반한 방법으로 Hady et al.(2009)가 제시한 방법을 사용하였다. CoBCreg로 명명되는 Hady ea al.(2009)의 방법은 기존의 표준 Co-training 방법에서 확장된 개념으로 배깅(Bagging)으로 구축되는 방사기저함수(Radial Basis Function: RBF) 네트워크 회귀분석기의 앙상블을 토대로 하고 유클리디언(Euclidean) 거리 대신 민코스키(Minkowski) 거리에 기반한 새로운 가우시안 기저함수(Gaussian basis function)를 사용하는 것이 특징이다. Co-training에 기반한 준지도 회귀 방법 대한 설명은 Hady et al.(2009)의 논문을 참고하여 작성하였다.

CoBCreg는 <그림 4-2>, <그림 4-3>와 같은 두 파트의 알고리즘으로 구성되어 있다. L = $\{(x_i, y_i)\}$ m_i = 1 은 레이블이 지정된 학습용 데이터 세트이고 U = $\{x_i\}$ n_i = 1은 레이블이 지정되지 않은 데이터 세트를 나타낸다. y_i는 레이블이 있는 데이터 세트 L에서 인스턴스 x_i에 대한 종속변수의 실제값인 반면 데이터 세트 U에서는 인스턴스 x_i에 대한 종속변수의 실제값은 알지 못한다. CoBCReg는 처음 H로 표시되는 N개의 회귀분석기의 앙상블로 구성되며, 배깅을 통해 L로부터 데이터 세트가 할당된다. 다음으로 최대 반복 횟수 T에 도달하거나 U가 비워 질 때까지 다음 단계가 반복된다. 각각의 반복 t에서는 각 앙상블 멤버 h_i에 대해 U의 데이터 세트로부터 무작위로 u개의 예제로 추출되어 U'을 생성한다. SelectRelevantExamples 절차는 컴패니언 커미티 H_i(h_i를 제외한 모든 멤버로 구성된 앙상블)가 레이블이 없는 각 U' 예제의 출력값을 추정하는 과정에 적용된다. 그런 다음 h_i는 gr개의 가장 근사한 예제를 학습용 데이터 세트 L_i에 추가하여 h_i를 수정한다. 예측 단계(Prediction Phase)에서는 마지막 CoBCReg 반복에서 생성된 회귀분석기의 출력값의

가중 평균을 계산한 후 지정된 인스턴스 x_i에 대한 회귀 추정치로 부여한다.

Algorithm 1. CoBC for Regression

Require: L - set of m labeled training examples
U - set of n unlabeled examples, T - maximum number of Co-Training iterations
N - number of committee members (*ensemble size*), u - pool size
gr - growth rate, k - number of RBF hidden nodes, α- RBF width parameter
p_i - distance order of the i^{th} regressor

Training Phase

for $i = 1$ to N **do**
  $\{L_i, V_i\} \leftarrow BootstrapSample(L)$ {L_i is bag and V_i is out-of-bag}
  $h_i \leftarrow RBFNN(L_i, k, \alpha, p_i)$
end for
for $t \in \{1 \dots T\}$ **do**
  if U is empty **then** $T \leftarrow t$-1 and abort loop **end if**
  for $i \in \{1 \dots N\}$ **do**
    Create a pool U' of u examples by random sampling from U
    $\pi_i \leftarrow SelectRelevantExamples(i, U', V_i, gr)$
    $U' \leftarrow U' \setminus \pi_i$ and $U \leftarrow U \cup U'$
  end for
  for $i \in \{1 \dots N\}$ **do**
    if π_i is not empty **then**
      $L_i \leftarrow L_i \cup \pi_i$
      $h_i \leftarrow RBFNN(L_i, k, \alpha, p_i)$
    end if
  end for
end for

Prediction Phase

return $H(x) \leftarrow \sum_{i=1}^{N} w_i h_i(x)$ for a given sample x

<그림 4-2> 회귀를 위한 CoBC 알고리즘(Hady et al., 2009)

Algorithm 2. SelectRelevantExamples

Require: j - the index of the regressor excluded from the committee
U' - pool of u unlabeled examples V_j - validation set gr - growth rate

Calculate validation error of h_j using V_j, ϵ_j
for each $x_u \in U'$ **do**
  $H_j(x_u) \leftarrow \frac{1}{N-1}\sum_{i=1,i\neq j}^{N} h_i(x_u)$
  $h'_j \leftarrow RBFNN(L_j \cup \{(x_u, H_j(x_u))\}, k, \alpha, p_j)$
  Calculate validation error ϵ'_j of h'_j using V_j, then $\Delta_{x_u} \leftarrow (\epsilon_j - \epsilon'_j)/\epsilon_j$
end for
$\pi_j \leftarrow \phi$
for gr times **do**
  if there exists $x_u \in U' \setminus \pi_j$ with $\Delta_{x_u} > 0$ **then**
    $\tilde{x}_j \leftarrow arg\max_{x_u \in U' \setminus \pi_j} \Delta_{x_u}$
    $\pi_j \leftarrow \pi_j \cup \{(\tilde{x}_j, H_j(\tilde{x}_j))\}$
  end if
end for
return π_j

<그림 4-3> CoBCreg에서 연관 예제 선택을 위한 알고리즘(Hady et al., 2009)

회귀분석기의 앙상블 조합은 다양성을 가지는 경우에만 효과적이다. 이를 위해 CoBCreg에서는 Brown et al.(2005)이 제시한 앙상블의 다양성을 만드는 여러 기법 중 회귀분석기에서 사용 가능한 인스턴트 x에 대한 앙상블 다양성(분산)의 개념인 식(4-1)을 이용하여 정량화하였다.

$$\bar{A}(x) = \sum_{i=1}^{N} w_i (h_i(x) - H(x))^2. \qquad (4\text{-}1)$$

Krogh & Vedelsby(1995)는 앙상블오차(E)는 앙상블 멤버의 가중평균 오차(¯E)와 주어진 인스턴스에 대한 출력값의 다양성이라는 두 가지 항으로 분해된다는 오차-모호성 분해(error-ambiguity decomposition) 개념을 소개하였다. 즉, E = ¯E −A¯이다. 오차-모호성 분해의 앙상블 오

류를 감소시키기 위해서는 앙상블 멤버의 평균 오차는 낮아야하고, 다양성은 높아야한다 것을 함의한다. CoBCReg에는 RBF 네트워크 회귀분석기의 다양성 생성을 위해 3 가지 방식을 사용한다: (1) 상이한 부트 스트랩 샘플, (2) RBF 중심에 대한 상이한 랜덤 초기값, (3) 상이한 거리 측정 방법에 의한 학습. 식(4-2)에 정의 된 두 개의 D-차원 변수 벡터 x1과 x2 사이의 Minkowski 거리는 서로 다른 RBF 네트워크 회귀분석기를 훈련시키기 위해 서로 다른 거리 차수 p을 사용한다. 일반적으로 차수 p가 작을수록 데이터 변동에 대한 거리 측정 결과가 더욱 강력해진다.

$$\|x_1 - x_2\|_p = \left(\sum_{i=1}^{D} |x_{1i} - x_{2i}|^p \right)^{1/p} \qquad (4\text{-}2)$$

CoBCReg는 회구분석기에 의해 선택된 예제가 U에서 제거되기 때문에 회귀분석기 간의 다양성을 해치지 않는다. 따라서 회귀분석기의 학습용 데이터 세트를 유사하게 유지하지 않는 다른 회귀분석기에 의해 더 이상 선택 될 수 없게 된다. 학습용 데이터 세트가 유사해지더라도 회귀분석기는 거리 측정 방식에 의해 서로 다른 인스턴스로 작용할 수 있기 때문에 여전히 다양성을 유지할 수 있다.

Co-training 스타일 알고리즘의 성능에 영향을 미치는 가장 중요한 요소 중 하나는 레이블이 없는 특정 예제의 신뢰도를 측정하는 방법이다. 부정확한 신뢰도 추정된 레이블이 지정되지 않은 예제가 레이블이 지정된 학습용 데이터 세트에 추가될 경우 준지도 학습 알고리즘의 성능에 부정적인 영향을 줄 수 있다. 분류(Classification)의 문제에서는 많은 분류기가 Naive Bayes 분류기와 같은 클래스(class) 사후확률을 추정하거나 신경망 및 의사결정나무의 경우에도 클래스에 포함될 확률을 계산하는 것이 쉽기 때문에 0.5를 기준으로 클래스를 분류하고 신뢰도를 계산하는 방식으로 알고리즘의 구현이 비교적 간단하다. 그러나 회귀(Regression)의 문제에서는 가능한 예측치가 특정 값으로 정의되지 않기

때문에 신뢰도 추정이 어렵다. Krogh & Vedelsby(1995)에서는 앙상블 멤버의 추정치 간의 높은 분산이 높은 평균 오차를 초래하기 때문에 능동 학습 과정에서 효과적인 선택 기준으로 분산을 사용된다. 그러나 분산이 낮다는 것이 반드시 평균 오차가 낮다는 것을 의미하지는 않는다. 즉, 커미티 멤버가 동의한 예측치가 목표 실측값에 가깝다는 것을 의미하는 것은 아니다. CoBCReg에서는 레이블 부여에 대한 신뢰도(labeling confidence)는 측정하지 않지만 <그림 4-3>에서와 같이 선택 신뢰도(selection confidence)라는 신뢰도 측정을 제공한다. 가장 적절하게 선택된 예제는 검증 세트에서 회귀 오류를 최소화하는 것이어야 하므로 각 회귀분석기 h_j에 대해 레이블이 지정되지 않은 예제풀 U'를 만든다. 그런 다음 h_j의 제곱근평균오차(RMSE) ε_j를 먼저 평가한다. 그런 다음 U'의 각 예제 x_u에 대해 h_j는 $(x_u, H_j(x_u))$의 형태의 새로운 회귀분석기 h'_j를 만든다. 이제 h_j의 RMSE ε'_j를 평가한다. 여기서 $H_j(x_u)$는 h_j의 컴패니언 커미티(h_j를 제외한 H의 다른 모든 앙상블 멤버)가 추정한 실제 출력이다. 마지막으로 RMSE(Δx_u)의 상대적 개선을 최대화하는 레이블이 없는 예제 x~j가 컴패니언 커미티 H_j에 의해 레이블을 부여받은 가장 관련성 있는 예제로 선택된다.

<그림 4-2>, <그림 4-3>에서 기술되어 있는 RBFNN함수는 회귀분석기의 학습에 사용되는 방사기저함수(RBF)를 이용하는 2단계 학습 알고리즘으로 Schwenker & Kestler(2001)에서 제안한 다변량 가우시안 방사기저함수(g)를 활성 함수(activation function)로 사용하여 회귀분석기를 학습하는 알고리즘이다. 1단계에서는 RBF의 중심은 민코스키(Minkowski) 거리를 사용하여 k-평균 군집화를 수행하여 결정된다. 가우시안 중심 세트는 L에서 무작위로 선택한 학습 예제로 초기화된다. j번째 RBF 뉴런(σ_j)의 너비는 센터 c_j와 가장 가까운 두 가우시안 중심 사이의 평균 민코스키(Minkowski) 거리에 α를 곱하여 중심 사이에서 중첩되는 정도를 조정한다. 2단계에서는 제한된 레이블이 있는 예제에 가장 근접한 출력층의 가중치 W를 행렬 의사-역(matrix pseudo-inverse) 기법을 이용하여 결정한다. 이 때 $W = H + T$ 이고, 여기서 T는 학습

예제의 목표 출력이고 식(4-3)의 H는 활성 행렬(activation matrix)이다.

$$H_{ij} = g(x_i; c_j, \sigma_j, p) = g(\|x - c_j\|_p / \sigma_j) = exp(-\frac{\|x - c_j\|_p^2}{2\sigma_j^2}) \quad (4\text{-}3)$$

학습 과정에서 경사 하강 오차 역전파(gradient-descent error backpropagation) 학습 방법은 계산 부하를 피하기 위해 고려하지 않았다. 반면에, W의 직접적인 계산은 더 쉽고 네트워크의 즉각적인 훈련을 제공하기 때문에 효율적으로 새로운 레이블이 지정된 예제를 사용하여 회귀분석기 성능을 개선할 수 있게 한다.

2. 연구 모형

본 연구에서 준지도 회귀 기법에 의한 예측 성능 개선 효과를 검증하기 위한 연구모형은 에세이1에서 베이지안 모형 평균화 기법을 이용하여 선별한 21개의 월 단위 변수로 구성된 모형(식 3-8)과, 40개의 15일 단위 변수로 구성된 모형(식 3-9)이다. 에세이1의 선행연구에 기반한 비교모형 대신 베이지안 모형 평균화에 의한 모형을 선택한 이유는 첫째, 구성된 변수가 대부분 작물 단수와의 관계에서 유의한 것으로 나타났기 때문이다. 이는 준지도 학습에 사용되는 예측모형에 유의하지 않은 변수에 의한 부정적인 영향을 최소화하기 위한 것으로 예측 성능 평가에서 준지도 회귀 기법에 의한 순수한 효과를 반영하기 위함이다. 둘째 이유는 샘플 수가 부족한 상황에서 모형의 복잡도가 준지도 회귀 기법의 성능에 영향을 미치는지 검증하기 위해 변수의 수가 많은 연구 모형의 선정하는 최적의 방법이 베이지안 모형 평균화 방법이기 때문이다. 유의한 변수를 많이 포함하는 모형을 만들기 위해서는 월 단위보다 짧은 기간의 집계변수가 고려되어야 하고, 이 때 다수의 유의한 변수로만 구성된 복잡한 모형을 만들기 위해서는 베이지안 모형 평균화가 가장 효과적인 방법이라고 판단하여 에세이1의 15일 단위 변수로 구성된 모형을 채용하였다.

연구 모형 식(4-1), 식(4-2)에서 사용된 변수 중 tm_avg는 평균기온, tm_min은, 최저기온, tm_max는 최고기온을 의미한다. hm_avg는 월 평균상대습도, pr_sum은 월 총강수량이고, ss_dura는 월 총일조시간을 의미한다. 변수명 끝의 숫자의 의미는 해당 월을 의미하고, a·b는 월의 상반기·하반기를 의미한다. 비교모형에 포함된 변수 중 기상 변수는 예측하려는 단수와 같은 연도에 관측된 변수들만 고려하여서 종속변수와 동일한 t시점의 관측치이며, 전기 단수 변수는 t-1시점의 종속변수이기 때문에 Y_{t-1}로 표시하였다. 비교 모형의 식은 선형회귀모형 추정을 고려하여 변수의 추정 계수를 포함하도록 표현하였으나 SVR, k-NN과 같은 머신러닝 기법의 경우 설명변수별로 계수를 추정하는 것이 아니기 때문에 해당 변수를 입력변수로 사용하는 모형이라는 의미로 식(4-1), 식

(4-2)을 해석해야 한다.

$$Yt = Yt-1 + tm_avg4t + tm_avg8t + tm_avg9t + tm_avg10t + tm_max4 + tm_max5 + tm_max7 + tm_min4t + tm_min5t + tm_min8t + hm_avg3t + hm_avg6t + hm_avg10 + ss_dura5t + ss_dura10t + pr_sum1t + pr_sum6t + pr_sum7t + pr_sum8t + pr_sum9t \quad (4-1)$$

$$Yt = Yt-1 + tm_avg1at + tm_avg2bt + tm_avg3at + tm_avg4bt + tm_avg5at + tm_avg7at + tm_avg7bt + tm_avg10at + tm_max1at + tm_max2bt + tm_max3at + tm_max3bt + tm_max7bt + tm_max8at + tm_min1at + tm_min3at + tm_min3bt + tm_min5at + tm_min8at + tm_min8bt + tm_min9at + tm_min9bt + tm_min10bt + hm_avg2at + hm_avg6at + hm_avg10at + hm_avg10bt + pr_sum1bt + pr_sum3at + pr_sum4at + pr_sum6at + pr_sum7bt + pr_sum8at + pr_sum10at + ss_dura1at + ss_dura5at + ss_dura6bt + ss_dura8bt + ss_dura9bt \quad (4-2)$$

3. 분석 자료

본 연구에서는 사용하는 쌀 단수 자료와 기상 관측 자료는 각각 통계청과 기상청에서 제공하는 공식자료를 사용하였다.

쌀 단수 데이터는 통계청 농작물 생산조사 자료에 1965년부터 최근까지 시·군 단위와 광역도 단위로 집계되고 있으나 1977년부터 1995년까지의 19년간은 시·군 단위 단수는 집계되지 않고 광역도 단위 단수만 집계되었다. 따라서 본 연구에서는 단수 데이터의 연속성과 과거 1980년대 이전의 쌀 품종과 1990년대 이후의 쌀 품종의 차이로 인한 기상 요인의 영향 차이를 고려하여 1997년부터 2018년까지의 연도별로 집계된 157개 시·군과 강원, 경기, 경남, 경북, 전남, 전북, 제주, 충남, 충북의 9개 광역도 단수 데이터를 사용하였다.

기상 관측 자료는 기상청의 기상자료개방포털에서 1973년 이후의 기상관측소 관측 데이터를 제공하고 있다. 제공되는 기상 요인 중 본 연구에서는 기온, 습도, 강수량, 일사량에 대한 일별 관측치 데이터를 사용하여 월 단위 최고온도, 최저온도와 같은 파생 변수를 생성하였다. 기상 데이터의 경우 쌀 단수 데이터의 시작 시점인 1997년에는 68개의 관측소가 운영되었으나 최근까지 꾸준히 증가하여 2018년에는 전국에 95개의 관측소가 운영되고 있다. 따라서 연도별로 관측소 수가 다른 문제와 68~95개의 관측소 관측치를 단수 데이터가 있는 전국 157개 지역을 대표하는 기상 정보가 될 수 있도록 매핑이 필요한 문제가 있다. 본 연구에서는 매년 157개 시·군 지역의 GIS 중심 좌표에 대하여 인근의 최근접한 기상 관측소 3개소의 GIS 좌표와의 거리 역수의 가중평균으로 계산된 기상 관측치를 각 지역의 기상 정보로 간주하였다. 기상 변수의 고려 범위는 일반적인 경우 쌀의 이앙기가 5월이고 수확기가 10월~11월인 점을 고려했을 때 기상이 쌀의 생육에 직접적인 영향을 줄 수 있는 5월~10월의 기상 변수만 고려될 수 있는 점과 기존 연구들에서도 대부분 5월~10월 기상 변수만 사용되고 있다. 그러나 베이지안 모형 평균화 기법에 의한 모형의 경우 선행연구의 고려 범위를 넘어서 1월~4월의 기상 변수도

후보 변수로 선정되었다. 이는 1월~4월 기상 조건에 의해 병해충과 관련한 요인이 영향을 받거나 1월~4월 기상 조건이 5월~10월 기상과 상관관계를 가질 수 있는 가능성이 존재하기 때문으로 본 연구에서는 이앙기 이전인 1~10월 기상 요인을 예측 모형에 사용하였다.

<표 4-1> 준지도 회귀 쌀 단수 예측 모형 사용 변수

변수	변수 설명	변수 범위
쌀 단수	단위면적(ha)당 쌀 생산량(kg)	
전기 쌀 단수	1년 전 시점의 쌀 단수	
평균기온	해당월의 일평균 기온의 평균값	1월~10월
최고기온	해당월의 일최고 기온의 평균값	1월~10월
최저기온	해당월의 일최저 기온의 평균값	1월~10월
누적일조시간	해당월의 일별 일조시간의 합계	1월~10월
강수량	해당월의 일별 강수량의 합계	1월~10월
상대습도	해당월의 일별 상대습도의 평균값	1월~10월

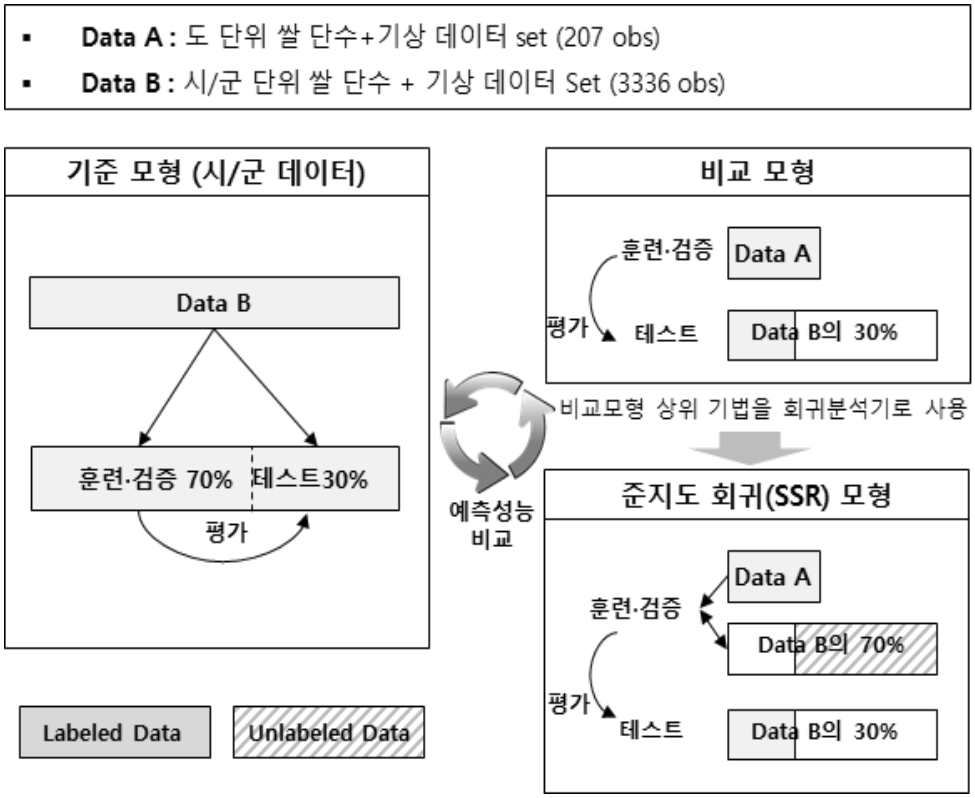
4. 준지도 회귀를 이용한 학습과 예측 성능 평가 방법

준지도 회귀 기법을 이용한 쌀 단수 예측 모형의 학습과 예측 성능 평가를 위해 광역도 단위 데이터 세트와 시·군 단위 데이터 세트의 두 가지 종류의 데이터를 사용한다. 이와 같이 넓은 지역 범위의 광역도 단위 데이터와 광역도 하위의 시·군 단위로 구분한 이유는 쌀을 제외한 대부분 작물의 단수 데이터가 광역도 단위로만 수집되고 있는 상황에서 준지도 회귀 기법을 이용하여 단수가 측정되지 않는 시·군 단위 지역의 기상 변수를 레이블이 없는 데이터로 활용하여 예측 모형을 개발했을 때 준지도 회귀 기법을 쓰지 않고 예측모형을 개발할 때보다 예측 성능이 더 나아질 수 있는지를 검증하는 것과 동시에 레이블이 없는 시·군 단위 지역에 준지도 회귀 방법을 이용한 단수 예측치를 레이블로 부여했을 때 예측에 의해 부여된 레이블 즉, 시·군 단위에서는 측정되지 않았지만 준지도 회귀 모형이 해당 지역의 기상 변수만으로 예측한 단수 데이터가 신뢰할만한가를 판단하기 위해서이다.

이와 같이 준지도 회귀 방법으로 예측한 시·군 단위 단수 예측치의 정확도를 평가하기 위해서는 시·군 단위의 단수가 측정된 데이터가 필요한데, 국내에서 집계된 통계 자료에서는 쌀 작물이 유일하게 시·군 단위의 단수 데이터가 수집되었기 때문에 본 연구의 대상 작물로 선정하였다.

데이터 세트(Data A)는 전국 9개 도의 쌀 단수와 해당 도의 기상 변수로 구성되었고 207개의 관측치를 가진다. 시·군 단위 데이터 세트(Data B)는 전국 157개 시·군의 쌀 단수와 해당 시·군의 기상 변수로 구성되었고 3,336개의 관측치를 가진다. <그림 4-4>은 데이터가 풍부한 상황을 의미하는 시·군 데이터를 이용하는 기준 모형, 데이터가 부족한 상황을 의미하는 도 단위 데이터의 비교모형, 레이블이 있는 데이터는 부족하지만 레이블이 없는 데이터는 풍부한 준지도 회귀 모형의 3가지 모형의 학습(Training), 검증(Validation), 테스트(Test)에 사용되는 데이터와 비교 모형과 준지도 회귀 모형의 관계를 나타내는 도면이다.

기준 모형은 시·군 단위 데이터(Data B)를 모두 사용하며 전체 데이터의 70%인 2,335개의 데이터를 다시 모형의 훈련에 필요한 데이터 80%, 하이퍼파라메터(hyperparameter) 조정을 위한 검증 데이터 20%로 분리하여 사용하였다. 테스트 데이터로 남겨둔 30%의 1,001개 데이터는 최종 학습 모형의 예측 성능 평가를 위한 테스트 데이터로 사용하였다. 비교 모형은 207개의 관측치를 가지는 광역도 단위 데이터(Data A)를 사용하여 Data A의 90%는 모형 학습용으로 10%는 하이퍼파라메터 검증용으로 사용하였다. 최종 학습 모형의 예측 성능 평가는 3,336개 관측치를 가지는 시·군 단위 데이터 중 30%인 1,001개 데이터를 사용하였다. 준지도 회귀 모형은 비교 모형과 동일하게 207개 관측치를 가지는 광역도 단위 데이터(Data A)를 Data A의 90%는 모형 학습용으로 10%는 하



<그림 4-4> 준지도 회귀 모형 검증을 위한 데이터 할당 방법

이퍼파라메터 검증용으로 사용하였다. 준지도 회귀 모형의 학습에 사용되는 레이블이 없는 데이터는 시·군 단위 데이터(Data B) 중 비교 모형의 테스트 데이터로 사용된 30%를 제외한 나머지 70%의 2,335개 데이터를 사용하였다. 실제로는 시·군 단위로 측정된 단수 레이블을 가지는 데이터이나 연구 목적에 맞게 레이블 정보를 제거하여 레이블이 없는 데이터로 만들었다. 준지도 회귀 기법에 의한 최종 학습 모형의 예측 성능 평가는 비교모형과 동일한 데이터 세트로 시·군 단위 데이터 중 30%인 1,001개 데이터를 사용하였다.

기준모형과 비교모형의 예측모형 학습에 고려된 기법은 선형회귀모형으로는 최소제곱법(Ordinary Least Squares: OLS)을 이용하였고, 머신러닝 기법으로는 서포트벡터회귀(Support Vector Regression: SVR), k-최근접이웃(k-nearest neighbors), 인공신경망(Artificial Neural Network: ANN) 기법을 사용하였다. OLS의 경우 모수적 방법론으로 최적화가 필요한 하이퍼파라메터 조정 과정이 필요 없기 때문에 학습용(Training) 데이터와 검증용(Validation) 데이터를 분리하지 않고 테스트용 데이터를 제외한 나머지 데이터 전체를 학습용 데이터로 사용하였다. 반면 나머지 SVR, k-NN, ANN의 경우 테스트용 데이터를 제외한 데이터를 다시 학습용 데이터와 검증용 데이터로 나누어 하이퍼파라메터를 최적화하였다.

준지도 회귀 모형은 Co-Training 방식이 적용되기 위해서는 2개 이상의 회귀분석기를 지정해야 하는데 본 연구에서는 비교 모형의 예측 성능 평가에서 우수한 성능을 보여주는 상위 기법들을 준회귀 모형의 Co-Training을 위한 회귀분석기로 사용하였다.

지금까지는 광역도 단위의 207개 데이터(Data A) 전체를 사용하여 준지도 회귀 모형을 학습하고 기준 모형, 비교 모형과 예측 성능을 비교하는 방법을 설명하였는데 대부분의 준지도 학습과 관련한 연구들에서는 학습에 사용하는 레이블이 있는 데이터의 수에 의한 준지도 학습 기법의 예측 성능을 평가하고 있다. 레이블이 있는 데이터가 얼마나 적어질 때까지 만족할만한 예측 성능을 확보할 수 있는가에 대한 검증을 위해 이와 같은 분석 과정이 추가된다. 본 연구에서는 레이블이 있는 데이터

(Data A) 207개를 100%(207개), 75%(155개), 50%(104개) 25%(52개)로 줄이면서 준지도 회귀 모형의 예측 성능을 검증하였다. 레이블 데이터의 크기를 줄이는 방법은 랜덤 샘플링으로 해당 비율에 맞게 추출하는 방법을 사용하였다. 이와 같이 레이블 데이터의 수를 줄이면서 예측성능을 평가하기 때문에 앞서 설명한 40개의 설명변수를 가지고 있는 복잡도가 높은 예측 모형이 50개, 100개 수준의 레이블 데이터만 있는 상황에서도 준지도 학습 방법으로 예측 성능을 높일 수 있는지 확인할 수 있다.

모형의 예측 성능 평가를 위해 사용된 지표는 평균제곱근오차(Root Mean Square Error: RMSE), 평균절대백분율오차(Mean Absolute Percentage Error: MAPE), 실측치와 예측치의 R^2(실측치(x축)와 예측치(y축) 산포도에 대한 R^2값) 3가지 지표를 사용하였다.

본 연구에서는 준지도 회귀 기법을 수행하기 위해 R 프로그래밍 언어와 R CRAN에서 제공하는 ssr 패키지(https://cran.r-project.org/web/packages/ssr/vignettes/ssr-package-vignette.html)를 이용하였다.

제 4 절 분석결과

207개의 레이블 데이터와 2,335의 레이블이 없는 데이터를 이용하는 준지도 회귀 모형의 예측 성능을 2,335개의 레이블 데이터를 모형 학습에 이용하는 기준 모형과 207개의 레이블 데이터만 모형 학습에 이용하는 비교 모형의 예측 성능과 비교한 결과는 <표 4-2>와 같다.

모형 학습에 시·군 단위의 많은 데이터를 사용하는 기준 모형의 경우 월 단위 변수 모형과 15일 단위 변수 모형 모두에서 RMSE, MAPE, R^2 세 지표에서 일관되게 SVR, k-NN, OLS 순으로 예측 정확도가 높았고 SVR, k-NN은 월 단위 모형과 15일 단위 모형의 예측 정확도가 유사하게 나타났다. 반면 OLS의 경우 월 단위 모형에서의 예측 정확도가 k-NN과 차이가 큰 편이었으나 15일 단위 변수 모형에서는 차이가 크지 않았다. 데이터가 충분히 많은 상황에서는 기계 학습 기법에서는 모형의 복잡도가 큰 영향을 받지 않지만 OLS의 경우에는 복잡도가 큰 모형에서 예측 성능이 더 개선되었다.

모형 학습에 적은 데이터를 사용하는 비교 모형의 경우 월 단위 모형에서는 100%, 75%의 레이블 데이터까지는 SVR이 가장 RMSE, MAPE, R^2 세 지표 모두에서 가장 높은 예측 정확도를 보였으나 50%, 25% 레이블 데이터에서는 OLS가 더 나은 예측 정확도를 보였다. 또한 OLS는 25% 레이블 데이터를 제외하면 50%~100% 레이블 데이터에서의 예측 성능이 크게 나빠지지 않았다. 반면 SVR, k-NN, ANN은 데이터의 양이 줄어들수록 예측 성능이 나빠지는 경향을 보였다. 15일 단위 모형에서는 RMSE, MAPE의 경우 SVR이 모든 크기의 레이블 데이터 범위에서 가장 나은 성능을 보여주었다. 그러나 R^2는 100%, 75%, 25% 레이블 데이터에서는 OLS보나 앞섰으나 50% 레이블 데이터에서는 OLS의 예측 성능이 더 나았다. 25% 레이블 데이터를 제외하면 50%~100% 레이블 데이터에서의 예측 성능이 크게 나빠지지 않는 OLS의 특징과 데이터 수가 줄어들수록 예측 성능이 떨어지는 기계 학습 기법의 특정은 월 단위 모형과 동일하였다. 15일 단위의 복잡한 모형의 경우 25% 레이블 데이

터로 학습하는 경우 OLS의 예측 성능이 급격하게 나빠지는 것으로 나타났고 SVR도 예측 정확도의 감소 폭은 컸으나 월 단위 변수와 비교하면 비슷한 수준이었다.

준지도 회귀 모형의 예측 성능은 비교 모형 중 가장 우수한 성능을 보여주는 경우와 비교했을 때 RMSE, MAPE, R^2 세 지표 모두에서 전반적으로 앞서는 것으로 나타났다. 월 단위 모형에서는 100%, 75% 레이블 데이터에서는 비교 모형 대비 준지도 회귀 모형의 예측 성능의 차이가 크게 앞서지는 않았으나 50% 레이블 데이터에서는 차이가 크게 나타났고 25% 레이블 데이터에서는 다시 소폭 앞서는 경향을 보였다. 15일 단위 모형에서도 월 단위 모형과 유사한 예측 정확도 추세를 보였다. 모형의 복잡도가 높은 15일 단위 모형에서 25% 레이블 데이터를 사용한 경우에 예측 성능 감소 추세가 큰 것은 준지도 회귀 모형에서도 비교 모형과 유사한 패턴을 보였다.

준지도 회귀 모형의 예측 성능을 기준 모형과 비교를 해보면 RMSE, MAPE, R^2 모든 지표에서 월 단위 모형에서의 기준 모형과 100% 레이블 데이터의 준지도 회귀모형 간의 예측 오차의 차이보다 15일 단위 모형에서의 예측 오차의 차이가 더 작은 것으로 나타났는데 데이터가 충분히 많은 상황에서의 월 단위 모형과 15일 단위 모형의 예측 성능이 차이가 없는 것을 고려했을 때 준지도 회귀 모형의 경우 모형의 복잡도가 높아지더라도 유의한 설명 변수를 많이 추가하는 것이 예측 성능 개선에 도움이 될 수 있음을 시사한다.

<표 4-2> 준지도 회귀 모형과 기준 모형 및 비교 모형의 예측 성능 비교

지표	기법	월 단위 변수 (21 Variables)					15 일 단위 변수 (40 Variables)				
		기준 모형	Labeled 100%	Labeled 75%	Labeled 50%	Labeled 25%	기준 모형	Labeled 100%	Labeled 75%	Labeled 50%	Labeled 25%
RMSE	OLS	27.20	28.18	28.6	29.61	32.12	25.91	27.15	27.15	28.61	47.9
	SVR	**21.77**	25.55	27.49	30.17	33.27	**21.58**	23.86	25.55	27.94	32.28
	kNN	24.64	31.04	32.22	34.37	39.03	24.83	32.09	33.36	35.83	41.28
	ANN	31.21	38.25	40.71	42.95	47.25	33.42	37.21	38.35	39.5	45.55
	SSR		**24.88**	**26.47**	**27.58**	**30.23**		**23.67**	**24.69**	**25.91**	**31.56**
MAPE	OLS	4.2%	4.5%	4.6%	4.8%	5.1%	3.9%	4.1%	4.3%	4.6%	7.6%
	SVR	**3.4%**	4.0%	4.4%	4.9%	5.3%	**3.4%**	3.7%	4.0%	4.4%	5.2%
	kNN	3.8%	5.1%	5.3%	5.7%	6.6%	3.9%	5.2%	5.5%	5.9%	7.0%
	ANN	5.0%	6.2%	6.6%	7.1%	7.9%	5.3%	6.1%	6.4%	6.6%	7.4%
	SSR		**3.9%**	**4.2%**	**4.4%**	**4.8%**		**3.7%**	**3.9%**	**4.1%**	**4.9%**
R^2	OLS	60.7%	58.9%	58.0%	55.8%	50.3%	64.3%	62.4%	62.6%	60.3%	30.5%
	SVR	**74.9%**	65.4%	59.9%	51.5%	41.1%	**75.1%**	69.8%	65.2%	58.5%	46.2%
	kNN	68.3%	55.6%	51.9%	46.9%	41.2%	67.5%	52.3%	49.3%	43.6%	38.3%
	ANN	50.9%	36.5%	31.8%	27.4%	23.8%	51.2%	37.8%	40.9%	34.3%	14.2%
	SSR		**67.6%**	**62.6%**	**59.4%**	**51.6%**		**70.5%**	**67.5%**	**64.8%**	**50.3%**

제 5 절 소결

지금까지 레이블이 있는 데이터가 부족한 상황에서 레이블이 없는 데이터를 모형의 학습에 이용하여 예측 성능을 높이는 준지도 회귀 기법을 쌀 단수 예측 모형에 적용했을 때의 결과를 살펴보았다. 작물 단수의 중요한 영향 요인인 기상 데이터는 측정 및 보간 기술의 발전으로 조밀한 범위의 지역 단위에서도 데이터 수집이 가능한 반면 작물의 단수 데이터의 측정을 위해서는 많은 비용이 발생하기 때문에 많은 수의 데이터 수집이 현실적으로 쉽지 않다. 이와 같이 레이블 데이터와 레이블이 없는 데이터의 불균형이 심한 작물 단수 데이터의 상황은 준지도 학습 방법이 적용될 수 있는 대표적인 사례로 볼 수 있다. 본 연구에서는 데이터가 충분한 상황에서의 예측 모형만큼의 성능까지는 따라갈 수 없지만 종속변수에 유의한 영향을 미치는 설명변수를 학습 데이터 수의 50%를 넘지 않는 선에서 최대한 많이 포함시키는 모형을 설계하고 변수 간의 비선형 관계를 잘 포착할 수 있는 기계 학습 기법을 회귀분석기로 포함하는 준지도 회귀 모형을 사용할 경우 본 연구의 쌀 단수의 예측 상황에서는 MAPE 기준으로는 0.5%point 차이, R^2 기준으로는 5% 이내 수준에서 예측 성능을 개선할 수 있음을 확인하였다. 준지도 회귀 기법을 이용할 경우 레이블 데이터가 100개인 상황에서도 2000개 이상의 데이터를 사용하는 기준 모형의 예측 성능과도 15일 단위 변수를 사용하는 경우에는 MAPE 0.8%point, R^2 10% 정도의 차이만 보였다는 점에서 준지도 학습 방법이 예측을 목적으로 하는 경우에는 좋은 선택이 될 수 있을 것이다.

쌀 단수 예측 모형에 한하여 모형의 복잡도는 데이터 수가 설명변수의 개수보다 2배 이상 많은 상황에서는 모형의 복잡도가 큰 것이 예측 성능에서는 득이되는 것을 확인하였다. 반대로 레이블 데이터 수의 제약이 큰 경우에는 모형의 복잡도가 크지 않도록 모형을 설계해야 하는 것이 바람직하다.

실용적인 관점에서는 시·군 단위에서는 측정되지 않는 작물의 단수 통계치를 준지도 회귀 모형의 예측치로 추정하는 것이 가능해지기 때문

에 시·군 단위의 쌀 단수 또는 생산량에 대한 통계가 필요할 경우 준지도 회귀 모형을 활용할 수 있을 것이다.

본 연구의 한계점으로는 레이블이 있는 데이터가 나머지 레이블이 없는 데이터를 대표할 수 있는지에 대한 통계적 방법론을 이용한 검증이 선행되지 않은 것으로 대표성에 대한 검증 방법을 이용하여 연구의 결과를 더 보완할 수 있을 것이다. 또한 쌀 작물 외에 다른 작물의 경우 시·군 단위의 단수 데이터가 없어서 준지도 회귀 방법이 시·군 단위의 기상 정보를 토대로 해당 지역의 단수를 예측하는 것이 신뢰할만한지는 검증이 어려운 한계를 가진다. 그렇지만 도 단위로 수집된 단수 데이터에 레이블이 없는 시·군 지역의 기상 데이터를 준지도 회귀 모형에 사용할 경우 예측 성능이 어느 정도 개선되는지는 쌀 외의 다른 작물 데이터로도 확인이 가능하다. 쌀 외의 다른 작물은 콩, 고구마, 양파, 마늘 4개 작물에 대하여 준지도 회귀 방법을 적용한 결과는 <부록 5>과 같이 도출되었다. 도 단위 집계 데이터 중 10%는 평가용 데이터로 제외시킨 후 나머지 90%의 데이터로 준지도 회귀 모형과 OLS, SVR을 비교하였다. 작물별로 예측 성능 지표별로 조금씩 차이는 있으나 준지도 회귀 방법을 이용할 경우 OLS, SVR을 단독으로 사용하는 경우와 비교했을 때 개선 효과가 있는 것이 확인되었다. 콩의 경우 도 단위 데이터 전체로 준지도 회귀를 했을 때에는 SVR과 비교하여 RMSE 개선은 뚜렷하지 않았으나 MAPE는 0.9%point, R^2는 3.7%point 개선 효과가 있었고 학습용 데이터가 더 부족해지는 상황인 90%의 절반인 45%의 데이터로 준지도 회귀를 했을 때 모든 지표에서 SVR 대비 성능 개선이 뚜렷하였다. 특히 R^2는 12.2%point나 개선되는 효과를 보였다. 고구마, 양파, 마늘 단수 예측 모형에서는 RMSE에서는 소폭 개선되었고 MAPE에서는 양파, 마늘은 90% 데이터 학습 시에만 개선되었고 45% 데이터에서는 개선효과가 없었다. 고구마의 경우는 90% 데이터에서 준지도 회귀 모형이 다소 떨어지는 성능을 보여주기도 하였다. 그러나 R^2에서는 고구마 45% 데이터만 제외하고 고구마, 양파, 마늘 모두에서 개선 효과가 나타났다.

향후 연구에서는 본 연구에서는 적용되지 않은 k-fold 교차 검증을

통해 샘플링 편향로 인한 영향을 최소화할 필요가 있다. 또한 쌀 작물 외에 시·군 단위 단수가 측정되지 않는 작물의 경우 준지도 회귀 모형에 의한 시·군 단위 예측치의 검증을 위해 위성 이미지에서 계산되는 식생지수(Normalized Difference Vegetation Index: NDVI) 변수와의 상관관계를 분석하는 방법을 시도해볼 수 있을 것이다. 작물 단수 데이터 외에도 레이블 데이터와 레이블이 없는 데이터 간의 불균형이 크면서 레이블 데이터가 연속형 변수인 경우에는 준지도 회귀 방법은 좋은 대안이 될 수 있는 만큼 다양한 분야에서의 실증 연구에 적용될 수 있을 것이다.

제 5 장 쌀 단수의 조기 예측을 위한 두 접근법에서의 기계 학습 방법 적용

제 1 절 서론

작물의 단수를 조기에 예측하여 생산량의 과잉 또는 부족에 대한 적절한 정책을 수립하는 것에 대한 필요성은 오래전부터 제기되고 있다. 그에 반해 작물 단수의 조기 예측을 위한 방법에 대해서는 다양하게 연구되지 못하고 있다. 전 세계적으로 기후변화에 대응의 중요성이 커지면서 글로벌 기상 모형을 구축하고 이를 이용한 연구들이 활발하게 수행되고 있으나 작황의 조기 예측을 다루는 많은 연구들에서 미래의 기상 정보를 사용하지 않고 있다(Alvarez et al., 2009; Everingham et al., 2016; Mathieu & Airesa, 2018; Qian et al., 2009) 기후 또는 계절 내 기상 예측은 작황 사전 예측에도 사용될 수 있으며(Ceglar et al., 2017), 농업 분야에서는 실제로 다수의 국가에서 지역별 작물 수확량 예측 시스템에서 기상 예측 모형을 하위 시스템으로 사용하고 있다(Cantelaube & Terres, 2005; Capa-Morocho et al., 2016; Challinor et al., 2005; Iizumi et al., 2013; Sivakumar et al., 2000). 그러나 기상 예측 모형에 대한 연구가 발전하면서 기상 예측 시스템도 지난 20 년간 지속적으로 개선되었지만(Doblas-Reyes et al., 2013; Saha et al, 2006, 2014), 아직까지 조기 예측 기간이 2주에서 1개월 정도로 긴 리드타임을 제공하지 못하고 있다. 또한 글로벌 기상 모형이 전지구 단위의 예측을 수행하기 때문에 지역 규모에서 작물의 단수 예측을 어느 정도 향상시킬 수 있을지에 대해서는 정확한 판단이 어렵다(Peng et al., 2018).

이와 같이 작황에 대한 조기 예측이 현장에 적용되기 위해서는 지역별 예측이 가능해야 하고, 조기 예측의 리드타임이 충분히 길어야 하지만 이러한 조건들을 만족시킬 수 있는 적절한 방법이나 조건에 대한 관련 연구가 부족하고 특히 국내에는 농촌경제연구원과 같은 연구기관에서

기상 요인의 시뮬레이션에 의한 단수 예측은 연구하였으나 미래의 기상 정보가 없는 상황에서의 작물 단수의 조기 예측에 대한 연구는 수행되지 않고 있다. 또다른 측면에서는 대부분의 국가에서와 마찬가지로 농산물 작황에 대한 데이터가 연간 단위로 조사되고 있고 수확이 종료된 후 자료 집계에도 시간이 소요되고 있어 통계청에서 제공하는 단수 데이터의 현장의 활용도가 떨어지므로 조기 예측 모형의 리드타임을 줄이고 수확 시기까지의 기상 데이터를 활용하게 되면 Nowcasting의 기능도 수행할 수 있게 된다.

본 연구는 작물 단수 예측 모형의 조기 예측을 위한 방법별 차이를 비교하고 각 방법별로 최적의 예측 성능을 달성하기 위한 조건을 탐색하는 것을 목적으로 한다. 이를 위해 1개월 단위로 조기 예측의 리드타임을 늘려가며 예측을 할 때 얼마나 앞선 시점에서 신뢰할만한 예측 성능을 보이는지를 조기 예측 방법별로 평가한다.

제 2 절 선행연구 검토

작물 단수의 조기 예측에 관한 연구들에서 사용된 사전 예측 방법은 3가지 유형으로 구분된다. 첫째, 미래의 기상 데이터를 사용하지 않고 예측 시점 이전의 기상 관련 변수만 모델링에 사용하는 연구들이다. (예: Alvarez et al., 2009; Everingham et al., 2016; Matsumura et al., 2015; Mathieu & Airesa, 2018; Oettli et al., 2018; Qian et al., 2009; Schauberger et al., 2018) 이들 연구는 사전 예측 기간이 다를 경우 서로 다른 모형을 구축해야하기 때문에 리드타임별 예측 시점의 수만큼 변수 선별과 예측력 평가를 해야 하는 노력이 필요하다. 또한 미래의 기상 요인의 변동성이 반영되지 않기 때문에 현재 시점까지의 기상 패턴과 유사한 과거의 단수의 평균치에 수렴할 가능성이 높다.

둘째, 전체 작기 데이터를 이용하여 단일 예측 모형을 개발하고 예측 시점에서 사용할 수 없는 미래의 기상 변수는 과거의 기상 데이터의 평균 또는 중위값을 이용하는 연구가 있다(예: Ceglar et al., 2017; Gouache et al., 2015). 이들 연구는 미래의 기상 요인을 과거의 평균으로 가정하기 때문에 기상 요인이 평균을 벗어날 경우 예측치의 오차가 커질 가능성이 높다.

셋째, 단일 예측 모형을 개발은 동일하고, 예측 시점에서 사용할 수 없는 미래의 기상 변수는 기상 예측 모형의 예측치를 사용하는 연구가 있다(예: Iizumi et al., 2013; Peng et al., 2018). 그러나 기상 예측 모형은 일반적으로 글로벌 수준의 모형이라 지역적인 편차가 반영되기 어려운 한계를 가지고 있다. 또한 기상 예측 모형에서는 기온과 강수량만 예측치를 제공하기 때문에 다른 기상 요인을 예측 모형에서 사용하기 위해서는 해당 기상 요인에 대한 예측치를 다른 방법을 이용해서 구해야 하는 한계도 가지고 있다.

그러나 이와 같이 서로 다른 유형의 사전 예측 방법들을 예측 성능의 관점에서 비교한 연구는 많지 않으며, 쌀 단수 모델에 대한 실증 분석 연구는 없는 것으로 조사되었다.

제 3 절 연구방법

1. 연구 모형

기상 요인 변수를 이용하여 쌀 단수를 조기에 예측하기 위한 방법으로 선행 연구에서 다루고 있는 접근법은 미래의 기상 변수를 사용하지 않는 모형을 만드는 접근법과 미래의 기상 변수의 관측치를 과거의 기상 데이터로부터 계산된 평균 또는 중위값으로 사용하는 접근법이 있다. 본 연구에서는 이와 같은 두 접근법에서의 단수 조기 예측 모형의 예측 성능을 비교하는 것을 기본 목표로 한다. 또한 기존의 선행 연구에서 사용하고 있는 방법보다 보다 발전된 예측 방법을 탐색하기 위해 미래 기상 변수를 사용하지 않는 접근법에서는 각 예측 시점별 개별 모형의 변수 선별에 베이지안 모형 평균화 기법을 적용하였다. 미래의 기상 변수를 사용하는 접근법에서도 선행 연구에서 사용하고 있는 방법과의 비교를

<그림 5-1> 단수 조기 예측 방법 검증을 위한 4가지 유형의 연구 모형 개요도

위해 기상 요인의 예측치를 구하는 모형에 베이지안 모형 평균화 기법을 적용하였다.

미래의 기상 변수를 사용하지 않고 예측 시점 이전의 기상 관련 변수만 모델링에 사용하는 접근법의 첫 번째 방식은 선행 연구에 기반한 예측 모형에서 1개월 단위의 리드타임으로 조기 예측의 시점이 앞당겨질 때마다 조기 예측의 시점 이후에 관측되는 기상 변수는 소거하는 방식이다. 따라서 10월 모형이 모든 영향 변수를 포함하는 모형이라면 9월 모형부터 6월 모형까지는 10월 Full 모형에서 해당 예측 시점 이후에 관측되는 변수들을 제거하여 만들었다. 이와 같은 방법으로 만들어진 미래의 기상 변수를 사용하지 않는 선행연구 기반의 모형은 식(5-1)~식(5-5)과 같다.

[10월] yield = yield_l1 + GDD5a8b + tm_dif8c9c +
tm_avg5 + tm_avg6 + tm_avg7 + tm_avg8 + tm_avg9 + tm_avg10 +
tm_max7 + tm_max8 + tm_max9 + tm_min7 + tm_min8 + tm_min9 +
pr_sum5 + pr_sum6 + pr_sum7 + pr_sum8 + pr_sum9 + pr_sum10 +
ss_dura7+ ss_dura8+ ss_dura9 (5-1)

[9월] yield = yield_l1 + GDD5a8b +
tm_avg5 + tm_avg6 + tm_avg7 + tm_avg8 + tm_avg9 +
tm_max7 + tm_max8 + tm_max9 + tm_min7 + tm_min8 + tm_min9 +
pr_sum5 + pr_sum6 + pr_sum7 + pr_sum8 + pr_sum9 +
ss_dura7+ ss_dura8+ ss_dura9 (5-2)

[8월] yield = yield_l1 + GDD5a8b + tm_avg5 + tm_avg6 + tm_avg7 + tm_avg8 +
tm_max7 + tm_max8 + tm_min7 + tm_min8 +
pr_sum5 + pr_sum6 + pr_sum7 + pr_sum8 + ss_dura7+ ss_dura8 (5-3)

[7월] yield = yield_l1 + tm_avg5 + tm_avg6 + tm_avg7 + tm_max7 + tm_min7 +
pr_sum5 + pr_sum6 + pr_sum7 + ss_dura7 (5-4)

[6월] yield = yield_l1 + tm_avg5 + tm_avg6 + pr_sum5 + pr_sum6 (5-5)

미래의 기상 변수를 사용하지 않고 예측 시점 이전의 기상 관련 변수만 모형에 사용하는 접근법 중 두 번째 방식은 예측 시점별로 선행연구에 기반한 예측 모형이 아닌 베이지안 모형 평균화를 이용하여 예측 모형을 만드는 것이다. 베이지안 모형 평균화에 의한 예측 모형은 1개월 단위의 리드타임으로 조기 예측의 시점이 앞당겨질 때마다 후보 변수의 범위를 예측 시점 과거의 변수만으로 제한하고 주어진 변수 범위 내에서 가장 설명력이 높은 모형을 예측 모형으로 정하는 것이다. 베이지안 모형 평균화에 의한 예측 모형은 선행연구에 기반한 예측 모형과 같이 선험적으로 결정되는 것이 아니기 때문에 분석 결과 파트에서 조기 예측 시점별로 도출된 예측 모형을 설명한다.

쌀 단수를 조기에 예측하기 위한 두 번째 접근법인 미래의 기상 변수를 예측하여 사용하는 접근법도 다시 두 가지 방식으로 구분하였는데, 첫 번째 방식은 기존의 선행연구에서 사용되었던 방법으로 지역별 과거 기상 데이터를 이용하여 기상 변수의 과거 평균 또는 중위수를 미래의 기상 변수로 사용하는 것이다. 기상 요인의 과거 평균을 사용할 경우 일반적인 범위를 많이 벗어나는 이상치가 존재할 경우 편향이 발생할 가능성이 있고, 중위수를 사용할 경우에는 해당 기상 변수의 분포의 왜도(skewness)가 커질수록 편향이 발생할 가능성이 높아지기 때문에 평균과 중위수 각각을 사용하는 예측 모형을 설계하였다.

쌀 단수를 조기에 예측하기 위해 미래의 기상 변수를 예측하여 사용하는 접근법 중 두 번째 방식은 미래의 기상 변수를 평균과 중위수가 아닌 예측 모형에 의한 예측치를 사용하는 것이다. 예를 들어 단수 예측 모형은 10월까지의 기상 변수를 사용하는데 9월 말에 단수 예측을 해야하는 상황에서 예측 모형에 포함된 10월 최저기온을 알 수 없는 상황이라고 가정하면 10월 최저기온 변수에 대한 예측 모형을 10월 이전의 과거 기상 변수로 만들고 이 모형에서 10월 최저기온을 예측하여 단수 예측 모형의 10월 최저기온 값으로 사용하는 것이다. 이와 같은 방식은 조기 예측 시점이 1개월 단위로 리드타임이 길어질 경우 예측해야하는 기상 변수의 수가 늘어나게 되고 각각의 기상 변수를 예측할 수 있는 예측

모형을 만들어야 하기 때문에 기존 연구들에서 설명변수를 탐색하는 방법을 이용할 경우에는 많은 시간과 노력이 필요하다. 따라서 본 연구에서는 기상 변수의 예측 모형을 가장 잘 설명할 수 있는 변수를 효율적으로 선별하기 위해 에세이1에서 소개되었던 베이지안 모형 평균화 기법을 이용하여 기상 변수의 예측 모형을 만드는 방법을 도입하였다. 예측의 대상은 평균기온, 최고기온, 최저기온 3가지 변수에 대한 예측 모형에만 베이지안 모형 평균화를 적용하였다. 강수량, 일사량, 상대습도의 경우 대기의 상태에 따른 구름의 형성과 강우의 시점 및 강수량이 가지는 불확실성이 크고 실제로 예측 모형 검증한 결과 신뢰할 수 있는 예측치를 얻지 못했기 때문에 베이지안 모형 평균화를 이용한 예측 모형에서는 제외하였다. 대신 이들 변수들은 과거의 기상 데이터로부터 평균을 구하여 예측치로 대체하였다.

2. 분석 자료

본 연구에서는 사용하는 쌀 단수 자료와 기상 관측 자료는 각각 통계청과 기상청에서 제공하는 공식자료를 사용하였다.

쌀 단수 데이터는 통계청 농작물 생산조사 자료에 1965년부터 최근까지 시·군 단위로 집계되고 있으나 1977년부터 1995년까지의 19년간은 시·군 단위 단수는 집계되지 않고 광역도 단위 단수만 집계되었다. 따라서 본 연구에서는 단수 데이터의 연속성과 과거 1980년대 이전의 쌀 품종과 1990년대 이후의 쌀 품종의 차이로 인한 기상 요인의 영향 차이를 고려하여 1997년부터 2018년까지의 연도별로 집계된 157개 시·군의 단수 데이터만 사용하였다.

기상 관측 자료는 기상청의 기상자료개방포털에서 1973년 이후의 기상관측소 관측 데이터를 제공하고 있다. 제공되는 기상 요인 중 본 연구에서는 기온, 습도, 강수량, 일사량에 대한 일별 관측치 데이터를 사용하여 월 단위 최고온도, 최저온도와 같은 파생 변수를 생성하였다. 기상 데이터의 경우 쌀 단수 데이터의 시작 시점인 1997년에는 68개의 관측소가 운영되었으나 최근까지 꾸준히 증가하여 2018년에는 전국에 95개의 관측소가 운영되고 있다. 따라서 연도별로 관측소 수가 다른 문제와 68~95개의 관측소 관측치를 단수 데이터가 있는 전국 157개 지역을 대표하는 기상 정보가 될 수 있도록 매핑이 필요한 문제가 있다. 본 연구에서는 매년 157개 시·군 지역의 GIS 중심 좌표에 대하여 인근의 최근접한 기상 관측소 3개소의 GIS 좌표와의 거리 역수의 가중평균으로 계산된 기상 관측치를 각 지역의 기상 정보로 간주하였다. 기상 변수의 고려 범위는 일반적인 경우 쌀의 이앙기가 5월이고 수확기가 10월~11월인 점을 고려했을 때 기상이 쌀의 생육에 직접적인 영향을 줄 수 있는 5월~10월의 기상 변수만 고려될 수 있는 점과 기존 연구들에서도 대부분 5월~10월 기상 변수만 사용되고 있다는 점에서 선행연구에 기반한 기본 모형은은 5월~10월 기상 변수만 사용하여 구성하였다. 기본 모형은 10월 말 시점에 10월까지의 기상 정보를 이용하여 쌀 단수를 예측하는 모형으로

식(5-1)과 같다.

<표 5-1> 조기 예측 연구를 위한 쌀 단수 예측 모형 사용 변수

변수	변수 설명	변수 범위
쌀 단수	단위면적(ha)당 쌀 생산량(kg)	
전기 쌀 단수	1년 전 시점의 쌀 단수	
유효적산온도 (GDD)	(최고온도+최저온도)/2 - 기준온도의 누적값, 기준온도보다 낮은 경우는 0의 값을 가짐	5월~9월
평균기온	해당월의 일평균 기온의 평균값	5월~10월
최고기온	해당월의 일최고 기온의 평균값	5월~10월
최저기온	해당월의 일최저 기온의 평균값	5월~10월
일교차	해당월의 일최고-최저기온 차이의 평균값	5월~10월
누적일조시간	해당월의 일별 일조시간의 합계	5월~10월
강수량	해당월의 일별 강수량의 합계	5월~10월
상대습도	해당월의 일별 상대습도의 평균값	5월~10월

한편 기상 요인의 예측 모형에서 사용하는 변수는 <표 5-2>에서와 같이 기온, 일조시간, 강수량, 상대습도 4개 요인의 6개 변수를 고려하였다. 설명변수의 집계 기간은 1년 전의 기상 패턴도 현재의 기상에 영향을 미칠 수 있을 것으로 가정하고 전년도 1월부터 현재 연도 10월까지의 기간동안의 측정치를 월 단위로 집계하였다.

<표 5-2> 기상 요인 예측 모형 사용 변수

변수	변수 설명	변수 범위
평균기온	해당월의 일평균 기온의 평균값	전연도 1월~10월
최고기온	해당월의 일최고 기온의 평균값	전연도 1월~10월
최저기온	해당월의 일최저 기온의 평균값	전연도 1월~10월
누적일조시간	해당월의 일별 일조시간의 합계	전연도 1월~10월
강수량	해당월의 일별 강수량의 합계	전연도 1월~10월
상대습도	해당월의 일별 상대습도의 평균값	전연도 1월~10월

3. 예측 모형 학습과 예측 성능 평가 방법

예측모형의 학습에 고려된 기법은 선형회귀모형으로는 최소제곱법(Ordinary Least Squares: OLS)을 이용하였고, 머신러닝 기법으로는 서포트벡터회귀(Support Vector Regression: SVR), k-최근접이웃(k-nearest neighbors) 기법을 사용하였다. OLS의 경우 모수적 방법론으로 최적화가 필요한 하이퍼파라메터(hyperparameter) 조정 과정이 필요없기 때문에 학습용(Training) 데이터와 검증용(Validation) 데이터를 분리하지 않고 테스트용 데이터를 제외한 나머지 데이터 전체를 학습용 데이터로 사용하였다. 반면 나머지 SVR, k-NN의 경우 테스트용 데이터를 제외한 데이터를 다시 80:20으로 학습용 데이터와 검증용 데이터로 나누어 80%의 데이터로 예측 모형을 학습한 후 20%의 데이터로 하이퍼파라메터를 최적화하였다. 이 때 샘플링 편향(Sampling bias)를 최소화하기 위해 5-fold 교차검증(Cross Validation)을 수행하였다.

예측모형의 학습에 사용되지 않은 테스트용 데이터를 이용한 예측모형의 학습기법별 예측 성능 비교는 두 가지 방법으로 평가하였다. 첫 번째 방법은 10-fold 교차검증 방법으로 전체 기간 1997~2018년의 데이터를 먼저 램덤 샘플링으로 10개 sub-sample로 분리한다. 10개 sub-sample은 다시 90:10의 학습·검증용:평가용 데이터로 구성되는 10개의 데이터 세트를 만든다. 10개의 데이터 세트에 대하여 90%의 데이터로 예측 모형을 최적으로 학습시킨 후 10개 세트의 10%의 테스트 데이터로 계산된 예측 성능 평가지표 평균값으로 각 학습 기법의 성능을 비교하는 방법이다.

두 번째 방법은 과거 시점의 데이터로 예측모형의 학습 및 검증을 한 후 최근 시점의 데이터를 테스트용으로 하여 예측 성능 평가지표의 평균값을 각 학습 기법별로 비교하는 방법이다. 학습 및 검증용 데이터는 1997~2011년 데이터를 사용하였고 2012~2018년 데이터는 평가용 데이터로 사용하였다. 2012년을 기준으로 한 것은 한석호(2016)의 연구에서 2012년의 경우 태풍의 영향으로 예측 오차가 커진 것으로 나타났는데,

본 연구에서도 예측력이 떨어질 수 있는 조건을 포함시키기 위하여 2012년이 테스트 데이터에 포함되도록 최근 7년 간의 쌀 단수 예측에 대한 예측 성능을 평가하도록 설정하였다. 이와 같이 과거 데이터로 학습한 후 최근 데이터로 평가하는 방법은 일반적으로 현업에서 모형을 학습은 과거 데이터로 하고 실제 예측은 미래 시점을 대상으로 하는 시계열 데이터 특성을 반영할 수 있다.

모형의 예측 성능 평가를 위해 사용된 지표는 평균제곱근오차(Root Mean Square Error: RMSE), 평균절대백분율오차(Mean Absolute Percentage Error: MAPE), 실측치와 예측치의 R^2(실측치(x축)와 예측치(y축) 산포도에 대한 R^2값) 3가지 지표를 사용하였다.

제 4 절 분석결과

쌀 단수 조기 예측을 위한 두 접근법 중 먼저 미래 기상 변수를 사용하지 않고 예측 시점별 개별 모형을 가지는 경우의 예측 성능을 평가하였다. 10-fold 교차 검증 방법에 의한 평가 결과는 <표 >와 같다. 학습 기법에서는 서포트 벡터 회귀(SVR)에 의한 예측이 대부분의 모형에서 RMSE, MAPE, R^2 세 지표 모두에서 우수한 성능을 보여었다. 선행연구 기반 모형에서 예측 시점 이전 변수를 소거한 6월말 예측 모형에서는 k-최근접이웃(k-NN) 기법의 예측 성능이 SVR보다 근소하게 앞섰다. 해당 모형은 설명변수의 수가 5개로 모형의 복잡도가 낮았는데 k-NN 기법의 경우 모형이 단순할수록 예측 성능이 좋아지는 것으로 보인다. 10월 말 예측 모형의 MAPE 예측 오차 3.2%를 기준으로 했을 때, 9월 말, 8월 말 예측 모형의 MAPE도 3.2%로 예측 성능의 변화가 거의 없었다. 7월 말 예측 모형에서는 3.6%로 소폭 감소하였으나 수확 3개월 전인 7월 말의 예측에서도 예측 오차의 증가 폭이 크지 않은 것으로 나타났다. 6월 말 예측에서는 MAPE가 4.2%로 크게 감소하였고 R^2는 10월 말에서 7월 말 예측까지 76.3%에서 70.7%로 줄어든 것과 비교했을 때 6월 말 예측에서는 59.8%로 10%point 이상 급감하였다.

베이지안 모형 평균화를 이용해서 예측 시점별로 최적의 모형을 만든 경우는 RMSE, MAPE, R^2 세 지표 모두에서 OLS, k-NN에서는 예측 시점이 빨라질수록 예측 성능이 떨어졌으나 SVR은 높은 수준의 예측 성능을 유지하였다. 10월 말 모형의 MAPE가 3.2%인 것과 비교했을 때 6월 말~9월 말 예측 모형 모두에서 3.3%로 차이가 없었다. RMSE, R^2에서도 10월 말 모형과 6월 말 모형의 예측 성능은 차이를 거의 보이지 않았다. 이와 같이 예측 시점별 모형들에서 성능 차이를 보이지 않는 것은 기계 학습 방법인 SVR, k-NN에서만 공통적으로 나타난 특징으로 OLS의 경우는 예측 시점이 빨라지면서 예측 오차도 커지는 경향성을 보였다.

<표 5-3> 예측 시점별 변수 소거 모형과 BMA 모형의 10-fold 교차검증에 의한 예측 성능 평가 결과

지표	기법	변수 소거 모형					BMA 모형				
		10 월 말 예측	9 월 말 예측	8 월 말 예측	7 월 말 예측	6 월 말 예측	10 월 말 예측	9 월 말 예측	8 월 말 예측	7 월 말 예측	6 월 말 예측
RMSE	OLS	27.54	27.68	28.21	33.05	34.92	25.70	26.09	26.17	28.62	29.45
	SVR	**21.07**	**20.84**	**21.07**	**23.41**	28.81	**20.96**	**21.18**	**21.12**	**21.03**	**21.26**
	kNN	23.92	23.50	23.31	23.81	**27.34**	23.98	23.72	23.65	23.82	23.72
MAPE	OLS	4.3%	4.3%	4.4%	5.3%	5.6%	4.0%	4.1%	4.1%	4.5%	4.7%
	SVR	**3.2%**	**3.2%**	**3.2%**	**3.6%**	4.4%	**3.2%**	**3.3%**	**3.3%**	**3.3%**	**3.3%**
	kNN	3.7%	3.7%	3.6%	3.6%	**4.2%**	3.7%	3.7%	3.7%	3.7%	3.7%
R^2	OLS	59.5%	59.1%	57.5%	41.8%	35.1%	64.7%	63.6%	63.4%	56.2%	53.7%
	SVR	**76.3%**	**76.8%**	**76.3%**	**70.7%**	55.6%	**76.6%**	**76.1%**	**76.2%**	**76.4%**	**75.9%**
	kNN	69.5%	70.5%	71.1%	69.7%	**59.8%**	69.3%	70.0%	70.2%	69.7%	69.9%

미래 기상 변수를 사용하지 않고 예측 시점별 개별 모형을 가지는 경우의 최근 7년 간의 예측 성능 평가 결과는 <표 5-3>와 같다. 학습 기법에서는 서포트 벡터 회귀(SVR)에 의한 예측이 전반적으로 우수한 성능을 보였으나 선행연구 기반 모형에서 예측 시점 이전 변수를 소거한 7월 말, 6월 말 예측 모형에서는 OLS가 근소하게 앞섰다. 예측 모형이 단순해지는 경우에는 SVR보다는 OLS가 더 나을 수 있음을 시사한다.

선행연구 기반 모형에서 예측 시점 이전 변수를 소거한 예측 모형들에서는 예측 시점이 빨라질수록 RMSE, MAPE 오차가 커지고 R^2 값도 줄어드는 추세를 보였는데 특히 7월 말 예측 시점에서부터 예측 성능이 급격히 나빠지는 것으로 나타났다. 10월 말 예측 모형의 MAPE는 3.8%이고 9월 말, 8월 말 모형까지는 4.1%, 4.2%로 0.3~0.4%point 수준에서 오차가 크게 증가하지 않은 반면 7월 말, 6월 말 모형은 5.2%, 6.0%로 예측 성능이 급격이 나빠진다. 즉 8월 말 시점에서의 조기 예측까지는 10월 말 시점에 예측치와 큰 차이를 보이지 않는 것이므로 이르면 8월 말 시점에서는 조기 예측을 통한 예측치를 어느정도는 믿을 수 있다는 말이다.

베이지안 모형 평균화를 이용해서 예측 시점별로 최적의 모형을 만든 경우에서는 RMSE, MAPE, R^2 세 지표 모두에서 예측 시점이 빨라질수록 예측 성능이 나빠지는 경향성을 보였다. 10월 말 예측 모형의 RMSE, MAPE는 선행연구 기반 모형과 비슷한 수준이었고 R^2는 BMA 모형이 1.2%point 앞섰다. 9월 말 모형에서는 RMSE, MAPE, R^2 세 지표 모두에서 선행연구 기반의 모형 대비 BMA 모형이 근소하게 앞서는 것으로 나타났다. 그러나 8월 말 예측 이전 시점의 모형에서는 BMA 모형 예측 성능이 선행연구 기반 모형보다 더 나빴다. 이는 BMA 모형에서 과적합이 발생하여 모형 학습에 사용되지 않은 새로운 데이터에 대한 예측에서 선행연구 기반의 단순한 모형보다 예측력이 떨어졌기 때문으로 추정된다. <표 5-4>의 연도별 예측 오차는 RMSE, MAPE, R^2 세 지표에서 고르게 예측 성능이 우수한 학습 기법으로 예측 모형의 오차를 구한 결과이다. 2012~2018년의 연도별 MAPE 예측 오차를 보면 과적합이 의심되

<표 5-4> 예측 시점별 변수 소거 모형과 BMA 모형의 최근 7년에 대한 예측 성능 평가 결과

지표	기법	변수 소거 모형					BMA 모형				
		10 월 말 예측	9 월 말 예측	8 월 말 예측	7 월 말 예측	6 월 말 예측	10 월 말 예측	9 월 말 예측	8 월 말 예측	7 월 말 예측	6 월 말 예측
RMSE	OLS	26.33	27.19	28.07	**33.42**	37.56	25.49	27.09	30.83	43.10	39.96
	SVR	**25.45**	26.42	**27.36**	34.64	38.72	25.61	**25.98**	29.15	37.56	**36.83**
	kNN	35.10	34.33	34.92	40.75	46.64	36.11	36.16	36.87	42.16	39.59
MAPE	OLS	4.0%	4.2%	4.4%	**5.2%**	6.0%	3.9%	4.2%	4.6%	6.8%	6.4%
	SVR	**3.8%**	4.1%	**4.2%**	5.4%	6.2%	3.9%	**3.9%**	4.4%	6.0%	**5.9%**
	kNN	5.3%	5.1%	5.5%	6.4%	7.4%	5.6%	5.7%	5.8%	6.4%	6.1%
R^2	OLS	57.6%	54.8%	52.6%	**40.8%**	42.8%	60.2%	57.8%	52.6%	15.7%	18.0%
	SVR	59.9%	57.1%	**54.8%**	40.5%	**43.3 %**	**61.1%**	**58.8%**	53.3%	33.4%	39.0%
	kNN	32.1%	35.7%	38.3%	25.6%	17.6%	27.4%	33.0%	30.0%	21.9%	22.8%
연도별 예측오차 (MAPE)	2012	5.6%	5.5%	5.4%	5.7%	5.6%	5.5%	5.4%	5.4%	5.6%	6.0%
	2013	4.3%	4.5%	5.2%	5.4%	6.6%	4.4%	4.5%	4.4%	5.8%	6.2%
	2014	3.8%	4.4%	4.0%	4.3%	6.0%	**3.2%**	3.3%	3.1%	5.6%	5.6%
	2015	3.6%	3.3%	**3.6%**	8.0%	8.0%	3.7%	3.1%	2.9%	8.2%	8.0%
	2016	3.1%	3.8%	3.8%	4.9%	7.0%	**3.2%**	**2.6%**	**2.7%**	7.2%	7.1%
	2017	**3.0%**	**3.1%**	4.1%	4.0%	5.5%	3.8%	5.6%	8.8%	**4.8%**	**4.1%**
	2018	3.5%	3.9%	**3.6%**	**3.9%**	**5.1%**	3.4%	3.2%	3.1%	5.0%	4.2%
	7년평균	3.8%	4.1%	4.2%	5.2%	6.2%	3.9%	3.9%	4.4%	6.0%	5.9%

는 수치를 확인할 수 있다. 9월 말 BMA 모형의 MAPE을 오름차순으로 나열하면 2.6%, 3.1%, 3.2%, 3.3%, 4.5%, 5.4%, 5.6%이고 9월 말 선행연구 기반 모형은 3.1%, 3.3%, 3.8%, 3.9%, 4.4%, 4.5%, 5.5%이다. BMA 모형이 선행 연구 기반 모형보다 양극단으로 치우친 것을 알 수 있다. 또한 2017년에는 선행 연구 기반 모형은 3.1%인 것에 반해 BMA 모형은 5.6%로 큰 차이를 보인다. 즉 BMA 모형은 과거 데이터로부터 2014~2016년의 단수 예측은 정확하게 할 수 있으나 2017년의 기상 변수들의 패턴은 과거 데이터에서 적절히 학습이 되지 않아 예측오차가 커진 것으로 추측된다.

다음으로 미래 기상 변수를 사용하는 접근법에서의 경우의 최근 7년간의 예측 성능 평가 결과는 <표 5-5>와 같다. 미래 기상 변수의 예측을 위한 예측 모형 선별과 최적화 절차가 복잡하여 10-fold 교차검증은 수행하지 않았다. 미래 기상 변수를 사용하는 방법에서의 단수 예측 모형은 모든 조건에서 동일하게 사용된다. 따라서 10월 말 모형은 과거 선행 연구 기반의 Full 모형으로 어떠한 기상 변수도 예측치를 사용하지 않는 경우이다. 9월 말 모형에서는 미래 기상 변수에 과거 평균 및 중위수, 예측 모형의 예측치를 사용한 경우의 예측 오차의 차이는 크지 않았다. 유일한 차이는 10월 평균 기온 변수만 기상 예측 모형에 의해 예측되기 때문에 당연한 결과로 볼 수 있다. 8월 말, 7월 말 예측 모형으로 예측 시점이 빨라지면서는 미래 기상 변수를 기상 예측 모형의 예측치를 사용한 경우의 예측 성능이 조금씩 나아지는 것을 확인할 수 있다. 7월 말 모형에서는 기상 예측치를 사용하는 경우의 RMSE, MAPE의 차이가 평균 및 중위수를 사용하는 경우와 비교하여 크지는 않았으나 R^2는 50.1%로 평균을 사용한 경우의 43.2%보다 6.9%point 높았다. 그러나 6월 말 예측 모형에서는 기상 예측치를 사용하는 경우가 평균이나 중위수를 사용한 경우와 비교했을 때 예측 성능이 더 나쁜 것으로 나타났다. RMSE, MAPE에서는 중위수를 사용한 경우가 가장 우수했고, R^2에서는 평균을 사용한 경우가 가장 우수했다.

<표 5-5> 예측 시점별 평균, 중위수, 기상 예측모형 예측치 사용에 따른 최근 7년에 대한 예측 성능 평가 결과

지표	기법	10 월 말 예측	9 월 말 예측			8 월 말 예측			7 월 말 예측			6 월 말 예측		
			mean	median	predict	mean	median	predict	mean	median	predict	mean	median	predict
RMSE	OLS	26.33	26.22	26.33	26.10	27.16	27.09	27.25	33.13	32.43	33.21	36.57	33.60	38.98
	SVR	**25.45**	25.83	25.85	**25.79**	26.42	26.44	**25.86**	31.86	31.30	**29.47**	34.58	**32.11**	33.43
	kNN	35.10	35.51	35.10	35.46	35.57	35.06	35.19	41.78	40.66	40.26	38.02	38.06	37.17
MAPE	OLS	4.0%	4.0%	4.0%	4.0%	4.1%	4.1%	4.2%	5.1%	4.9%	5.1%	5.9%	5.3%	6.2%
	SVR	**3.8%**	**3.9%**	**3.9%**	**3.9%**	4.0%	4.0%	**3.9%**	4.9%	4.7%	**4.5%**	5.5%	**5.0%**	5.4%
	kNN	5.3%	5.4%	5.3%	5.3%	5.4%	5.4%	5.4%	6.6%	6.3%	6.3%	5.9%	5.9%	5.8%
R^2	OLS	57.6%	57.6%	57.6%	58.0%	55.3%	55.1%	56.7%	40.6%	37.2%	43.4%	47.3%	44.4%	38.3%
	SVR	**59.9%**	58.6%	58.6%	**58.8%**	57.0%	56.7%	**58.8%**	43.2%	41.1%	**50.1%**	**48.6%**	45.4%	44.8%
	kNN	32.1%	33.0%	32.1%	32.3%	37.2%	37.8%	37.6%	24.4%	22.9%	20.8%	27.5%	26.3%	23.7%
연도별 예측오차 (MAPE)	2012	5.6%	5.6%	5.5%	5.6%	5.7%	5.7%	5.6%	5.8%	6.3%	6.3%	5.7%	6.1%	5.5%
	2013	4.3%	4.3%	4.4%	4.3%	4.1%	4.2%	4.1%	4.7%	4.6%	4.1%	6.0%	5.5%	5.0%
	2014	3.8%	4.1%	3.2%	4.1%	4.2%	4.1%	4.1%	4.7%	4.3%	4.6%	5.7%	5.1%	6.1%
	2015	3.6%	3.3%	3.7%	3.3%	3.4%	3.3%	3.8%	7.6%	6.8%	5.4%	7.7%	6.4%	7.4%
	2016	3.1%	3.3%	3.2%	3.4%	3.6%	3.7%	3.2%	4.4%	3.8%	3.6%	5.3%	4.3%	4.4%
	2017	3.0%	**3.0%**	3.8%	**3.0%**	**2.9%**	3.0%	**2.9%**	3.4%	3.3%	**3.2%**	4.5%	4.0%	4.5%
	2018	3.5%	3.8%	3.4%	3.7%	3.8%	3.9%	3.4%	3.4%	3.9%	4.4%	3.8%	**3.7%**	4.6%
	7년평균	3.8%	**3.9%**	**3.9%**	**3.9%**	4.0%	4.0%	**3.9%**	4.9%	4.7%	**4.5%**	5.5%	**5.0%**	5.4%

<표 5-5>의 연도별 MAPE에서는 기상 예측 모형의 예측치를 사용하는 경우에서는 7월 말~9월 말 모형에서 평균 또는 중위수를 사용하는 경우와 비교했을 때 특별하게 예측 오차가 크거나 작은 패턴은 보이지 않는 평이한 예측 오차를 보여주었다. 구체적으로 살펴보면 2015년의 MAPE는 10월말 모형보다 9월 말, 8월 말 모형의 평균 또는 중위수를 사용한 경우가 더 오차가 적은 것을 확인할 수 있는데 이는 2015년의 8월과 9월 기상 변수가 과거 평균 또는 중위수의 값을 가질 때 예측 성능이 좋아진다는 것을 의미한다. 즉 10월 말 예측 모형에서 과거 데이터로 학습한 모형에서는 10월 기상 변수가 평균적인 수준보다 높거나 낮은 특정한 범위에서 단수에 영향을 주는 관계를 포착했기 때문으로 해석할 수 있다. 7월 말 모형의 평균을 사용한 경우를 보면 2012년, 2018년에는 다른 경우보다 MAPE가 낮고, 2015년에는 MAPE가 높은 것을 알 수 있는데 8월 말 모형에서 보이던 패턴과는 다른 경향을 보이고 있다. 이는 7월 말 예측 모형에서 사용하는 8월 기상 변수에 과거 평균값을 사용할 경우 나타나는 현상으로 볼 수 있다. 2012년, 2018년은 8월 기상 변수를 과거 평균으로 두었을 때 예측 오차가 줄어들기 때문에 2012, 2018년의 기상 패턴과 단수와의 관계에서는 과거 데이터로 학습한 모형에서 과적합이 발생했다고 볼 수 있다. 반대로 2015년의 경우를 보면 8월 기상 변수를 과거 평균으로 두었을 때 예측 오차가 증가하기 때문에 2015년의 기상 패턴과 단수와의 관계에서는 평균보다는 특정한 범위에서의 관계에 초점이 맞추어졌을 것으로 추측해볼 수 있다.

제 5 절 소결

지금까지 쌀 단수의 조기 예측을 위해 기존의 선행 연구에서 다루고 있는 두 가지 접근법에서 베이지안 모형 평균화 기법을 적용하여 예측 성능을 개선할 수 있는지를 검증하였다.

두 가지 접근법에서의 4 가지 비교 모형의 최근 7년 예측 성능을 비교한 주요 결과는 <표 5-6>와 같다. RMSE, MAPE, R^2 모든 지표 전반에서 미래 기상 변수를 사용하지 않는 모형보다 미래 기상 변수를 사용하는 모형의 예측 성능이 우월한 것으로 나타났다. 기존의 선행 연구에서는 두 접근법을 직접 비교한 연구가 없었는데 본 연구에서는 두 접근법 중 평균 또는 중위수를 쓰더라도 미래 시점의 기상 변수에 예측치를 적용하는 것이 예측력의 관점에서는 유리하다는 것을 확인하였다. 특히 조기 예측의 리드타임이 길어질수록 이와 같은 효과는 더 극명하게 나타난다.

<표 5-6> 두 가지 접근법에서의 4 가지 비교 모형의 최근 7년 예측 성능을 비교

지표	미래기상변수	Full 10월 (기준)	Full 9 월	Full 8 월	Full 7 월	Full 6 월
RMSE	미사용 (변수소거)	25.45	26.42	27.36	33.42	37.56
	미사용 (신규모형)		25.98	29.15	37.56	36.83
	과거 평균값		25.83	26.42	31.30	**32.11**
	과거기반 예측		**25.79**	**25.86**	**29.47**	33.43
MAPE	미사용 (변수소거)	3.8%	4.1%	4.2%	5.2%	6.0%
	미사용 (신규모형)		**3.9%**	4.4%	6.0%	5.9%
	과거 평균값		**3.9%**	4.0%	4.7%	**5.0%**
	과거기반 예측		**3.9%**	**3.9%**	**4.5%**	5.4%
R^2	미사용 (변수소거)	59.9%	57.1%	54.8%	40.8%	43.3%
	미사용 (신규모형)		**58.8%**	53.3%	33.4%	39.0%
	과거 평균값		58.6%	57.0%	43.2%	**48.6%**
	과거기반 예측		**58.8%**	**58.8%**	**50.1%**	44.8%

미래의 기상 변수를 사용하지 않는 모형에는 예측 시점별로 모형을 개발하는 과정에서 지나치게 복잡한 모형을 만들 경우 오히려 과적합에 의해 예측 시점의 리드타임이 긴 경우에는 예측 성능이 나빠질 수 있다는 것도 분석 결과에서 나타났다. 불확실성이 큰 경우는 복잡도가 낮은 모형이 오히려 성능이 더 우수할 수 있는 것이다.

미래의 기상 변수를 사용하는 모형에서는 예측 시점이 일정 수준의 리드타임을 가지는 경우(본 연구에서는 수확 2~4개월 전)에서는 미래 기상 변수에 예측 모형에 의한 예측치를 적용하는 것이 예측 성능 개선에 도움이 될 수 있음을 확인하였다. 하지만 일정 수준 이상의 긴 리드타임(본 연구에서는 수확 5개월 전)에서의 단수 예측 모형에서는 기상 예측 모형에 의한 예측치 사용이 큰 이득을 주지 않는 것으로 나타났다. 이와 같은 현상의 원인은 3가지 측면에서 생각해볼 수 있다. 첫째, 긴 리드타임으로 인해 예측되는 기상변수가 증가함에 따라 모형의 예측 불확실성이 증가할 수 있다. 둘째, 6월 예측 모형에서 새롭게 예측을 해야 하는 7월 기상 변수의 예측이 8~10월 기상변수보다 어려울 수 있다. 셋째, 7월 기상 변수가 과거 시점과 비교하였을 때 최근 7년 사이에 구조적인 변화가 발생했을 가능성이 있다(예: 장마가 6월 말이 아닌 7월에 오는 빈도가 많아지는 경우). 두번째 원인은 <부록 7>의 기상 요인 예측 모형의 예측 성능 표를 보면 8월 기상 변수와 비교했을 때 7월 기상변수의 예측 오차가 특별히 나쁘지 않다는 점에서 가능성이 낮아진다. 세번째 원인은 기술통계 분석을 통해서 확인이 가능할 것이다.

모형의 학습 과정에서 모든 시점의 데이터를 학습에 활용할 수 있는 10-fold 교차검증에 의한 예측 성능 평가 결과 베이지안 모형 평균화를 이용한 월별 예측 모형의 예측 성능이 기대 이상으로 높은 이유는 데이터의 공간적 상관관계가 높은 것이 주요한 원인으로 추측된다. 즉 랜덤 샘플링으로 훈련·검증용 데이터와 테스트용 데이터를 분리하지만 훈련·검증용 데이터 포함된 지역과 동일한 연도의 인근 지역이 테스트 데이터에 포함될 경우, 학습된 모형은 이미 테스트 데이터와 유사한 데이터의 특징을 포착했기 때문에 예측 성능이 높게 나오는 것이다. 특히 면적이

넓지 않은 우리나라의 시·군 단위 기상 변수들은 동일한 시점에 유사한 패턴을 보일 가능성이 높기 때문에공간 상관성에서 자유로울 수 없다. 따라서 이와 같은 데이터의 특성을 고려한다면 학습용 데이터와 테스트용 데이터의 시간을 독립적으로 분리할 수 있는 Leave-one-out 교차검증 또는 Rolling Window와 같은 교차검증 방법에 의한 평가가 적절할 것이다. 그럼에도 불구하고 10-fold 교차 검증 결과에서 발견한 중요한 사실은 예측 시점이 빨라져서 7월 말 시점에 예측을 하는 상황에서도 기계 학습 기법을 이용하여 학습된 예측 모형의 경우 8~10월의 기상 정보를 전혀 활용하지 않고도 8~10월의 기상 정보를 활용하는 10월 말 예측모형과 비교했을 때 예측 성능이 크게 떨어지지 않는다는 점이다. 즉 5월에 이앙을 하는 쌀의 경우 5~7월의 기상 정보만으로도 10월 이후의 단수를 예측하는 것이 가능하다는 것을 보여주었다. 선행 연구 기반의 예측 모형에서 단순히 변수를 소거한 경우에도 이와 같은 가능성을 확인할 수 있었다. 만약 베이지안 모형 평균화를 적용한 모형과 같이 각 예측 시점에 맞는 최적 모형을 구축할 수 있다면 6월 말에 예측을 하더라도 이앙기 이전 기간을 포함하는 1~6월 기상정보로 만들어진 최적 모형의 예측 성능은 10월 말 예측한 예측치와 크게 다르지 않을 수 있다는 가능성을 보여준다. 이는 이앙기 이전의 1~4월의 기상 정보에서 그 해의 병해충이나 토양 수분의 가용성과 같은 요소들을 포착할 수 있고 이와 같이 새롭게 발견된 영향 관계로 인한 설명력이 7~10월의 기상 정보를 알지 못해서 잃게 되는 설명력을 보완할 수 있기 때문으로 추측할 수 있다. 이에 대한 검증을 위해서는 지역별 예측 오차의 차이가 10월 말 예측 모형과 6월 말 예측 모형이 상이하다는 것을 밝히는 것에서부터 시작될 수 있을 것이다.

<표 5-5>의 연도별 예측 오차 결과는 비교 기법들이 가지는 특성을 상세하게 보여준다. 따라서 연도별 예측 오차에 대한 분석을 통해 여러 가지 다양한 해석이 가능하고 특정 연도의 기상 변수 패턴과 단수와의 관계를 세부적으로 들어다볼 수 있는 좋은 정보를 제공할 수 있다. 예를 들어 연도별 예측 오차 표에서 학습한 데이터가 예측을 잘 하는 연도보

다는 예측을 잘 못하는 연도에 대하여 해당 데이터 샘플에 더 가중치를 두고 예측 모형을 학습시키는 기계 학습 방법을 적용하는 것도 고려할 수 있을 것이다. 또는 예측 시점의 리드타임을 늘려가면서 특정 시점 구간의 예측 오차에서 특이한 패턴을 보인다면 해당 시점을 대표하는 변수들의 패턴을 세부적으로 살펴보고 예측 모형 설계에서 조사된 패턴의 특징을 반영하는 것도 고려해볼 수 있을 것이다.

제 6 장 결 론

제 1 절 연구결과 요약 및 시사점

주곡인 쌀의 수급 안정화를 위해 정부는 다양한 정책을 통해 쌀 생산 체계를 유지해오고 있으며 장기적인 식량 안보의 관점에서도 쌀 자급률을 꾸준히 관리하고 있다. 그러나 최근 WTO 개발도상국 지위의 포기로 인해 쌀 생산 농가의 불안감이 확대되고 있고, 전세계적 기후변화에 의해 기상 조건의 변동성이 커지는 등 쌀 생산에서의 불확실성은 점차 커지는 추세이다. 또한 우리나라는 기후변화협약국으로 농업부문의 기후변화 적응을 위해 벼의 재배 작기 이동도 고려해야 하는 중요한 전환점을 맞이하고 있다.

안정적인 쌀 수급을 위해서는 미래 시점의 생산량 관측이 선행되어야 하며 이를 위해서는 신뢰할 수 있는 쌀 단수 예측 모형의 개발과 운영이 필요하다. 국내에서는 농촌경제연구원이 운영하는 농업관측본부에서 쌀 단수 예측 모형을 개발하여 매년 9월 말에 예측치를 발표하고 있고, 쌀 단수 예측과 관련한 연구들은 타 작물에 비해 비교적 활발하게 연구되고 있다. 그러나 최근 예측 분야에서 많이 사용되고 있는 기계 학습 방법은 아직까지 국내의 쌀 단수 예측 분야에서는 활발하게 검토되고 있지는 않고 있다. 따라서 본 연구는 기계 학습 방법을 이용하여 쌀 단수 예측 모형의 개선에 기여할 수 있는 연구 주제를 탐색하고 도출된 연구 질문을 해결하기 위한 새로운 연구 방법을 세 편의 에세이를 통해 제시하였다.

첫 번째 에세이에서는 쌀 단수 예측 모형의 변수 선별 방법으로 기존의 연구에서 사용되지 않은 기계 학습 방법인 베이지안 모형 평균화 기법을 이용하여 변수를 선별한 예측 모형이 기존의 예측 모형과 비교했을 때 예측력이 개선되는지 살펴보았다. 예측력의 평가에 앞서 베이지안 모형 평균화를 이용한 변수 선별이 월 단위로 집계되는 기상요인 변수보다 더 짧은 시간 단위로 집계되는 변수에서도 의미있는 예측 모형을 도출할

수 있는지 확인하기 위해 반기(15일), 순기(10일) 단위로 집계된 기상 요인 변수로도 예측 모형을 선별하였다. 베이지안 모형 평균화에 의해 선별된 기상 변수들은 회귀분석 결과 대부분 유의한 것으로 나타났고 예측 성능 평가에서는 10-fold 교차검증, Leave-one-out 교차검증에서 베이지안 모형 평균화에 의한 예측 모형의 성능이 기존의 선행연구 기반의 모형보다 우수한 것이 확인되었다. 기존의 통계 기반의 예측모형에서 사용되는 선형회귀모형(OLS)에 의한 예측보다는 서포트벡터회귀(SVR)와 같은 기계 학습 방법에 의해 학습된 예측 모형에서 더 나은 예측치를 제공하는 것으로 나타났다. 그러나 과거시점 데이터만으로 학습한 예측 모형으로 미래 시점을 예측한 최근 7년간 예측 성능 평가에서는 베이지안 모형 평균화에 의한 예측 모형의 성능이 기존 선행연구와 비교하여 제곱근 평균오차(RMSE), 평균절대백분율오차(MAPE)에서는 다소 앞서기는 하지만 유의미한 차이를 보이지 못하였고 실측치와 예측치의 R^2에서는 오히려 상대적으로 단순한 선행연구의 예측 모형보다 성능이 떨어지기도 하였다. 설명변수가 많은 복잡한 예측 모형은 불확실성이 큰 미래의 예측에서는 오히려 과적합 문제를 일으킬 가능성도 내포하고 있는 것이다. 이 외에도 베이지안 모형 평균화의 후보변수로 재배 기간 외의 기상 변수를 포함시켰을 때 겨울철의 기상요인이 유의한 변수로 선택되었는데 베이지안 모형 평균화 기법이 기존에 고려되지 않던 새로운 변수의 탐색에도 활용될 수 있는 가능성을 보여주었다. 그러나 새롭게 발견된 변수의 경우 선행연구 검토와 실증분석을 통한 추가적인 검증이 필요하다.

두 번째 에세이에서는 종속변수의 관측치를 가지는 레이블 데이터는 부족하고 레이블이 없는 관측 데이터는 풍부한 상황에서 레이블이 없는 데이터를 예측 모형 학습에 이용하여 예측 성능을 개선할 수 있는 방법을 쌀 단수 예측 모형에 적용하고 예측력의 관점에서 적용하는 것이 타당한지 살펴보았다. 연구 결과는 데이터가 충분한 상황에서의 예측 모형만큼의 성능까지는 따라갈 수 없지만 종속변수에 유의한 영향을 미치는 설명변수를 학습 데이터수의 50%를 넘지 않는 선에서 선별하고 변수 간의 비선형 관계를 잘 포착할 수 있는 기계 학습 기법을 준지도 학습의

회귀분석기로 포함시킬 경우 예측 성능이 개선되는 것을 확인하였다. 쌀 단수의 예측 상황에서는 준지도 회귀 방법을 사용하지 않을 경우와 비교했을 때 MAPE 기준으로는 0.5%point 차이, R^2 기준으로는 5% 이내 수준에서 예측 성능을 개선할 수 있음을 확인하였다. 레이블이 있는 시·군 데이터를 이용한 기준모형과의 비교에서는 준지도 회귀 기법을 이용할 경우 레이블 데이터가 100개인 상황에서도 2000개 이상의 데이터를 사용하는 기준 모형의 예측 성능 대비 MAPE는 0.8%point, R^2는 10% 정도의 차이만 보였다는 점에서 예측을 목적으로 하는 경우에는 레이블이 없는 데이터가 많은 상황에서는 준지도 학습 방법이 좋은 선택이 될 수 있음을 확인하였다. 쌀의 경우 평가용 데이터를 시·군 데이터를 사용하였기 때문에 시·군 단위 단수 관측치가 없는 상황에서 시·군을 대표할 수 있는 단수 예측치를 제공할 수 있는지를 평가할 수 있는데 MAPE 기준에서는 0.3~0.5%point, R^2 기준에서는 5~7%point 가량 더 큰 오차를 가지는 수준에서는 예측이 가능한 것으로 나타났다. 시·군 단위 단수 데이터가 없는 콩, 고구마, 양파, 마늘 품목에 대한 준지도 회귀 방법의 적용 결과에서도 작물별로 편차는 있었으나 레이블이 없는 시·군 기상 데이터만 추가하는 준지도 회귀 방법의 예측치에서 예측력이 개선되는 것이 확인되었다.

세 번째 에세이에서는 작물 단수의 조기 예측에 사용된 여러 방법들과 기존 연구에서 사용되지 않았던 기계 학습 기법을 적용한 방법을 비교했을 때 예측 성능이 개선되는지 평가하고, 조기 예측의 리드 타임별로 예측 성능의 차이를 비교하였다. 두 가지 접근법에서의 4 가지 비교 모형의 최근 7년 예측 성능을 비교한 결과에서 RMSE, MAPE, R^2 모든 지표 전반에서 미래 기상 변수를 사용하지 않는 모형보다 미래 기상 변수를 사용하는 모형의 예측 성능이 우월한 것으로 나타났다. 기존의 선행 연구에서는 두 접근법을 직접 비교한 연구가 없었는데 본 연구에서는 두 접근법 중 평균 또는 중위수를 쓰더라도 미래 시점의 기상 변수에 예측치를 적용하는 것이 예측력의 관점에서는 유리하다는 것을 확인하였다. 특히 조기 예측의 리드타임이 길어질수록 이와 같은 효과는 더 극명

하게 나타났다. 또한 미래의 기상 변수를 사용하는 모형에서는 예측 시점이 일정 수준의 리드타임을 가지는 경우(본 연구에서는 수확 2~4개월 전)에서는 미래 기상 변수에 예측 모형에 의한 예측치를 적용하는 것이 예측 성능 개선에 도움이 될 수 있음을 확인하였다. 하지만 일정 수준 이상의 긴 리드타임(쌀의 경우 수확 5개월 전)에서의 단수 예측 모형에서는 기상 예측 모형에 의한 예측치 사용이 큰 이득을 주지 않는 것으로 나타났다.

비록 쌀 단수 예측 모형에 국한된 실증분석 결과이나 베이지안 모형 평균화와 같은 기계 학습 방법에 의한 변수 선별 방법은 기존의 통계 기반의 작물 단수 예측에 관한 연구들에서 사용하고 있는 변수 선별 방법과는 차별점을 가진다. 본 연구에서 제안하는 방법은 작물의 단수에 영향을 미치는 요인에 대한 사전 지식이 없는 상황에서도 변수들 간의 관계에 대한 사후확률 계산을 통해서 가장 설명력 있는 변수들을 탐색한다는 점에서 데이터 주도형 모형 개발에서의 활용 가능성을 시사한다. 기상 요인의 월 단위 집계 변수를 사용한 모형과 15일, 10일 단위 집계 변수를 사용한 모형에서의 변수의 영향 구조가 상이한 점은 기상요인을 월 단위로 집계했을 때 해당 월의 특정 기간의 영향력이 반영되지 않거나 측정치가 특정 범위를 넘어가서 작물에 부정적 영향을 줄 수 있는 패턴들이 월 단위로 집계되면서 상쇄되었을 가능성을 내포하고 있다. 따라서 작물의 생육 주기를 고려하여 기상 요인의 집계 기간을 세분화하거나 측정치의 특정 범위 이탈에 대한 파생변수를 고려한다면 보다 정교한 예측 모델 개발이 가능할 수 있음을 시사한다. 이와 같이 작물 단수 예측 모형에서의 베이지안 모형 평균화 기법의 적용은 데이터에 기반한 변수 탐색과 예측력 개선이 가능하기 때문에 농업 분야 인공지능의 기초 기술로 활용될 수 있을 것이다.

시·군 단위 단수 조사 데이터는 없으나 기상 관측 데이터는 있는 경우에 대한 레이블이 없는 데이터 활용 방법으로써의 준지도 학습 방법은도 단위 단수 데이터만 이용한 예측 모형 학습 상황에서 예측력 개선을 위한 방법으로 활용될 수 있을 뿐만 아니라 쌀 단수 예측에서는 예측 오

차의 크기는 다소 커졌으나 도 단위로 조사된 단수 데이터를 이용하여 준지도 학습 방법을 사용할 경우 시·군 단위의 단수를 보다 정확하게 예측할 수 있는 가능성을 보여주었다. 이는 시·군 단위 단수 데이터가 집계되지 않는 작물에 대해서도 특정 시·군 지역의 과거의 기상 정보를 이용하여 작물 재배가 적합한 지역인지에 대한 판단 근거로 단수 예측값을 제공할 수 있다는 점에서 의의를 가진다.

쌀 단수에 대한 조기 예측 모형의 실증 분석 결과에 따르면 예측 시점 기준에서 측정되지 않는 미래 시점의 기상 변수에 대해서는 추정치 또는 기상에 대한 예측 모형에 의한 예측치를 사용하는 것이 권장된다. 만약 국가 단위의 작물별 작황 조기 예측 시스템이 개발된다면 조기 예측을 위한 기상 예보의 기능을 기상청과는 독립적으로 운영하면서 작물별 특성을 고려한 농업 기상 예보 모형을 발전시키는 방안도 검토가 필요할 것이다. 작황 조기 예측 시스템은 기후변화로 인한 농작물 생산 체계의 불확실성이 커져 가면서 그 필요성 또한 증대될 것으로 예상된다. 따라서 기후변화로 인한 생산 체계의 변동성이 커지기 전에 작물별 품종 및 작부체계를 고려한 기상 요인의 영향에 대한 분석과 데이터 축적이 선행되어야 한다. 또한 다양한 품목과 지역별 분석까지 고려했을 때 현재의 국내 농업 분야의 분석 인력이 부족한 문제를 해결하기 위해 농업 데이터 분석에 특화되어 있는 인공지능 기반 기술도 함께 연구하고 발전시켜야 할 것이다.

제 2 절 연구의 한계 및 향후 연구 방향

본 연구는 쌀 단수 예측 모형에서의 기계 학습 방법 적용이라는 새로운 시도에 중점을 두어 의미있는 연구 결과를 도출하였으나 통계 기반의 쌀 단수 예측 모형에서 고려되어야 할 중요한 영향 요인들이 고르게 다루어지지 못한 한계를 가진다. 국내에서 재배되고 있는 쌀 품종은 2018년 농촌진흥청 자료에 의하면 상위 10개 품종이 전체 재배면적의 78.5%로 21.5%는 수십 종류의 품종으로 구성되어 있을 만큼 다양하다. 품종에 의한 단수의 영향 요인 차이 외에도 지역별 관개시설의 유형과 수준, 병해충에 대한 방제방식, 토양 성분에서도 차이를 보일 수 있다. 이와 같은 차이는 지역별 농협 및 생산 조직의 구성에 따라서 인접한 지역에서는 유사한 특징을 보일 수 있기 때문에 지역에 대한 통계적 처리를 통해서 지역 간의 차이를 어느정도 반영할 수는 있다. 지역별 더미변수를 추가하거나 공간 패널 회귀 모형을 이용하여 지역 간 단수의 차이를 예측 모형의 절편 값으로 추정하여 지역별 품종 및 경작기술의 차이를 반영할 수 있을 것이다. 그러나 지역별로 재배 시기가 다른 경우에 대한 차이는 지역에 의한 차이를 통제하더라도 해결되기가 어려운 문제를 가진다. 예를들어 모내기 시기가 5월 하순인 곳과 6월 상순인 곳은 모내기 직후에 영향을 주는 기상 변수가 5월과 6월로 달라지게 된다. 이와 같이 변수 간 영향 구조가 달라지는 문제는 본 연구에와 같은 예측 모형의 접근 방식에서는 해결하기가 어렵다. 이 경우에는 지역별 이앙 시기 조사 데이터가 필요하며 지역별 기상 요인 변수를 이앙 시기만큼의 편차를 보정할 수 있도록 날짜에 이앙 시기 편차를 추가하여 집계될 수 있도록 재가공이 필요하다. 향후 연구에서는 지역별 주요 품종, 경작 기술에 관한 추가 데이터의 수집이 고려되어야 하며 추가 데이터 수집이 어렵다면 지역별 더미변수 또는 공간 패널 회귀 모형을 이용한 지역별 특성을 통제하는 방법을 고려해야 한다. 또한 지역별 재배 기간을 반영할 수 있는 이앙 시기 데이터 수집과 가공을 통해서 현재의 쌀 단수 예측 모형이 가지는 구조적인 문제도 개선이 필요하다.

본 연구에서는 베이지안 모형 평균화 방법의 가능성과 한계점을 검토하기 위해 벼 생육 기간을 벗어난 기간의 기상 변수도 설명력을 가지는지를 살펴보았고 실제로 몇몇 겨울철 기상 변수들이 유의한 영향을 미치는 것이 확인되었다. 이와 같은 발견은 기존의 이론이 설명하지 못하는 새로운 영역에서의 요인 간 관계를 발견하고 새로운 연구 주제를 제시한다는 점에서는 긍정적으로 평가할 수 있으나 새로운 변수의 영향에 대한 심도 있는 검증이 필요하는 점에서는 해결해야 할 과제를 남긴다. 본 연구에서는 겨울철 기상 변수가 단수에 직접적인 영향을 주기 보다는 병해충 발생의 영향 변수이거나 1, 2월의 최저기온이 9월의 최저기온과 상관관계가 높음으로 인해서 미칠 수 있는 영향만 확인되었다. 따라서 이와 같은 변수들이 중요한 의사결정에 영향을 줄 수 있는 예측 모형에서 사용되기 위해서는 실제 겨울철 기상과 병해충 발생의 관계와 해당 병해충 발생에 의해서 벼의 생육이 영향을 받았는지에 대한 충분한 검증이 선행되어야 한다. 베이지안 모형 평균화와 같은 기계 학습 방법이 인공지능에 적용되어 효율적으로 영향 변수를 탐색하고 그로 인해 예측력을 높인다하더라도 설명 가능한 인공지능이 되지 못한다면 실제 현장에서는 새로운 기술을 채택하는 것을 주저하게 될것이다. 그러므로 새로운 변수 선별 방법이 아무리 예측력을 높일지라도 기존의 선행연구에 기반한 예측 모형의 토대 위에서 충분한 검증이 끝나기 전까지는 참고 자료로 활용하는 것이 바람직할 것이다.

시·군 단위 단수 데이터가 존재하는 쌀의 경우에는 도 단위 단수 데이터만을 이용한 준지도 학습 방법에 의한 예측력 개선과 시·군 단위 단수 예측치의 정확도 수준을 확인할 수 있었으나 다른 작물의 경우 예측력 개선만 확인할 수 있는 한계를 가진다. 향후 연구에서는 통계청에서는 공개되지 않는 시·군 단위 단수 데이터를 지자체나 농협을 통해서 수집하여 준지도 학습에 의한 시·군 단수 예측치가 어느정도 정확한지에 대한 평가가 필요하다. 준지도 학습이 정책적으로 활용될 수 있는 방안으로는 현재는 농업관측본부에서 9월 말에 쌀 단수를 예측을 하고 있는데 8월 이전의 시기에 주요 재배 지역을 표본추출하고 해당 지역의 벼

생육에 대한 정밀 조사를 통해 그 지역의 단수를 예측할 수 있다면 표본 추출된 지역이 레이블데이터가 되는 준지도 학습 모형이 적용될 수 있을 것이다. 과거에는 농업관측본부에서 지역별 작황 상태를 조사하여 전체 단수에 동일한 가중치로 반영하였다면 준지도 학습 방법을 활용하면 지역별로 기상 조건이 다른 영향을 반영할 수 있기 때문에 보다 정확한 예측이 가능 할 것으로 기대해볼 수 있다.

쌀 단수의 조기 예측 모형 연구에서는 예측 시점 이후의 알 수 없는 기상 변수를 잘 예측할 경우 평균값 또는 중위수를 이용하는 경우보다 조기 예측의 정확도를 높일 수 있는 것은 확인되었으나 기상 변수의 예측이라는 더 어려운 과제가 남게 되는 한계점을 가지고 있다. 따라서 향후 연구에서는 미래 시점의 기상을 예측하기보다는 기상 조건별 시나리오를 확률적으로 분류하고 각 시나리오별 예측치가 실제 과거 데이터로 평가하였을 때 어느정도 정확한지를 평가하여 실제 현장에서 활용 가능한 연구 결과를 도출하는 것이 필요하다.

참 고 문 헌

권오상, 김창길. 2008. "기후변화가 쌀 단수변화 미치는 영향: 비모수적 및 준모수적 분석," 농업경제연구 (49:4), pp. 45-64.

남영식, 양승룡, 송용호, 박호정. 2012. "기후변화로 인한 한국 쌀 (정곡) 생산 변화에 관한 연구: Rcp 8.5 시나리오를 중심으로," 농업경제연구 (53:4), pp. 61-88.

농촌경제연구원. 2005a. "사과 생산관측 개선방안 연구".

농촌경제연구원. 2005b. "기상 요인이 청과물 수급에 미치는 영향".

농촌경제연구원. 2011. "기상요인을 고려한 단수예측모형 개발 연구".

농촌경제연구원. 2012. "기후변화가 식량공급에 미치는 영향분석과 대응방안".

농촌경제연구원. 2013. "배추· 무 예측모형 고도화 방안".

이병훈, 문한필. 2012. "공간계량경제 방법론을 이용한 단수반응모형의 예측력에 관한 비교연구," 농업경영· 정책연구 (39:3), pp. 23-43.

조현경, 조은빛, 권오상, 노재선. 2013. "기후변수와 쌀 생산성: 패널 지역자료를 이용한 준모수적 분석," 농업경제연구 (54:3), pp. 71-94.

한석호. 2016. "쌀 단수 예측모형 개발연구," 농업생명과학연구 (50:3), pp. 219-229.

Almaraz, J. J., Mabood, F., Zhou, X., Gregorich, E. G., and Smith, D. L. 2008. "Climate Change, Weather Variability and Corn Yield at a Higher Latitude Locale: Southwestern Quebec," Climatic change (88:2), pp. 187-197.

Alvarez, R. 2009. "Predicting Average Regional Yield and Production of Wheat in the Argentine Pampas by an Artificial Neural Network Approach," European Journal of Agronomy (30:2), pp. 70-77.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P. J., Rötter, R. P., and

Cammarano, D. 2013. "Uncertainty in Simulating Wheat Yields under Climate Change," Nature climate change (3:9), pp. 827-832.

Bagherzadeh, J., and Asil, H. 2019. "A Review of Various Semi-Supervised Learning Models with a Deep Learning and Memory Approach," Iran Journal of Computer Science (2:2), pp. 65-80.

Bakker, M. M., Govers, G., Ewert, F., Rounsevell, M., and Jones, R. 2005. "Variability in Regional Wheat Yields as a Function of Climate, Soil and Economic Variables: Assessing the Risk of Confounding," Agriculture, ecosystems & environment (110:3-4), pp. 195-209.

Barnett, B. J. 2004. "Agricultural Index Insurance Products: Strengths and Limitations."

Barnett, T., Arpe, K., Bengtsson, L., Ji, M., and Kumar, A. 1997. "Potential Predictability and Amip Implications of Midlatitude Climate Variability in Two General Circulation Models," Journal of climate (10:9), pp. 2321-2329.

Basso, B., Ritchie, J., Pierce, F., Braga, R., and Jones, J. 2001. "Spatial Validation of Crop Models for Precision Agriculture," Agricultural Systems (68:2), pp. 97-112.

Blum, A., and Mitchell, T. 1998. "Combining Labeled and Unlabeled Data with Co-Training," Proceedings of the eleventh annual conference on Computational learning theory: ACM, pp. 92-100.

Bornn, L., and Zidek, J. V. 2012. "Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies," Agricultural and forest meteorology (152), pp. 223-232.

Brown, G., Wyatt, J., Harris, R., and Yao, X. 2005. "Diversity Creation Methods: A Survey and Categorisation," Information Fusion (6:1), pp. 5-20.

Cantelaube, P., and Terres, J.-M. 2005. "Seasonal Weather Forecasts for Crop Yield Modelling in Europe," Tellus A: Dynamic Meteorology and Oceanography (57:3), pp. 476-487.

Capa-Morocho, M., Ines, A. V., Baethgen, W. E., Rodríguez-Fonseca, B., Han, E., and Ruiz-Ramos, M. 2016. "Crop Yield Outlooks in the Iberian Peninsula: Connecting Seasonal Climate Forecasts with Crop Simulation Models," Agricultural systems (149), pp. 75-87.

Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., and Dentener, F. 2016. "Impact of Meteorological Drivers on Regional Inter-Annual Crop Yield Variability in France," Agricultural and forest meteorology (216), pp. 58-67.

Ceglar, A., Turco, M., Toreti, A., and Doblas-Reyes, F. J. 2017. "Linking Crop Yield Anomalies to Large-Scale Atmospheric Circulation in Europe," Agricultural and forest meteorology (240), pp. 35-45.

Challinor, A., Slingo, J., Wheeler, T., and Doblas-Reyes, F. 2005. "Probabilistic Simulations of Crop Yield over Western India Using the Demeter Seasonal Hindcast Ensembles," Tellus A: Dynamic Meteorology and Oceanography (57:3), pp. 498-512.

Chen, C., Wang, E., and Yu, Q. 2010. "Modelling the Effects of Climate Variability and Water Management on Crop Water Productivity and Water Balance in the North China Plain," Agricultural Water Management (97:8), pp. 1175-1184.

Chen, K., O'Leary, R. A., and Evans, F. H. 2019. "A Simple and Parsimonious Generalised Additive Model for Predicting Wheat Yield in a Decision Support Tool," Agricultural Systems (173), pp. 140-150.

Conradt, T., Gornott, C., and Wechsung, F. 2016. "Extending and Improving Regionalized Winter Wheat and Silage Maize Yield

Regression Models for Germany: Enhancing the Predictive Skill by Panel Definition through Cluster Analysis," Agricultural and forest meteorology (216), pp. 68-81.

Cortes, C., and Mohri, M. 2007. "On Transductive Regression," Advances in Neural Information Processing Systems, pp. 305-312.

Doblas Reyes, F. J., García Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. 2013. "Seasonal Climate Predictability and Forecasting: Status and Prospects," Wiley Interdisciplinary Reviews: Climate Change (4:4), pp. 245-268.

Draper, N. R., and Smith, H. 1998. Applied Regression Analysis. John Wiley & Sons.

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S. 2007. "Multi-Model Ensemble Hydrologic Prediction Using Bayesian Model Averaging," Advances in Water Resources (30:5), pp. 1371-1386.

Everingham, Y., Sexton, J., Skocaj, D., and Inman-Bamber, G. 2016. "Accurate Prediction of Sugarcane Yield Using a Random Forest Algorithm," Agronomy for sustainable development (36:2), p. 27.

Feng, P., Wang, B., Li Liu, D., Xing, H., Ji, F., Macadam, I., Ruan, H., and Yu, Q. 2018. "Impacts of Rainfall Extremes on Wheat Yield in Semi-Arid Cropping Systems in Eastern Australia," Climatic change (147:3-4), pp. 555-569.

Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., and Obersteiner, M. 2019. "Spatio-Temporal Downscaling of Gridded Crop Model Yield Estimates Based on Machine Learning," Agricultural and forest meteorology (264), pp. 1-15.

Genell, A., Nemes, S., Steineck, G., and Dickman, P. W. 2010. "Model Selection in Medical Research: A Simulation Study Comparing Bayesian Model Averaging and Stepwise Regression," BMC

medical research methodology (10:1), p. 108.
Gitay, H., Brown, S., Easterling, W., Jallow, B., Antle, J., Apps, M., Beamish, R., Chapin, T., Cramer, W., and Frangi, J. 2001. "Ecosystems and Their Goods and Services,".
Gornott, C., and Wechsung, F. 2016. "Statistical Regression Models for Assessing Climate Impacts on Crop Yields: A Validation Study for Winter Wheat and Silage Maize in Germany," Agricultural and Forest Meteorology (217), pp. 89-100.
Gouache, D., Bouchon, A.-S., Jouanneau, E., and Le Bris, X. 2015. "Agrometeorological Analysis and Prediction of Wheat Yield at the Departmental Level in France," Agricultural and Forest Meteorology (209), pp. 1-10.
Hady, M. F. A., Schwenker, F., and Palm, G. 2009. "Semi-Supervised Learning for Regression with Co-Training by Committee," International Conference on Artificial Neural Networks: Springer, pp. 121-130.
Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. 1999. "Bayesian Model Averaging: A Tutorial," Statistical science), pp. 382-401.
Hung, C., Nieto, J., Taylor, Z., Underwood, J., and Sukkarieh, S. 2013. "Orchard Fruit Segmentation Using Multi-Spectral Feature Learning," 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems: IEEE, pp. 5314-5320.
Krogh, A., and Vedelsby, J. 1995. "Neural Network Ensembles, Cross Validation, and Active Learning," Advances in neural information processing systems, pp. 231-238.
Iizumi, T., Sakuma, H., Yokozawa, M., Luo, J.-J., Challinor, A. J., Brown, M. E., Sakurai, G., and Yamagata, T. 2013. "Prediction of Seasonal Climate-Induced Variations in Global Food Production,"

Nature climate change (3:10), p. 904.
Innes, P., Tan, D., Van Ogtrop, F., and Amthor, J. 2015. "Effects of High-Temperature Episodes on Wheat Yields in New South Wales, Australia," Agricultural and Forest Meteorology (208), pp. 95-107.
IPCC. 2007. "Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group Ii to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change," Cambridge Uni-versity Press, Cambridge, UK.
Jacobson, T., and Karlsson, S. 2004. "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," Journal of Forecasting (23:7), pp. 479-496.
Jayakumar, M., Rajavel, M., and Surendran, U. 2016. "Climate-Based Statistical Regression Models for Crop Yield Forecasting of Coffee in Humid Tropical Kerala, India," International journal of biometeorology (60:12), pp. 1943-1952.
Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., and Reddy, V. R. 2016. "Random Forests for Global and Regional Crop Yield Predictions," PLoS One (11:6), p. e0156571.
Kass, R. E., and Raftery, A. E. 1995. "Bayes Factors," Journal of the american statistical association (90:430), pp. 773-795.
Kaul, M., Hill, R. L., and Walthall, C. 2005. "Artificial Neural Networks for Corn and Soybean Yield Prediction," Agricultural Systems (85:1), pp. 1-18.
Leamer, E. E., and Leamer, E. E. 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley New York.
Leng, G., and Huang, M. 2017. "Crop Yield Response to Climate Change Varies with Crop Spatial Distribution Pattern," Scientific reports (7:1), p. 1463.

Li, M., and Zhou, Z.-H. 2007. "Improve Computer-Aided Diagnosis with Machine Learning Techniques Using Undiagnosed Samples," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans (37:6), pp. 1088-1098.

Li, X., Takahashi, T., Suzuki, N., and Kaiser, H. M. 2011. "The Impact of Climate Change on Maize Yields in the United States and China," Agricultural Systems (104:4), pp. 348-353.

Li, Y., Guan, K., Yu, A., Peng, B., Zhao, L., Li, B., and Peng, J. 2019. "Toward Building a Transparent Statistical Model for Improving Crop Yield Prediction: Modeling Rainfed Corn in the Us," Field Crops Research (234), pp. 55-65.

Liu, H., Shi, T., Chen, Y., Wang, J., Fei, T., and Wu, G. 2017. "Improving Spectral Estimation of Soil Organic Carbon Content through Semi-Supervised Regression," Remote Sensing (9:1), p. 29.

Lobell, D. B., and Asseng, S. 2017. "Comparing Estimates of Climate Change Impacts from Process-Based and Statistical Crop Models," Environmental Research Letters (12:1), p. 015001.

Lobell, D. B., and Burke, M. B. 2010. "On the Use of Statistical Models to Predict Crop Yield Responses to Climate Change," Agricultural and forest meteorology (150:11), pp. 1443-1452.

Lottes, P., and Stachniss, C. 2017. "Semi-Supervised Online Visual Crop and Weed Classification in Precision Farming Exploiting Plant Arrangement," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS): IEEE, pp. 5155-5161.

Mathieu, J. A., and Aires, F. 2018. "Assessment of the Agro-Climatic Indices to Improve Crop Yield Forecasting," Agricultural and forest meteorology (253), pp. 15-30.

Matsumura, K., Gaitan, C. F., Sugimoto, K., Cannon, A. J., and Hsieh, W. W. 2015. "Maize Yield Forecasting by Linear Regression and

Artificial Neural Networks in Jilin, China," The Journal of Agricultural Science (153:3), pp. 399-410.

Moriondo, M., Giannakopoulos, C., and Bindi, M. 2011. "Climate Change Impact Assessment: The Role of Climate Extremes in Crop Yield Simulation," Climatic change (104:3-4), pp. 679-701.

Mutshinda, C. M., Finkel, Z. V., and Irwin, A. J. 2013. "Which Environmental Factors Control Phytoplankton Populations? A Bayesian Variable Selection Approach," Ecological Modelling (269), pp. 1-8.

Oettli, P., Behera, S. K., and Yamagata, T. 2018. "Climate Based Predictability of Oil Palm Tree Yield in Malaysia," Scientific reports (8:1), p. 2271.

Pérez-Ortiz, M., Peña, J., Gutiérrez, P. A., Torres-Sánchez, J., Hervás-Martínez, C., and López-Granados, F. 2015. "A Semi-Supervised System for Weed Mapping in Sunflower Crops Using Unmanned Aerial Vehicles and a Crop Row Detection Method," Applied Soft Computing (37), pp. 533-544.

Peng, B., Guan, K., Pan, M., and Li, Y. 2018. "Benefits of Seasonal Climate Prediction and Satellite Data for Forecasting Us Maize Yield," Geophysical Research Letters (45:18), pp. 9662-9671.

Potgieter, A., Hammer, G., Doherty, A., and De Voil, P. 2005. "A Simple Regional-Scale Model for Forecasting Sorghum Yield across North-Eastern Australia," Agricultural and Forest Meteorology (132:1-2), pp. 143-153.

Potgieter, A., Lobell, D. B., Hammer, G. L., Jordan, D. R., Davis, P., and Brider, J. 2016. "Yield Trends under Varying Environmental Conditions for Sorghum and Wheat across Australia," Agricultural and forest meteorology (228), pp. 276-285.

Pozdnoukhov, A., and Bengio, S. 2006. "Semi-Supervised Kernel

Methods for Regression Estimation," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings: IEEE, pp. V-V.

Prost, L., Makowski, D., and Jeuffroy, M.-H. 2008. "Comparison of Stepwise Selection and Bayesian Model Averaging for Yield Gap Analysis," Ecological modelling (219:1-2), pp. 66-76.

Qian, B., De Jong, R., Warren, R., Chipanshi, A., and Hill, H. 2009. "Statistical Spring Wheat Yield Forecasting for the Canadian Prairie Provinces," Agricultural and Forest Meteorology (149:6-7), pp. 1022-1031.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. 2005. "Using Bayesian Model Averaging to Calibrate Forecast Ensembles," Monthly weather review (133:5), pp. 1155-1174.

Raftery, A. E., Madigan, D., and Hoeting, J. A. 1997. "Bayesian Model Averaging for Linear Regression Models," Journal of the American Statistical Association (92:437), pp. 179-191.

Romaszewski, M., Głomb, P., and Cholewa, M. 2016. "Semi-Supervised Hyperspectral Classification from a Small Number of Training Samples Using a Co-Training Approach," ISPRS journal of photogrammetry and remote sensing (121), pp. 60-76.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., and Behringer, D. 2010. "The Ncep Climate Forecast System Reanalysis," Bulletin of the American Meteorological Society (91:8), pp. 1015-1058.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., and Iredell, M. 2014. "The Ncep Climate Forecast System Version 2," Journal of Climate (27:6), pp. 2185-2208.

Schauberger, B., Gornott, C., and Wechsung, F. 2017. "Global Evaluation of a Semiempirical Model for Yield Anomalies and Application to within Season Yield Forecasting," Global change biology (23:11), pp. 4750-4764.

Schwenker, F., Kestler, H. A., and Palm, G. 2001. "Three Learning Phases for Radial-Basis-Function Networks," Neural networks (14:4-5), pp. 439-458.

Schlenker, W., and Lobell, D. B. 2010. "Robust Negative Impacts of Climate Change on African Agriculture," Environmental Research Letters (5:1), p. 014010.

Schlenker, W., and Roberts, M. J. 2009. "Nonlinear Temperature Effects Indicate Severe Damages to Us Crop Yields under Climate Change," Proceedings of the National Academy of sciences (106:37), pp. 15594-15598.

Sharif, B., Makowski, D., Plauborg, F., and Olesen, J. E. 2017. "Comparison of Regression Techniques to Predict Response of Oilseed Rape Yield to Variation in Climatic Conditions in Denmark," European Journal of Agronomy (82), pp. 11-20.

Shi, W., Tao, F., and Zhang, Z. 2013. "A Review on Statistical Models for Identifying Climate Contributions to Crop Yields," Journal of geographical sciences (23:3), pp. 567-576.

Shine, P., Murphy, M. D., Upton, J., and Scully, T. 2018. "Machine-Learning Algorithms for Predicting on-Farm Direct Water and Electricity Consumption on Pasture Based Dairy Farms," Computers and electronics in agriculture (150), pp. 74-87.

Silva, J., Bacao, F., and Caetano, M. 2017. "Specific Land Cover Class Mapping by Semi-Supervised Weighted Support Vector Machines," Remote Sensing (9:2), p. 181.

Sivakumar, M., Gommes, R., and Baier, W. 2000. "Agrometeorology

and Sustainable Agriculture," Agricultural and Forest Meteorology (103:1-2), pp. 11-26.

Sun, B., and Van Kooten, G. 2014. "Weather Effects on Maize Yields in Northern China," The Journal of Agricultural Science (152:4), pp. 523-533.

Van der Velde, M., and Nisini, L. 2019. "Performance of the Mars-Crop Yield Forecasting System for the European Union: Assessing Accuracy, in-Season, and Year-to-Year Improvements from 1993 to 2015," Agricultural systems (168), pp. 203-212.

Villamil, M. B., Davis, V. M., and Nafziger, E. D. 2012. "Estimating Factor Contributions to Soybean Yield from Farm Field Data," Agronomy journal (104:4), pp. 881-887.

Wang, D., Zhang, W., and Bakhai, A. 2004. "Comparison of Bayesian Model Averaging and Stepwise Methods for Model Selection in Logistic Regression," Statistics in medicine (23:22), pp. 3451-3467.

Wang, M., Hua, X.-S., Song, Y., Dai, L.-R., and Zhang, H.-J. 2006. "Semi-Supervised Kernel Regression," Sixth International Conference on Data Mining (ICDM'06): IEEE, pp. 1130-1135.

Wang, F., Fraisse, C., Kitchen, N., and Sudduth, K. 2003. "Site-Specific Evaluation of the Cropgro-Soybean Model on Missouri Claypan Soils," Agricultural Systems (76:3), pp. 985-1005.

Wang, X., Fu, L., and Ma, L. 2011. "Semi-Supervised Support Vector Regression Model for Remote Sensing Water Quality Retrieving," Chinese Geographical Science (21:1), pp. 57-64.

Wang, X., Ma, L., and Wang, X. 2010. "Apply Semi-Supervised Support Vector Regression for Remote Sensing Water Quality Retrieving," 2010 IEEE International Geoscience and Remote Sensing Symposium: IEEE, pp. 2757-2760.

Whisler, F., Acock, B., Baker, D., Fye, R., Hodges, H., Lambert, J.,

Lemmon, H., McKinion, J., and Reddy, V. 1986. "Crop Simulation Models in Agronomic Systems," in Advances in Agronomy. Elsevier, pp. 141-208.

Wright, J. H. 2008. "Bayesian Model Averaging and Exchange Rate Forecasts," Journal of Econometrics (146:2), pp. 329-341.

Yang, Y., Liu, J., Xu, S., and Zhao, Y. 2016. "An Extended Semi-Supervised Regression Approach with Co-Training and Geographical Weighted Regression: A Case Study of Housing Prices in Beijing," ISPRS International Journal of Geo-Information (5:1), p. 4.

Yeung, K. Y., Bumgarner, R. E., and Raftery, A. E. 2005. "Bayesian Model Averaging: Development of an Improved Multi-Class, Gene Selection and Classification Tool for Microarray Data," Bioinformatics (21:10), pp. 2394-2402.

Zellner, A. 1971. An Introduction to Bayesian Inference in Econometrics. Wiley New York.

Zeugner, S., and Feldkircher, M. 2015. "Bayesian Model Averaging Employing Fixed and Flexible Priors: The Bms Package for R," Journal of Statistical Software (68:4), pp. 1-37.

Zhang, Q., Zhang, J., Guo, E., Yan, D., and Sun, Z. 2015a. "The Impacts of Long-Term and Year-to-Year Temperature Change on Corn Yield in China," Theoretical and applied climatology (119:1-2), pp. 77-82.

Zhang, Y., Zhao, Y., Chen, S., Guo, J., and Wang, E. 2015b. "Prediction of Maize Yield Response to Climate Change with Climate and Crop Model Uncertainties," Journal of Applied Meteorology and Climatology (54:4), pp. 785-794.

Zhou, Z.-H., and Li, M. 2005. "Tri-Training: Exploiting Unlabeled Data Using Three Classifiers," IEEE Transactions on Knowledge

& Data Engineering:11), pp. 1529-1541.

Zhou, Z.-H., and Li, M. 2007. "Semisupervised Regression with Cotraining-Style Algorithms," IEEE Transactions on Knowledge and Data Engineering (19:11), pp. 1479-1493.

부 록

<부록 1> 통계 기반 작물 단수 예측 모형 리뷰 논문 목록

Almaraz, J. J., Mabood, F., Zhou, X., Gregorich, E. G., and Smith, D. L. 2008. "Climate Change, Weather Variability and Corn Yield at a Higher Latitude Locale: Southwestern Quebec," Climatic change (88:2), pp. 187-197.

Alvarez, R. 2009. "Predicting Average Regional Yield and Production of Wheat in the Argentine Pampas by an Artificial Neural Network Approach," European Journal of Agronomy (30:2), pp. 70-77.

Bakker, M. M., Govers, G., Ewert, F., Rounsevell, M., and Jones, R. 2005. "Variability in Regional Wheat Yields as a Function of Climate, Soil and Economic Variables: Assessing the Risk of Confounding," Agriculture, ecosystems & environment (110:3-4), pp. 195-209.

Bornn, L., and Zidek, J. V. 2012. "Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies," Agricultural and forest meteorology (152), pp. 223-232.

Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., and Dentener, F. 2016. "Impact of Meteorological Drivers on Regional Inter-Annual Crop Yield Variability in France," Agricultural and forest meteorology (216), pp. 58-67.

Ceglar, A., Turco, M., Toreti, A., and Doblas-Reyes, F. J. 2017. "Linking Crop Yield Anomalies to Large-Scale Atmospheric Circulation in Europe," Agricultural and forest meteorology (240), pp. 35-45.

Chatfield, C. 1995. "Model Uncertainty, Data Mining and Statistical Inference," Journal of the Royal Statistical Society: Series A (Statistics in Society) (158:3), pp. 419-444.

Chen, K., O'Leary, R. A., and Evans, F. H. 2019. "A Simple and Parsimonious Generalised Additive Model for Predicting Wheat Yield in a Decision Support Tool," Agricultural Systems (173), pp. 140-150.

Chipanshi, A. C. 1998. "Monitoring the Effects of Drought on Wheat Yields

in Saskatchewan,".

Conradt, T., Gornott, C., and Wechsung, F. 2016. "Extending and Improving Regionalized Winter Wheat and Silage Maize Yield Regression Models for Germany: Enhancing the Predictive Skill by Panel Definition through Cluster Analysis," Agricultural and forest meteorology (216), pp. 68-81.

De Silva, A. M., and Leong, P. H. 2015. Grammar-Based Feature Generation for Time-Series Prediction. Springer.

Douak, F., Melgani, F., and Benoudjit, N. 2013. "Kernel Ridge Regression with Active Learning for Wind Speed Prediction," Applied energy (103), pp. 328-340.

Everingham, Y., Sexton, J., Skocaj, D., and Inman-Bamber, G. 2016. "Accurate Prediction of Sugarcane Yield Using a Random Forest Algorithm," Agronomy for sustainable development (36:2), p. 27.

Feng, P., Wang, B., Li Liu, D., Xing, H., Ji, F., Macadam, I., Ruan, H., and Yu, Q. 2018. "Impacts of Rainfall Extremes on Wheat Yield in Semi-Arid Cropping Systems in Eastern Australia," Climatic change (147:3-4), pp. 555-569.

Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., and Obersteiner, M. 2019. "Spatio-Temporal Downscaling of Gridded Crop Model Yield Estimates Based on Machine Learning," Agricultural and forest meteorology (264), pp. 1-15.

Gornott, C., and Wechsung, F. 2016. "Statistical Regression Models for Assessing Climate Impacts on Crop Yields: A Validation Study for Winter Wheat and Silage Maize in Germany," Agricultural and Forest Meteorology (217), pp. 89-100.

Gouache, D., Bouchon, A.-S., Jouanneau, E., and Le Bris, X. 2015. "Agrometeorological Analysis and Prediction of Wheat Yield at the Departmental Level in France," Agricultural and Forest Meteorology (209), pp. 1-10.

Guyon, I., and Elisseeff, A. 2003. "An Introduction to Variable and Feature Selection," Journal of machine learning research (3:Mar), pp. 1157-1182.

Hansen, J., and Jones, J. 2000. "Scaling-up Crop Models for Climate

Variability Applications," Agricultural Systems (65:1), pp. 43-72.

Hawkins, E., Osborne, T. M., Ho, C. K., and Challinor, A. J. 2013. "Calibration and Bias Correction of Climate Projections for Crop Modelling: An Idealised Case Study over Europe," Agricultural and forest meteorology (170), pp. 19-31.

Hocking, R. R. 1976. "A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression," Biometrics (32:1), pp. 1-49.

Hoffman, A. L., Kemanian, A. R., and Forest, C. E. 2018. "Analysis of Climate Signals in the Crop Yield Record of Sub Saharan Africa," Global change biology (24:1), pp. 143-157.

Iizumi, T., Sakuma, H., Yokozawa, M., Luo, J.-J., Challinor, A. J., Brown, M. E., Sakurai, G., and Yamagata, T. 2013. "Prediction of Seasonal Climate-Induced Variations in Global Food Production," Nature climate change (3:10), p. 904.

Innes, P., Tan, D., Van Ogtrop, F., and Amthor, J. 2015. "Effects of High-Temperature Episodes on Wheat Yields in New South Wales, Australia," Agricultural and Forest Meteorology (208), pp. 95-107.

Jayakumar, M., Rajavel, M., and Surendran, U. 2016. "Climate-Based Statistical Regression Models for Crop Yield Forecasting of Coffee in Humid Tropical Kerala, India," International journal of biometeorology (60:12), pp. 1943-1952.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., and Reddy, V. R. 2016. "Random Forests for Global and Regional Crop Yield Predictions," PLoS One (11:6), p. e0156571.

Kaul, M., Hill, R. L., and Walthall, C. 2005. "Artificial Neural Networks for Corn and Soybean Yield Prediction," Agricultural Systems (85:1), pp. 1-18.

Leng, G., and Huang, M. 2017. "Crop Yield Response to Climate Change Varies with Crop Spatial Distribution Pattern," Scientific reports (7:1), p. 1463.

Li, X., Takahashi, T., Suzuki, N., and Kaiser, H. M. 2011. "The Impact of

Climate Change on Maize Yields in the United States and China," Agricultural Systems (104:4), pp. 348-353.

Li, Y., Guan, K., Yu, A., Peng, B., Zhao, L., Li, B., and Peng, J. 2019. "Toward Building a Transparent Statistical Model for Improving Crop Yield Prediction: Modeling Rainfed Corn in the Us," Field Crops Research (234), pp. 55-65.

Lobell, D. B. 2013. "Errors in Climate Datasets and Their Effects on Statistical Crop Models," Agricultural and Forest meteorology (170), pp. 58-66.

Lobell, D. B., and Asseng, S. 2017. "Comparing Estimates of Climate Change Impacts from Process-Based and Statistical Crop Models," Environmental Research Letters (12:1), p. 015001.

Lobell, D. B., and Burke, M. B. 2010. "On the Use of Statistical Models to Predict Crop Yield Responses to Climate Change," Agricultural and forest meteorology (150:11), pp. 1443-1452.

Mathieu, J. A., and Aires, F. 2018. "Assessment of the Agro-Climatic Indices to Improve Crop Yield Forecasting," Agricultural and forest meteorology (253), pp. 15-30.

Matsumura, K., Gaitan, C. F., Sugimoto, K., Cannon, A. J., and Hsieh, W. W. 2015. "Maize Yield Forecasting by Linear Regression and Artificial Neural Networks in Jilin, China," The Journal of Agricultural Science (153:3), pp. 399-410.

Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., and Foley, J. A. 2012. "Closing Yield Gaps through Nutrient and Water Management," Nature (490:7419), p. 254

Nicholls, N. 1997. "Increased Australian Wheat Yield Due to Recent Climate Trends," Nature (387:6632), pp. 484-485.

Oettli, P., Behera, S. K., and Yamagata, T. 2018. "Climate Based Predictability of Oil Palm Tree Yield in Malaysia," Scientific reports (8:1), p. 2271.

Peng, B., Guan, K., Pan, M., and Li, Y. 2018. "Benefits of Seasonal Climate Prediction and Satellite Data for Forecasting Us Maize Yield,"

Geophysical Research Letters (45:18), pp. 9662-9671.

Potgieter, A., Hammer, G., Doherty, A., and De Voil, P. 2005. "A Simple Regional-Scale Model for Forecasting Sorghum Yield across North-Eastern Australia," Agricultural and Forest Meteorology (132:1-2), pp. 143-153.

Potgieter, A. B., Lobell, D. B., Hammer, G. L., Jordan, D. R., Davis, P., and Brider, J. 2016. "Yield Trends under Varying Environmental Conditions for Sorghum and Wheat across Australia," Agricultural and forest meteorology (228), pp. 276-285.

Qian, B., De Jong, R., Warren, R., Chipanshi, A., and Hill, H. 2009. "Statistical Spring Wheat Yield Forecasting for the Canadian Prairie Provinces," Agricultural and Forest Meteorology (149:6-7), pp. 1022-1031.

Schauberger, B., Gornott, C., and Wechsung, F. 2017. "Global Evaluation of a Semiempirical Model for Yield Anomalies and Application to within Season Yield Forecasting," Global change biology (23:11), pp. 4750-4764.

Schlenker, W., and Roberts, M. J. 2009. "Nonlinear Temperature Effects Indicate Severe Damages to Us Crop Yields under Climate Change," Proceedings of the National Academy of sciences (106:37), pp. 15594-15598.

Semenov, M. A., and Barrow, E. M. 1997. "Use of a Stochastic Weather Generator in the Development of Climate Change Scenarios," Climatic change (35:4), pp. 397-414.

Shalev-Shwartz, S., and Ben-David, S. 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge university press.

Sharif, B., Makowski, D., Plauborg, F., and Olesen, J. E. 2017. "Comparison of Regression Techniques to Predict Response of Oilseed Rape Yield to Variation in Climatic Conditions in Denmark," European Journal of Agronomy (82), pp. 11-20.

Shi, W., Tao, F., and Zhang, Z. 2013. "A Review on Statistical Models for Identifying Climate Contributions to Crop Yields," Journal of geographical sciences (23:3), pp. 567-576.

Shine, P., Murphy, M. D., Upton, J., and Scully, T. 2018. "Machine-Learning

Algorithms for Predicting on-Farm Direct Water and Electricity Consumption on Pasture Based Dairy Farms," Computers and electronics in agriculture (150), pp. 74-87.

Sun, B., and Van Kooten, G. 2014. "Weather Effects on Maize Yields in Northern China," The Journal of Agricultural Science (152:4), pp. 523-533.

Van der Velde, M., and Nisini, L. 2019. "Performance of the Mars-Crop Yield Forecasting System for the European Union: Assessing Accuracy, in-Season, and Year-to-Year Improvements from 1993 to 2015," Agricultural systems (168), pp. 203-212.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., and Freckleton, R. P. 2006. "Why Do We Still Use Stepwise Modelling in Ecology and Behaviour?," Journal of animal ecology (75:5), pp. 1182-1189.

Zhang, Q., Zhang, J., Guo, E., Yan, D., and Sun, Z. 2015a. "The Impacts of Long-Term and Year-to-Year Temperature Change on Corn Yield in China," Theoretical and applied climatology (119:1-2), pp. 77-82.

Zhang, Y., Zhao, Y., Chen, S., Guo, J., and Wang, E. 2015b. "Prediction of Maize Yield Response to Climate Change with Climate and Crop Model Uncertainties," Journal of Applied Meteorology and Climatology (54:4), pp. 785-794.

<부록 2> 베이지안 모형 평균화 분석 결과

<부표 1> 월 단위 변수의 베이지안 모형 평균화 분석 결과

Variable	PIP	Post Mean	Post SD	Cond.Pos.Sign
tm_avg4	1.000	-31.925	3.780	0.000
tm_avg8	1.000	21.658	3.936	1.000
tm_min4	1.000	9.954	1.798	1.000
tm_min8	1.000	-32.446	2.606	0.000
tm_max4	1.000	15.775	2.100	1.000
tm_max5	1.000	-9.982	1.768	0.000
pr_sum1	1.000	-0.299	0.033	0.000
pr_sum6	1.000	0.048	0.008	1.000
pr_sum8	1.000	-0.031	0.004	0.000
pr_sum9	1.000	-0.029	0.006	0.000
hm_avg10	1.000	0.974	0.181	1.000
ss_dura5	1.000	4.820	0.947	1.000
yield_l1	1.000	0.523	0.013	1.000
hm_avg6	0.997	-0.853	0.252	0.000
pr_sum7	0.995	-0.019	0.005	0.000
tm_avg10	0.961	5.749	3.601	1.000
tm_min5	0.945	7.023	2.107	1.000
ss_dura10	0.941	4.408	1.671	1.000
tm_max7	0.903	6.668	2.306	1.000
tm_min9	0.805	2.605	1.430	1.000
tm_avg1	0.691	2.923	4.155	0.997
hm_avg3	0.592	-0.268	0.247	0.000
ss_dura3	0.458	1.291	1.541	1.000
tm_min10	0.429	-2.080	2.801	0.014
tm_min1	0.372	-0.085	2.037	0.764
pr_sum3	0.364	-0.021	0.031	0.000
ss_dura2	0.312	0.563	0.921	1.000
tm_max1	0.308	-1.217	2.239	0.016
ss_dura6	0.225	-0.378	0.777	0.000

tm_avg9	0.216	0.711	1.536	0.953
tm_avg7	0.123	1.583	4.725	0.987
tm_max8	0.116	-0.523	1.661	0.000
tm_min7	0.109	-0.742	2.373	0.061
tm_max10	0.088	-0.049	1.031	0.441
tm_avg5	0.080	0.854	3.244	0.991
hm_avg8	0.079	-0.031	0.129	0.001
ss_dura1	0.072	0.124	0.577	0.916
tm_max3	0.066	0.067	0.307	0.993
tm_min2	0.064	-0.194	0.968	0.097
tm_avg2	0.062	0.192	1.027	0.823
ss_dura8	0.057	0.075	0.378	0.994
tm_avg6	0.057	-0.097	0.489	0.010
tm_min3	0.052	0.061	0.322	0.985
tm_avg3	0.052	0.057	0.314	0.966
tm_max6	0.049	-0.057	0.323	0.026
hm_avg2	0.046	-0.006	0.042	0.131
tm_min6	0.043	-0.050	0.323	0.025
ss_dura4	0.035	-0.046	0.311	0.000
hm_avg1	0.035	0.004	0.036	0.839
pr_sum4	0.031	0.000	0.003	1.000
pr_sum2	0.030	-0.001	0.005	0.042
hm_avg7	0.030	0.004	0.048	0.766
pr_sum10	0.029	0.000	0.003	0.904
hm_avg9	0.028	0.003	0.044	0.680
tm_max9	0.025	-0.010	0.243	0.403
ss_dura7	0.023	-0.001	0.125	0.495
hm_avg5	0.021	0.003	0.046	0.811
hm_avg4	0.020	0.001	0.029	0.609
ss_dura9	0.019	0.010	0.141	0.884
tm_max2	0.019	0.004	0.177	0.528
pr_sum5	0.018	0.000	0.002	0.344

<부표 2> 15일 단위 변수의 베이지안 모형 평균화 분석 결과

* 변수명 끝의 숫자는 월의 전반기(1), 후반기(2)를 의미함

Variable	PIP	Post Mean	Post SD	Cond.Pos.Sign
tm_avg11	1.000	29.120	4.362	1.000
tm_avg101	1.000	8.914	1.443	1.000
tm_min11	1.000	-16.830	2.417	0.000
tm_min31	1.000	12.049	2.662	1.000
tm_min51	1.000	11.656	1.975	1.000
tm_max11	1.000	-13.075	2.362	0.000
tm_max32	1.000	6.583	1.460	1.000
pr_sum31	1.000	-0.247	0.043	0.000
pr_sum72	1.000	-0.053	0.009	0.000
pr_sum101	1.000	-0.173	0.029	0.000
hm_avg61	1.000	-1.314	0.260	0.000
hm_avg101	1.000	1.844	0.267	1.000
ss_dura82	1.000	5.764	0.844	1.000
ss_dura92	1.000	-4.202	1.034	0.000
yield_l1	1.000	0.542	0.013	1.000
tm_avg31	0.998	-16.156	6.676	0.000
tm_min32	0.997	-7.383	1.403	0.000
pr_sum81	0.996	-0.032	0.008	0.000
tm_min81	0.991	-4.551	1.271	0.000
tm_avg71	0.985	8.269	4.653	1.000
tm_avg51	0.981	-14.338	2.962	0.000
tm_max92	0.971	7.528	2.192	1.000
ss_dura62	0.947	-2.835	1.011	0.000
ss_dura51	0.946	4.042	1.407	1.000
hm_avg21	0.927	-0.756	0.315	0.000
ss_dura11	0.922	4.260	1.730	1.000
tm_min91	0.892	3.442	1.616	1.000
hm_avg102	0.882	-0.791	0.368	0.000
tm_min102	0.879	-4.895	2.042	0.000
tm_min92	0.874	-4.838	2.395	0.000

tm_max22	0.855	-5.568	3.327	0.000
pr_sum21	0.834	0.163	0.088	1.000
pr_sum12	0.824	-0.222	0.121	0.000
pr_sum22	0.781	0.150	0.095	1.000
tm_avg42	0.775	-3.373	2.047	0.000
hm_avg11	0.755	0.528	0.346	1.000
tm_max72	0.711	5.623	4.110	1.000
tm_avg22	0.704	5.915	4.580	0.986
pr_sum61	0.687	0.029	0.022	1.000
tm_avg72	0.681	-8.147	6.100	0.005
pr_sum41	0.580	0.044	0.041	1.000
tm_max52	0.570	1.522	1.478	1.000
tm_max31	0.553	3.732	3.770	0.997
pr_sum32	0.539	-0.089	0.093	0.000
tm_min82	0.500	-2.183	2.400	0.000
pr_sum11	0.356	-0.064	0.093	0.000
ss_dura32	0.320	-1.278	2.008	0.000
tm_min72	0.313	-1.295	2.073	0.008
ss_dura102	0.273	0.867	1.548	1.000
tm_min42	0.252	-0.815	1.487	0.001
pr_sum62	0.243	-0.017	0.033	0.002
tm_avg82	0.241	-0.931	1.967	0.016
tm_max41	0.233	0.507	0.997	0.998
tm_max82	0.205	-0.578	1.265	0.010
tm_avg61	0.203	-0.716	1.527	0.000
tm_min71	0.203	-1.504	3.447	0.000
hm_avg41	0.191	0.120	0.270	0.995
pr_sum71	0.191	-0.003	0.007	0.000
tm_min22	0.188	0.526	1.236	0.987
tm_max61	0.175	-0.463	1.103	0.001
tm_avg92	0.168	-0.609	3.791	0.247
tm_avg41	0.152	0.452	1.164	0.993
ss_dura41	0.144	0.359	0.978	0.992

hm_avg32	0.141	-0.068	0.187	0.000
tm_avg102	0.129	-0.880	2.452	0.018
tm_min21	0.123	0.198	0.590	0.999
tm_min41	0.119	0.306	0.924	0.996
ss_dura21	0.099	-0.219	0.743	0.000
tm_min62	0.090	0.257	1.039	1.000
ss_dura22	0.077	0.188	0.727	0.991
tm_avg52	0.077	0.227	0.849	0.995
tm_max101	0.071	-0.249	1.025	0.000
ss_dura72	0.068	0.109	0.462	0.997
hm_avg71	0.068	0.041	0.171	0.998
tm_max71	0.063	-0.254	1.820	0.316
hm_avg62	0.055	-0.025	0.123	0.029
tm_min61	0.054	-0.121	0.587	0.043
tm_max102	0.052	0.151	0.822	0.902
tm_avg62	0.046	-0.095	0.894	0.502
tm_max42	0.044	0.017	0.461	0.542
pr_sum92	0.042	-0.001	0.005	0.002
pr_sum82	0.037	0.000	0.003	0.981
tm_max81	0.035	0.071	0.460	0.979
tm_avg21	0.030	0.035	0.261	0.960
hm_avg91	0.029	0.013	0.084	1.000
tm_avg91	0.028	-0.024	0.570	0.323
tm_max62	0.028	-0.002	0.295	0.698
ss_dura52	0.028	0.036	0.267	0.917
ss_dura42	0.025	-0.024	0.228	0.199
ss_dura71	0.024	0.022	0.173	0.973
tm_avg81	0.023	0.033	0.620	0.760
hm_avg72	0.023	0.000	0.048	0.464
tm_min101	0.022	0.008	0.382	0.703
ss_dura61	0.022	0.022	0.199	0.949
tm_max51	0.022	-0.138	1.035	0.050
ss_dura91	0.021	0.011	0.179	0.741

pr_sum42	0.021	0.001	0.008	0.864
pr_sum51	0.020	0.001	0.006	0.957
hm_avg92	0.020	0.003	0.041	0.878
hm_avg42	0.019	-0.001	0.029	0.280
hm_avg52	0.018	0.001	0.032	0.541
hm_avg12	0.017	-0.003	0.028	0.144
hm_avg31	0.016	0.002	0.026	0.842
hm_avg81	0.016	0.007	0.062	0.999
tm_min52	0.016	0.017	0.184	0.910
hm_avg22	0.015	0.002	0.032	0.728
hm_avg82	0.014	0.005	0.056	0.890
ss_dura81	0.014	-0.009	0.154	0.270
tm_avg12	0.014	0.007	0.095	0.790
ss_dura12	0.013	0.011	0.137	0.927
tm_max21	0.012	0.004	0.113	0.774
tm_max12	0.012	-0.006	0.079	0.005
tm_min12	0.012	0.004	0.072	0.645
tm_avg32	0.011	-0.057	0.801	0.133
pr_sum52	0.010	0.000	0.003	0.020
tm_max91	0.010	-0.003	0.118	0.271
ss_dura101	0.009	0.002	0.158	0.434
hm_avg51	0.008	0.003	0.039	0.883
pr_sum91	0.007	0.000	0.001	0.052
pr_sum102	0.005	0.000	0.002	0.392
ss_dura31	0.003	0.000	0.054	0.651

<부표 3> 10일 단위 변수의 베이지안 모형 평균화 분석 결과

* 변수명 끝의 숫자는 월의 상순(1), 중순(2), 하순(3)을 의미함

	PIP	Post Mean	Post SD	Cond.Pos.Sign
tm_avg52	1.000	-10.620	2.068	0.000
tm_avg102	1.000	13.502	2.729	1.000
tm_min52	1.000	10.157	1.658	1.000
tm_min102	1.000	-7.133	1.799	0.000
tm_max21	1.000	-4.293	1.662	0.000
tm_max42	1.000	3.975	1.074	1.000
tm_max82	1.000	8.335	1.376	1.000
pr_sum33	1.000	-0.359	0.076	0.000
pr_sum72	1.000	-0.053	0.012	0.000
pr_sum83	1.000	-0.052	0.009	0.000
hm_avg32	1.000	1.083	0.273	1.000
hm_avg62	1.000	-2.207	0.461	0.000
hm_avg102	1.000	1.040	0.209	1.000
ss_dura42	1.000	5.347	1.317	1.000
ss_dura62	1.000	-5.750	0.900	0.000
ss_dura63	1.000	3.652	0.890	1.000
yield_l1	1.000	0.570	0.015	1.000
ss_dura33	0.997	-5.306	1.152	0.000
pr_sum11	0.991	-0.407	0.108	0.000
ss_dura52	0.974	4.335	1.313	1.000
hm_avg42	0.939	0.856	0.337	1.000
tm_avg13	0.898	5.532	5.584	1.000
pr_sum63	0.895	0.051	0.023	1.000
hm_avg31	0.883	-0.645	0.286	0.000
pr_sum71	0.783	-0.029	0.018	0.000
tm_avg101	0.759	-4.167	2.662	0.000
tm_avg92	0.747	-3.636	2.402	0.002
hm_avg41	0.693	-0.480	0.374	0.000
hm_avg101	0.685	-0.660	0.510	0.000
ss_dura81	0.677	1.445	1.146	1.000

tm_min31	0.655	-2.400	1.919	0.000
ss_dura12	0.652	1.876	1.510	1.000
hm_avg82	0.611	0.535	0.468	1.000
tm_avg62	0.579	-2.714	2.601	0.001
tm_max32	0.577	1.551	1.490	1.000
tm_min83	0.522	-4.139	4.429	0.001
tm_avg103	0.446	1.496	1.856	1.000
pr_sum13	0.442	-0.147	0.181	0.000
tm_avg83	0.389	2.777	3.712	0.995
tm_min91	0.382	-1.224	1.659	0.002
tm_avg31	0.366	-1.669	2.402	0.019
tm_max41	0.365	-1.166	1.610	0.002
tm_min43	0.359	-1.186	1.756	0.000
hm_avg92	0.358	0.303	0.447	0.999
tm_max13	0.355	-2.188	3.289	0.005
pr_sum42	0.348	0.038	0.057	1.000
tm_min22	0.342	0.873	1.326	1.000
tm_avg43	0.331	-1.301	2.089	0.003
tm_avg33	0.321	-2.024	3.444	0.007
tm_min33	0.290	1.664	2.764	0.996
tm_avg21	0.284	1.311	2.234	1.000
hm_avg61	0.267	0.154	0.272	1.000
tm_min93	0.258	0.648	1.194	1.000
tm_min92	0.252	-1.221	2.275	0.010
ss_dura31	0.247	0.793	1.498	1.000
tm_min101	0.232	-0.589	1.432	0.091
tm_avg61	0.217	-0.707	1.420	0.005
tm_min13	0.212	-1.081	2.349	0.061
ss_dura61	0.203	-0.429	0.928	0.003
hm_avg12	0.192	-0.085	0.186	0.000
hm_avg71	0.184	0.114	0.266	1.000
tm_avg91	0.178	-0.568	1.320	0.012
ss_dura22	0.176	0.422	1.005	1.000

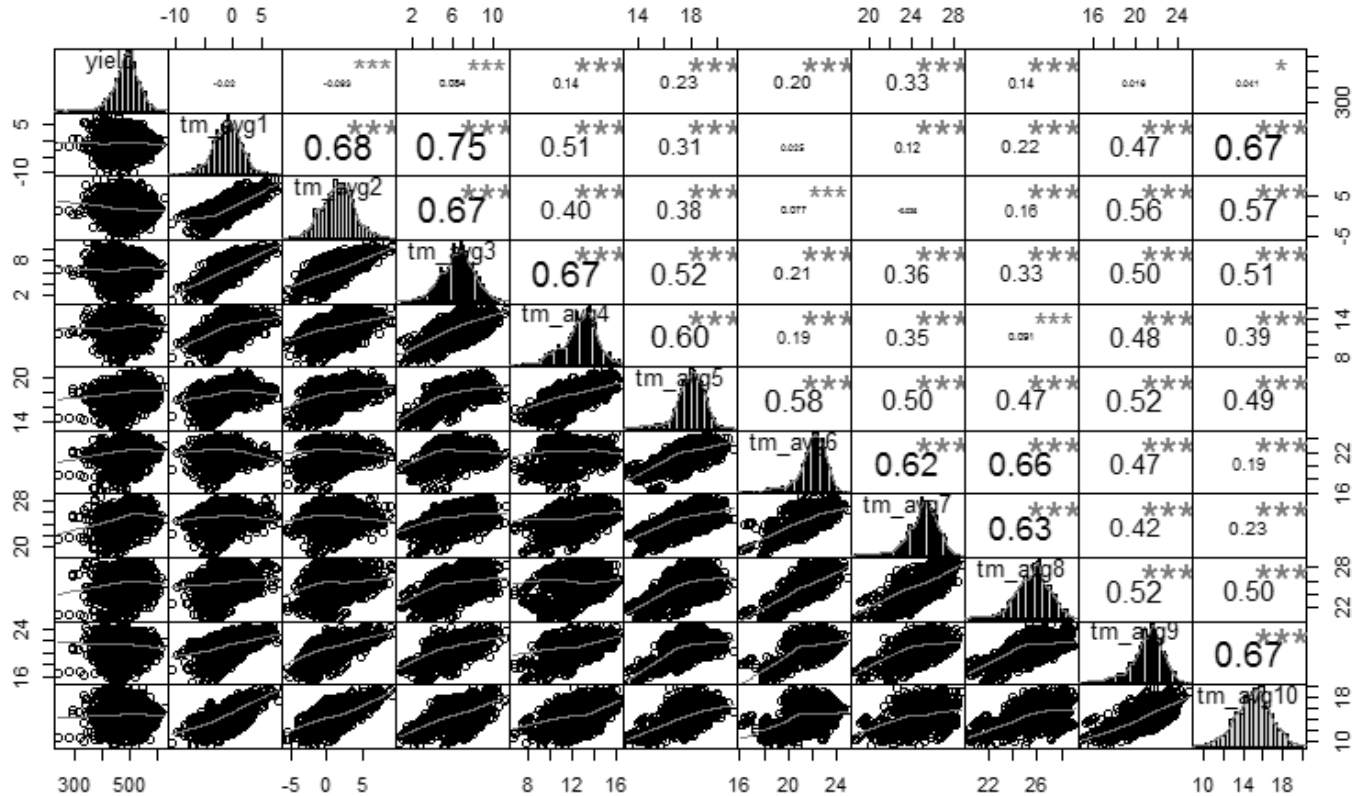
hm_avg63	0.173	0.115	0.271	0.992
tm_max33	0.169	-0.098	0.969	0.179
tm_max83	0.168	0.505	1.239	1.000
hm_avg43	0.168	-0.084	0.207	0.036
tm_max103	0.157	0.422	1.057	1.000
hm_avg11	0.156	-0.068	0.171	0.023
pr_sum103	0.145	-0.014	0.037	0.003
pr_sum43	0.143	0.012	0.032	1.000
tm_max23	0.138	0.288	0.831	0.997
ss_dura23	0.136	-0.351	0.993	0.001
ss_dura82	0.122	-0.294	0.848	0.000
tm_avg11	0.119	0.164	0.496	0.980
tm_max73	0.119	0.240	0.721	0.997
tm_avg41	0.098	-0.234	0.769	0.009
tm_min23	0.095	-0.194	0.729	0.133
ss_dura91	0.086	-0.142	0.515	0.000
tm_max62	0.084	-0.180	0.671	0.034
pr_sum12	0.083	-0.014	0.052	0.000
tm_avg32	0.082	0.120	0.759	0.869
hm_avg93	0.080	0.050	0.185	0.975
hm_avg13	0.072	0.023	0.095	0.992
tm_min81	0.071	-0.167	0.672	0.000
tm_max81	0.061	-0.156	0.655	0.004
tm_min11	0.061	0.054	0.297	0.897
tm_max12	0.056	-0.033	0.364	0.375
tm_min12	0.053	-0.062	0.298	0.000
tm_max61	0.051	-0.063	0.368	0.106
pr_sum73	0.051	-0.001	0.005	0.000
tm_avg22	0.049	0.046	0.522	0.853
tm_avg73	0.041	0.167	0.982	1.000
pr_sum62	0.040	-0.001	0.008	0.000
ss_dura92	0.039	-0.042	0.261	0.111
ss_dura41	0.038	-0.054	0.345	0.092

tm_max101	0.037	0.026	0.538	0.677
tm_avg81	0.036	-0.087	0.502	0.007
tm_max31	0.035	0.092	0.576	0.959
tm_max91	0.035	-0.050	0.298	0.000
tm_max51	0.034	0.038	0.241	1.000
tm_max71	0.034	0.082	0.544	0.937
tm_max11	0.032	0.001	0.193	0.538
tm_min61	0.032	-0.060	0.381	0.012
tm_min62	0.032	0.086	0.587	0.919
ss_dura11	0.032	-0.031	0.313	0.230
tm_max72	0.030	0.018	0.183	0.868
hm_avg51	0.027	-0.009	0.061	0.032
tm_min41	0.026	-0.024	0.201	0.057
tm_max22	0.025	0.005	0.226	0.759
tm_avg12	0.024	-0.024	0.215	0.127
ss_dura73	0.024	0.031	0.229	1.000
tm_min103	0.021	0.000	0.290	0.537
tm_min71	0.021	-0.001	0.281	0.414
tm_min73	0.021	-0.067	0.621	0.236
pr_sum91	0.021	0.001	0.004	0.927
tm_avg51	0.021	0.006	0.209	0.662
ss_dura71	0.020	0.031	0.265	0.966
ss_dura101	0.019	-0.035	0.281	0.000
tm_max43	0.019	0.011	0.267	0.555
tm_min53	0.019	0.025	0.216	0.960
tm_avg93	0.018	0.023	0.245	0.803
ss_dura21	0.018	0.012	0.124	0.928
ss_dura13	0.017	0.024	0.220	1.000
pr_sum51	0.016	0.001	0.006	0.871
tm_min21	0.016	0.021	0.187	1.000
tm_avg71	0.016	-0.020	0.343	0.320
hm_avg22	0.016	-0.003	0.030	0.058
tm_max92	0.015	0.024	0.257	0.965

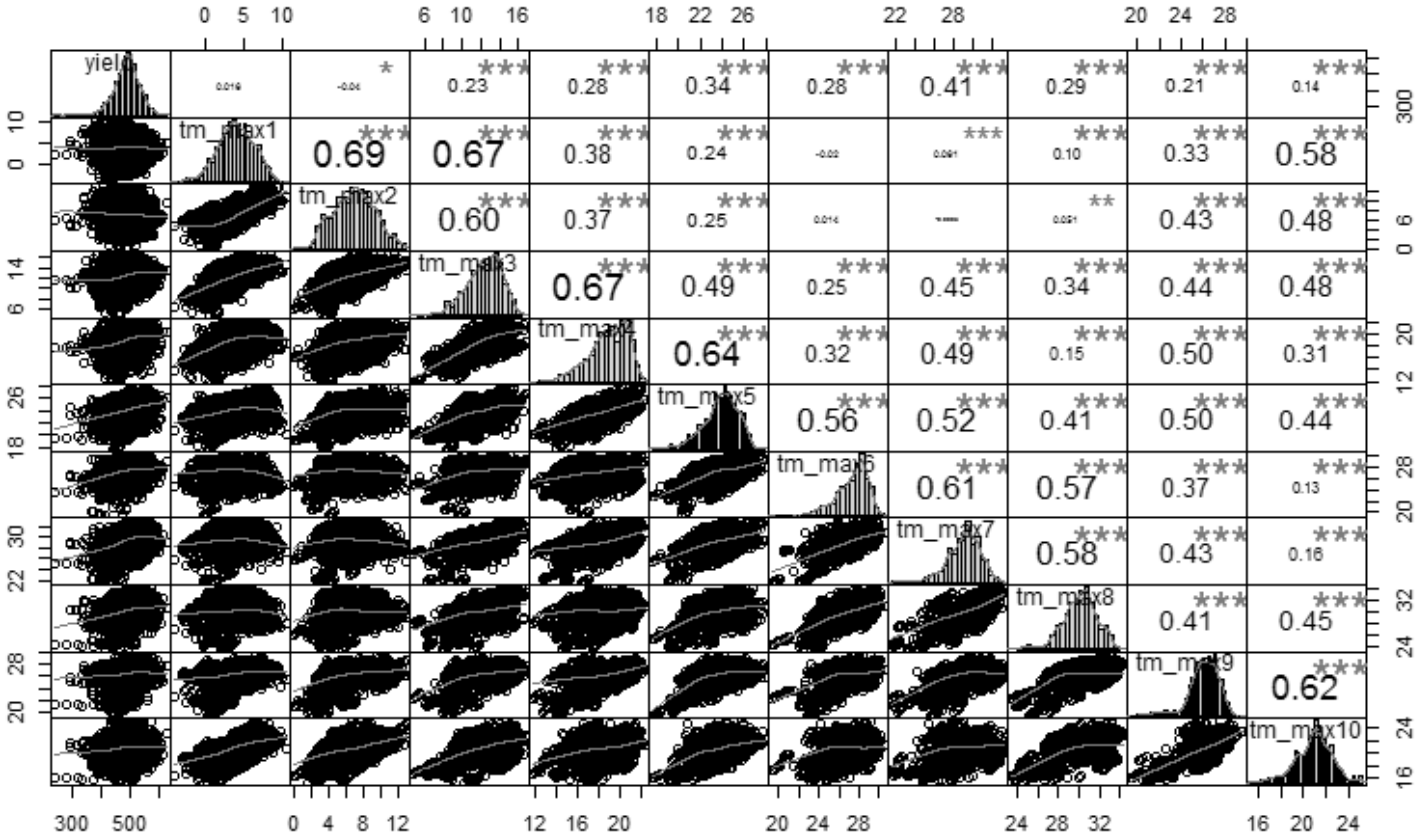
pr_sum41	0.015	-0.001	0.008	0.021
hm_avg52	0.015	-0.004	0.041	0.005
hm_avg33	0.014	0.000	0.021	0.376
pr_sum92	0.014	0.000	0.002	0.907
hm_avg91	0.014	0.004	0.041	1.000
ss_dura83	0.014	0.014	0.169	0.778
tm_min51	0.014	-0.004	0.127	0.339
pr_sum61	0.013	0.000	0.004	0.282
pr_sum32	0.013	0.001	0.008	0.944
pr_sum53	0.013	0.000	0.005	0.681
pr_sum81	0.013	0.000	0.001	0.752
pr_sum82	0.012	0.000	0.002	1.000
tm_min63	0.011	0.005	0.126	0.559
ss_dura53	0.010	-0.006	0.116	0.273
ss_dura102	0.010	0.006	0.116	0.790
tm_max102	0.010	-0.017	0.282	0.142
pr_sum101	0.010	0.000	0.004	0.000
hm_avg83	0.010	-0.002	0.029	0.201
ss_dura43	0.010	0.004	0.079	0.787
tm_min72	0.010	-0.002	0.081	0.383
tm_max53	0.010	-0.004	0.073	0.143
pr_sum31	0.009	0.000	0.006	0.968
hm_avg103	0.008	-0.002	0.026	0.000
tm_max63	0.008	0.006	0.096	1.000
tm_avg63	0.008	-0.002	0.101	0.320
tm_avg23	0.008	-0.006	0.270	0.625
tm_max93	0.008	-0.003	0.090	0.171
pr_sum21	0.008	0.000	0.007	0.175
tm_avg53	0.007	0.004	0.084	0.911
ss_dura103	0.007	0.009	0.112	1.000
pr_sum52	0.007	0.000	0.003	0.280
pr_sum22	0.007	0.000	0.006	0.639
tm_max52	0.006	0.005	0.158	0.836

tm_min32	0.006	0.003	0.146	0.425
pr_sum23	0.006	0.000	0.003	0.265
hm_avg23	0.006	0.000	0.015	0.491
tm_avg82	0.006	0.006	0.154	1.000
ss_dura93	0.006	0.000	0.062	0.451
ss_dura51	0.006	-0.001	0.053	0.492
pr_sum93	0.005	0.000	0.002	0.668
hm_avg81	0.005	0.001	0.030	0.686
ss_dura32	0.005	0.000	0.064	0.440
tm_avg72	0.005	0.001	0.145	0.796
hm_avg53	0.005	0.001	0.012	0.981
ss_dura72	0.005	-0.004	0.082	0.159
hm_avg73	0.004	0.000	0.020	0.248
tm_min82	0.004	0.001	0.076	0.706
tm_min42	0.004	-0.001	0.071	0.352
hm_avg72	0.003	0.000	0.009	0.738
pr_sum102	0.003	0.000	0.003	0.676
tm_avg42	0.000	0.000	0.033	0.000
hm_avg21	0.000	0.000	0.001	1.000

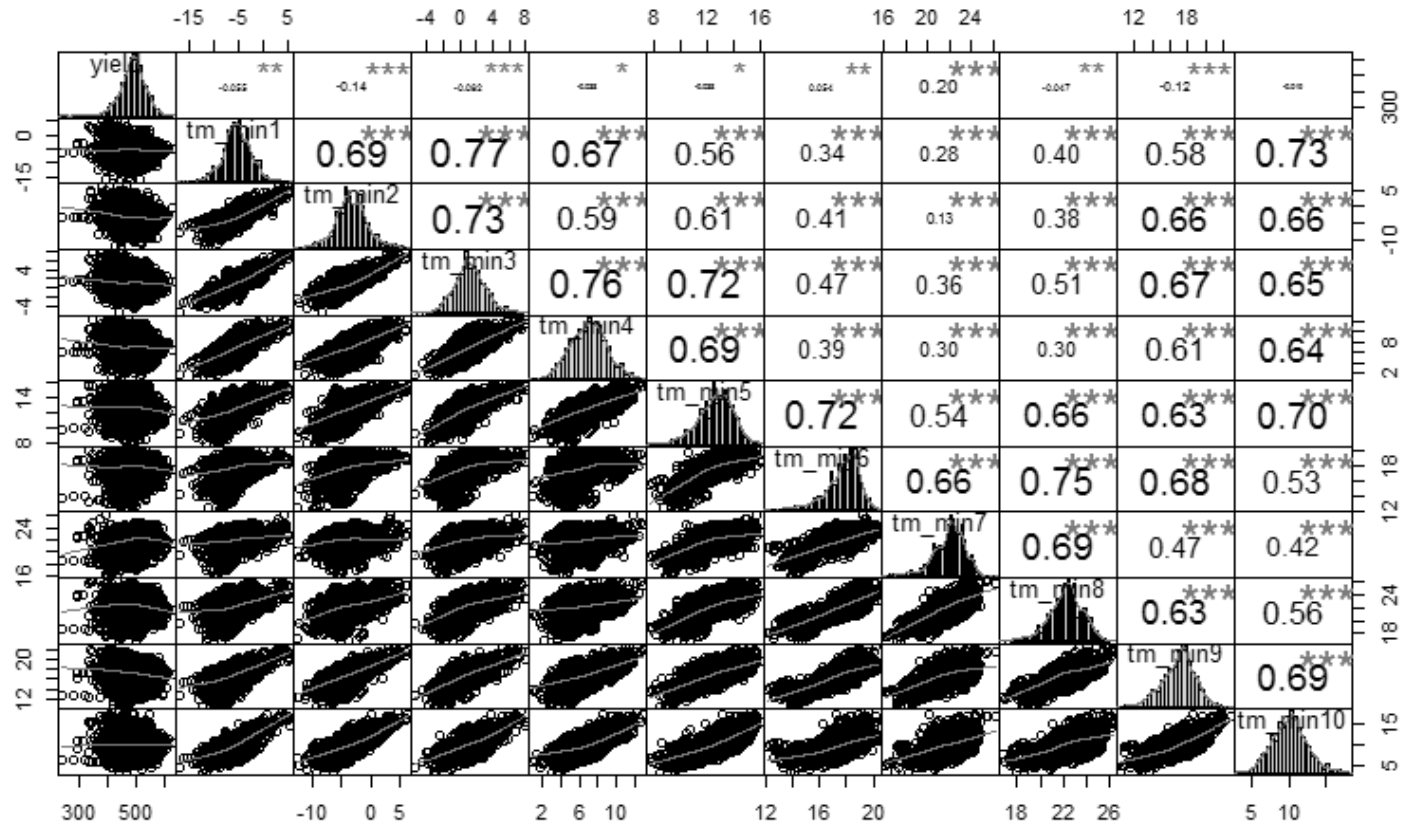
<부록 4> 쌀 단수와 월 평균 기온과의 상관계수 및 분포 특성



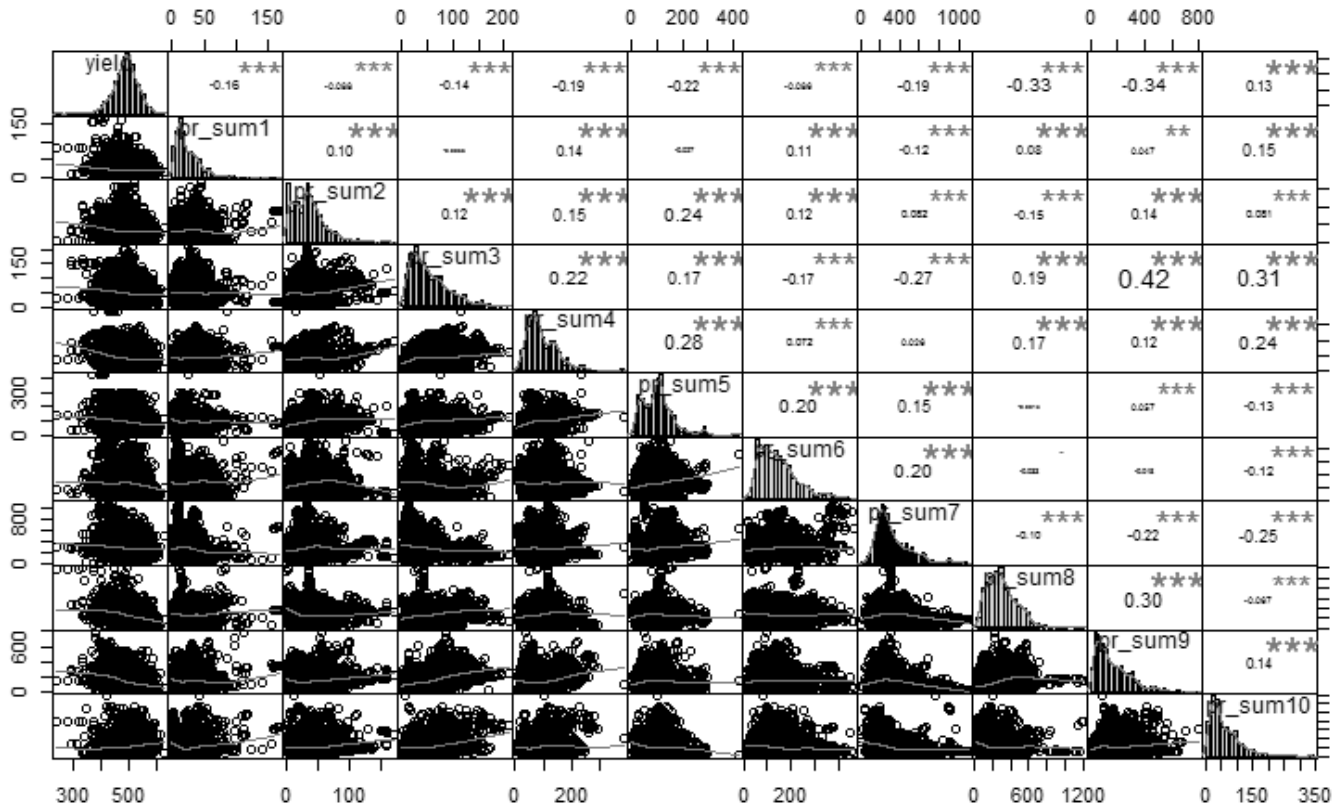
<그림 부-1> 쌀 단수와 월 평균 기온과의 상관계수 및 분포 특성



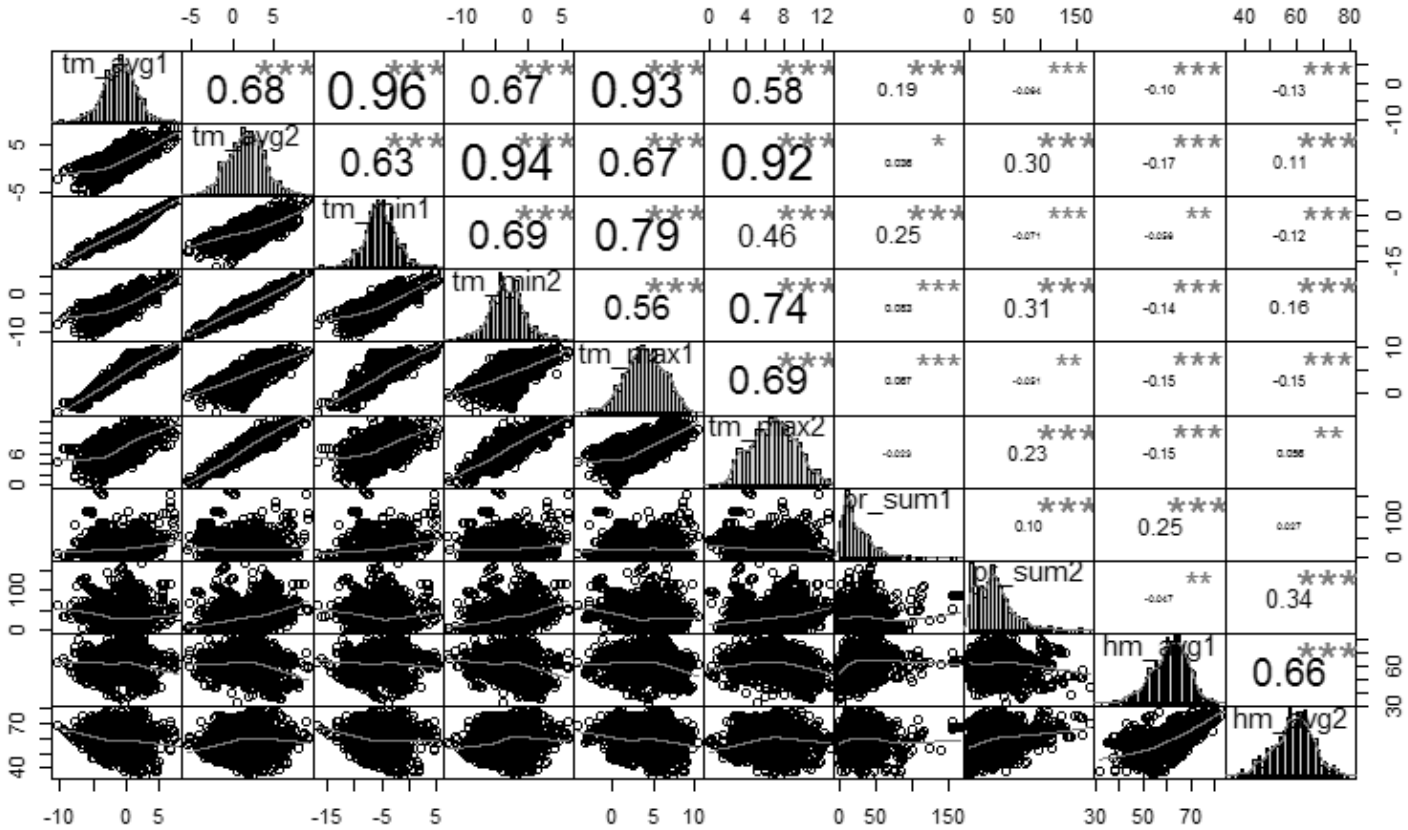
<그림 부-2> 쌀 단수와 월 최고 기온과의 상관계수 및 분포 특성



<그림 부-3> 쌀 단수와 월 최저 기온과의 상관계수 및 분포 특성



<그림 부-4> 쌀 단수와 월 누적 강수량의 상관계수 및 분포 특성



<그림 부-5> 1, 2월의 누적 강수량의 상관계수 및 분포 특성

<부록 5> 준지도 회귀 방법의 타 작물 적용을 위한 예측 모형과 예측력

1. 작물 선택 및 모형 도출 방법

- 작물 선정 기준 : 통계청에서 단수 집계 데이터를 제공하는 품목 중 계절별로 품목이 분리되지 않으면서(예: 봄감자, 가을감자) 전국적으로 노지 재배 비중이 높은 작물인 콩, 고구마, 양파, 마늘을 선정하였음
- 단수 모형 도출 방법 : 각 작물별 파종(정식)에서 수확까지의 기간에 대한 모든 기상요인 월 집계 변수를 후보변수로 두고 베이지안 모형 평균화 기법을 이용하여 설명력이 높은 변수를 중심으로 예측 모형을 설정하였음

2. 콩 단수 모형

yield = yield_l1 + tm_avg7 + tm_avg8 + tm_avg9 + hm_avg7 + hm_avg9 + pr_sum8+ pr_sum9

3. 고구마 단수 모형

yield = yield_l1 + tm_avg7 + tm_avg8 + tm_avg5 + tm_avg6 + tm_max5 + tm_max6 + tm_max7 + tm_min5 + tm_min6 + pr_sum5 + pr_sum6 + pr_sum7 + pr_sum8 + hm_avg5 + ss_dura5

4. 양파 단수 모형

yield = yield_l1 + tm_avg1 + tm_avg2 + tm_avg3 + tm_avg4 + tm_avg5 + tm_max1 + tm_max2 + tm_max3 + tm_max5 + tm_max9l + tm_min1 + tm_min2 + tm_min3 + tm_min4 + tm_min5 + r_sum2 + pr_sum10l + ss_dura3 + ss_dura4 + ss_dura11l + hm_avg1

5. 마늘 단수 모형

yield = yield_l1 + tm_avg1 + tm_avg3 + tm_avg4 + tm_avg5 + tm_avg9l + tm_avg10l + tm_avg11l + tm_max1 + tm_max3 + tm_max4 + tm_max9l+ tm_max10l + tm_min1 + tm_min3 + tm_min4 + tm_min9l + tm_min10l + tm_min11l + pr_sum9l + pr_sum4 + ss_dura10l + ss_dura3 + ss_dura5 + hm_avg1

<부표 4> 작물별 준지도 회귀(SSR) 방법과 OLS, SVR의 예측 성능 비교

지표	기법	콩		고구마		양파		마늘	
		100%	50%	100%	50%	100%	50%	100%	50%
RMSE	OLS	22.42	23.14	200.2	219.8	904.4	990	126	133.4
	SVR	19.1	19.15	184.5	200.9	855.6	779.9	123	136.5
	SSR	18.24	16.78	183.2	193.8	812.6	762.2	118.1	131
MAPE	OLS	10.9%	11.6%	9.2%	7.8%	14.3%	14.4%	9.9%	10.2%
	SVR	9.2%	10%	8.7%	8.5%	14.9%	12.2%	9.6%	9.2%
	SSR	8.3%	8.8%	9.0%	8.4%	13.9%	12.2%	8.4%	9.3%
R^2	OLS	31.2%	27.5%	51%	45.7%	61.9%	65.4%	83.6%	79.5%
	SVR	46.9%	54.9%	57.8%	58.4%	67.7%	75.9%	83.9%	79.6%
	SSR	50.6%	67.1%	62.4%	57.5%	71.0%	77.4%	84.7%	80.2%

<부록 6> 베이지안 모형 평균화 모형의 월별 예측 모형

10월

Yt = Yt-1 + tm_avg4t + tm_avg8t + tm_avg9t + tm_avg10t + tm_max4 + tm_max5 + tm_max7 + tm_min4t + tm_min5t + tm_min8t + hm_avg3t + hm_avg6t + hm_avg10 + ss_dura5t + ss_dura10t + pr_sum1t + pr_sum6t + pr_sum7t + pr_sum8t + pr_sum9t

9월

Yt = Yt-1 + hm_avg5 + hm_avg6 + pr_sum1 + pr_sum6 + pr_sum8 + ss_dura3 + ss_dura5 + tm_avg1 + tm_avg4 + tm_avg7 + tm_avg8 + tm_avg9 + tm_max4 + tm_max5 + tm_min4 + tm_min5 + tm_min6 + tm_min8

8월

Yt = Yt-1 + hm_avg2 + hm_avg3 + hm_avg5 + hm_avg6 + hm_avg7 + pr_sum1 + pr_sum6 + pr_sum8 + ss_dura3 + ss_dura5 + tm_avg1 + tm_avg4 + tm_avg6 + tm_avg7 + tm_avg8 + tm_max4 + tm_max5 + tm_min4 + tm_min5 + tm_min8

7월

Yt = Yt-1 + hm_avg2 + hm_avg3 + hm_avg4 + hm_avg6 + hm_avg7 + pr_sum1 + pr_sum4 + pr_sum5 + pr_sum6 + ss_dura4 + tm_avg2 + tm_avg4 + tm_avg6 + tm_avg7 + tm_max1 + tm_max2 + tm_max4 + tm_max7 + tm_min2 + tm_min5

6월

Yt = Yt-1 + hm_avg4 + hm_avg6 + pr_sum1 + pr_sum2 + pr_sum4 + pr_sum6 + ss_dura3 + ss_dura4 + ss_dura6 + tm_avg2 + tm_avg4 + tm_max1 + tm_max2 + tm_max4 + tm_max5 + tm_min2 + tm_min3 + tm_min5 + tm_min6

<부록 7> BMA를 이용한 기상 변수 예측 모형과 모형 예측 성능 비교

tm_avg10 = hm_avg2 + hm_avg3 + hm_avg4 + hm_avg8 + pr_sum2 +
pr_sum3 + pr_sum4 + ss_dura2 + ss_dura3 + ss_dura4 + ss_dura5 +
ss_dura7 + ss_dura8 + ss_dura9 + tm_avg2 + tm_avg4 + tm_avg5 +
tm_avg9 + tm_max1 + tm_max2 + tm_max3 + tm_max5 + tm_max6 +
tm_max7 + tm_min2 + tm_min3 + tm_min5 + tm_min8 + tm_min9

tm_avg9 = hm_avg1_l1 + hm_avg10_l1 + hm_avg2 + hm_avg2_l1 + hm_avg4 +
hm_avg6 + hm_avg7_l1 + hm_avg8_l1 + hm_avg9_l1 + pr_sum1_l1 +
pr_sum10_l1 + pr_sum5 + pr_sum6 + pr_sum7 + pr_sum8 + pr_sum8_l1 +
ss_dura1 + ss_dura1_l1 + ss_dura3_l1 + ss_dura4_l1 + ss_dura5_l1 +
ss_dura7 + ss_dura7_l1 + ss_dura8 + ss_dura8_l1 + ss_dura9_l1 +
tm_avg10.avg + tm_avg10_l1 + tm_avg12_l1 + tm_avg2 + tm_avg2_l1 +
tm_avg4_l1 + tm_avg5 + tm_avg5_l1 + tm_avg6 + tm_avg6_l1 +
tm_avg8 + tm_avg8_l1 + tm_avg9_l1 + tm_max1 + tm_max1_l1 +
tm_max12_l1 + tm_max2 + tm_max2_l1 + tm_max3 + tm_max3_l1 +
tm_max4 + tm_max5_l1 + tm_max7 + tm_max7_l1 + tm_max8 + tm_max8.avg +
tm_max8_l1 + tm_max9.avg + tm_min10_l1 + tm_min11_l1 + tm_min12_l1 +
tm_min2 + tm_min2_l1 + tm_min4_l1 + tm_min5 + tm_min5_l1 +
tm_min7 + tm_min8.avg + tm_min9.avg

tm_min9 = hm_avg1_l1 + hm_avg10_l1 + hm_avg12_l1 + hm_avg2 +
hm_avg2_l1 + hm_avg3 + hm_avg3_l1 + hm_avg4_l1 + hm_avg6 +
hm_avg7_l1 + hm_avg8 + hm_avg8_l1 + hm_avg9_l1 + pr_sum10_l1 +
pr_sum5 + pr_sum6 + pr_sum7 + pr_sum7_l1 + pr_sum8_l1 + ss_dura1 +
ss_dura11_l1 + ss_dura2 + ss_dura3_l1 + ss_dura4 + ss_dura4_l1 +
ss_dura5_l1 + ss_dura8 + tm_avg10.avg + tm_avg10_l1 + tm_avg11_l1 +
tm_avg2 + tm_avg2_l1 + tm_avg3_l1 + tm_avg5_l1 + tm_avg6_l1 +
tm_avg7_l1 + tm_avg8.avg + tm_avg8_l1 + tm_avg9_l1 + tm_max1 +
tm_max1_l1 + tm_max11_l1 + tm_max2 + tm_max2_l1 + tm_max3 +
tm_max4 + tm_max5 + tm_max5_l1 + tm_max6 + tm_max6_l1 + tm_max7 +
tm_max8 + tm_min11_l1 + tm_min12_l1 + tm_min2_l1 + tm_min3_l1 +
tm_min4_l1 + tm_min7_l1 + tm_min8 + tm_min8.avg + tm_min9.avg

tm_max9 = hm_avg1_l1 + hm_avg2 + hm_avg2_l1 + hm_avg3_l1 + hm_avg4_l1 +
hm_avg5 + hm_avg6 + hm_avg7_l1 + hm_avg9_l1 + pr_sum1_l1 +
pr_sum10_l1 + pr_sum2 + pr_sum3_l1 + pr_sum5 + pr_sum6 +
pr_sum7 + pr_sum8 + pr_sum9_l1 + ss_dura1 + ss_dura1_l1 +
ss_dura10_l1 + ss_dura2 + ss_dura3_l1 + ss_dura4_l1 + ss_dura5_l1 +
ss_dura7 + ss_dura8 + ss_dura9_l1 + tm_avg1 + tm_avg11_l1 +
tm_avg12_l1 + tm_avg2 + tm_avg3 + tm_avg3_l1 + tm_avg4 +
tm_avg4_l1 + tm_avg5 + tm_avg6 + tm_avg7 + tm_avg8 + tm_avg8_l1 +
tm_avg9.avg + tm_avg9_l1 + tm_max1 + tm_max1_l1 + tm_max11_l1 +
tm_max12_l1 + tm_max2 + tm_max2_l1 + tm_max3 + tm_max4 +
tm_max5 + tm_max6_l1 + tm_max7 + tm_max7_l1 + tm_max8.avg +
tm_max8_l1 + tm_max9.avg + tm_max9_l1 + tm_min1_l1 + tm_min12_l1 +
tm_min2 + tm_min3 + tm_min4 + tm_min4_l1 + tm_min5_l1 + tm_min7 +
tm_min7_l1 + tm_min8 + tm_min8_l1 + tm_min9.avg + tm_min9_l1

tm_avg8 = hm_avg1 + hm_avg12_l1 + hm_avg2 + hm_avg3 + hm_avg3_l1 +
hm_avg4 + hm_avg4_l1 + hm_avg5_l1 + hm_avg7 + hm_avg7_l1 +
hm_avg8_l1 + pr_sum1 + pr_sum1_l1 + pr_sum12_l1 + pr_sum2_l1 +
pr_sum3_l1 + pr_sum4 + pr_sum4_l1 + pr_sum5 + pr_sum5_l1 +
pr_sum7 + pr_sum8_l1 + pr_sum9_l1 + ss_dura1 + ss_dura10_l1 +
ss_dura11_l1 + ss_dura12_l1 + ss_dura2 + ss_dura3 + ss_dura4 +
ss_dura4_l1 + ss_dura5 + ss_dura5_l1 + ss_dura6_l1 + ss_dura7 +
ss_dura7_l1 + ss_dura8_l1 + tm_avg1 + tm_avg1_l1 + tm_avg10_l1 +
tm_avg2_l1 + tm_avg3_l1 + tm_avg4 + tm_avg6 + tm_avg6_l1 +
tm_avg7 + tm_avg7_l1 + tm_avg9.avg + tm_avg9_l1 + tm_max10_l1 +
tm_max3_l1 + tm_max4 + tm_max6 + tm_max9.avg + tm_max9_l1 +
tm_min1 + tm_min10_l1 + tm_min11_l1 + tm_min12_l1 + tm_min2_l1 +
tm_min3_l1 + tm_min5_l1 + tm_min6 + tm_min7 + tm_min7.avg +
tm_min8.avg + tm_min8_l1 + tm_min9.avg

tm_min8 = hm_avg10_l1 + hm_avg11_l1 + hm_avg2 + hm_avg3_l1 +
hm_avg4 + hm_avg4_l1 + hm_avg5_l1 + hm_avg6 + hm_avg8_l1 +
hm_avg9_l1 + pr_sum1 + pr_sum1_l1 + pr_sum10_l1 + pr_sum12_l1 +

pr_sum2 + pr_sum2_l1 + pr_sum3 + pr_sum3_l1 + pr_sum4 + pr_sum4_l1 + pr_sum6_l1 + pr_sum7 + ss_dura1 + ss_dura11_l1 + ss_dura3 + ss_dura4 + ss_dura4_l1 + ss_dura5 + ss_dura5_l1 + ss_dura6_l1 + ss_dura7_l1 + tm_avg1 + tm_avg10.avg + tm_avg10_l1 + tm_avg2_l1 + tm_avg3_l1 + tm_avg4 + tm_avg6 + tm_avg6_l1 + tm_avg7 + tm_avg8.avg + tm_avg9_l1 + tm_max10_l1 + tm_max2 + tm_max2_l1 + tm_max3_l1 + tm_max4 + tm_max7 + tm_max7.avg + tm_max7_l1 + tm_max9_l1 + tm_min1 + tm_min1_l1 + tm_min11_l1 + tm_min2_l1 + tm_min3 + tm_min4 + tm_min5_l1 + tm_min7.avg + tm_min8.avg + tm_min8_l1

tm_max8 ~ hm_avg1 + hm_avg12_l1 + hm_avg2 + hm_avg3 + hm_avg4_l1 + hm_avg5_l1 + hm_avg6_l1 + hm_avg7_l1 + hm_avg9_l1 + pr_sum1_l1 + pr_sum12_l1 + pr_sum2 + pr_sum2_l1 + pr_sum3_l1 + pr_sum4 + pr_sum5 + pr_sum5_l1 + pr_sum7 + pr_sum8_l1 + pr_sum9_l1 + ss_dura1 + ss_dura1_l1 + ss_dura11_l1 + ss_dura12_l1 + ss_dura2 + ss_dura3 + ss_dura3_l1 + ss_dura4 + ss_dura5 + ss_dura6 + ss_dura7 + ss_dura8_l1 + tm_avg10_l1 + tm_avg11_l1 + tm_avg2 + tm_avg2_l1 + tm_avg3 + tm_avg3_l1 + tm_avg4_l1 + tm_avg5_l1 + tm_avg6 + tm_avg6_l1 + tm_avg7 + tm_avg7_l1 + tm_avg8.avg + tm_avg8_l1 + tm_avg9.avg + tm_avg9_l1 + tm_max1 + tm_max1_l1 + tm_max10_l1 + tm_max12_l1 + tm_max2 + tm_max3 + tm_max3_l1 + tm_max4 + tm_max4_l1 + tm_max5_l1 + tm_max7 + tm_max7.avg + tm_max7_l1 + tm_max9.avg + tm_min1 + tm_min10_l1 + tm_min2_l1 + tm_min3 + tm_min3_l1 + tm_min4 + tm_min6 + tm_min6_l1 + tm_min7 + tm_min8_l1 + tm_min9.avg

tm_avg7 ~ hm_avg1 + hm_avg2 + hm_avg4 + hm_avg5 + pr_sum2 + pr_sum4 + pr_sum5 + pr_sum6 + ss_dura2 + ss_dura3 + ss_dura4 + ss_dura5 + tm_avg1 + tm_avg10.avg + tm_avg2 + tm_avg3 + tm_avg4 + tm_avg5 + tm_avg6 + tm_max1 + tm_max2 + tm_max5 + tm_max8.avg + tm_min1 + tm_min5

tm_min7 ~ hm_avg1 + hm_avg2 + hm_avg4 + hm_avg5 + hm_avg6 + pr_sum2 + pr_sum4 + pr_sum5 + ss_dura3 + ss_dura4 + ss_dura5 + tm_avg1 +

tm_avg10.avg + tm_avg4 + tm_avg6 + tm_max1 + tm_max3 + tm_min1 + tm_min2 + tm_min4 + tm_min5 + tm_min6

tm_max7 ~ hm_avg1 + hm_avg2 + hm_avg4 + hm_avg5 + hm_avg6 + pr_sum2 + pr_sum4 + pr_sum5 + ss_dura3 + ss_dura4 + ss_dura5 + tm_avg1 + tm_avg10.avg + tm_avg4 + tm_avg6 + tm_max1 + tm_max3 + tm_min1 + tm_min2 + tm_min4 + tm_min5 + tm_min6

<부표 5> BMA를 이용한 기상 변수 예측 모형의 변수 유형별 R^2

	당해연도 기상 변수	지역 평균변수 포함	지역평균 1년전 변수 포함
tm_avg10	**0.727**	0.578	0.652
tm_avg9	0.522	0.659	**0.677**
tm_min9	0.568	0.586	**0.629**
tm_max9	0.447	0.587	**0.68**
tm_avg8	0.408	0.303	**0.522**
tm_min8	0.506	0.417	**0.642**
tm_max8	0.282	0.238	**0.423**
tm_avg7	0.42	**0.633**	0.416
tm_min7	0.397	**0.504**	0.467
tm_max7	0.481	**0.529**	0.408

Abstract

Application of Machine Learning Methods for improving Rice Yield Prediction Models

Jang, Ikhoon

Major in Regional Information

Department of Agricultural Economics and Rural Development

The Graduate School

Seoul National University

The government has maintained a rice production system through various policies in order to stabilize the supply and demand of rice as main grain. However, the recent abandonment of the status of WTO developing countries has increased anxiety for rice producers. The volatility of weather conditions due to global climate change is increasing, and the uncertainty of rice production is gradually increasing. In addition, Korea is facing an important turning point in changing the rice planting schedule to adapt the agricultural sector to climate change.

In order to ensure stable rice supply and demand, future production observations must be preceded, and this requires the development and operation of reliable rice yield forecasting models. In Korea, the Agricultural Observation Center, which is run by the Korea Rural Economic Institute, develops a model predicting rice yields at the end of September and publishes the forecasts at the end of September.

However, the machine learning method that is widely used in the prediction field has not been actively studied in the field of rice yield prediction in Korea. Therefore, this study explores the research topics that can contribute to the improvement of rice yield prediction model using machine learning method and suggests new research method through the following empirical research to solve the research question.

First, we examined whether the predictive performance improved when the rice yield predictor was selected using Bayesian model averaging which is a machine learning method that has not been used in previous studies. Most of the meteorological variables selected by Bayesian model averaging were found to be significant by regression analysis, and the performance of the prediction model by Bayesian model averaging was superior to previous model based on cross-validation. The performance of the prediction model trained by machine learning methods such as support vector regression (SVR) is better than the prediction by the linear regression model (OLS) used in the conventional statistical model. However, when evaluating the predictive performance over the past 7 years with the prediction model learned from past data before 2012, the performance of the prediction model by Bayesian model averaging did not show any significant difference in the error-based indicators compared with the previous studies. In terms of explanatory power-based indicators, the performance was lower than that of the relatively simple preceding model. In the case of predicting the future with high uncertainty, we found that complex prediction models with many explanatory variables can lead poor predictive performance due to overfitting problems.

Second, in the situation where there is a lack of label data with observations of dependent variables, and there is abundance of unlabeled observation data, a semi-supervised regression method that

can improve the prediction performance by using unlabeled data for prediction model training was applied to the rice yield prediction model. The results showed that the predictive performance of the rice yield prediction model was improved by 4.6% on the error-based indicator and 5.8% on the explanatory power-based index, compared with the case without using the semi-superviesed regression method. The model that trained 200 labed province data using a semi-supervised regression method performed 12% lower on the error-based index, 8.9% lower on the explanatory power-based indicator than the reference model that uses more than 2,000 labeled city data for training. This is a meaningful result considering the number of training data. Thus, the semi-supervised regression method can be a good alternative to improve the performance of prediction models by using unlabeled data.

Third, when comparing existing methods for early prediction of crop yield with methods using machine learning techniques that were not used in previous studies, the prediction performance was found to be improved when using machine learning. In the comparison of the prediction performance by lead time of the early prediction, the prediction performance of the model using future weather variables is superior to the models without future weather variables in all four lead time models. Future weather variables were found to outperform predictive models that do not use future weather variables even if they use past mean or median value of weather variables. In particular, the longer the lead time of early prediction, the more pronounced this effect is. Also, in the model using future weather variables, if the forecasting point has a certain level of lead time (2-4 months before harvesting in this study), applying the predicted value by the prediction model to future weather variables will improve the predictive performance. However, the use of forecasts by weather predcition model does not show significant benefits in yield prediction

models with longer lead times (five months before harvest).

The results of the empirical research suggest that the application of Bayesian model averaging to the crop yield prediction model enables data-based variable search and predictive performance improvement. Therefore, it is expected to be developed as the basic technology of artificial intelligence in agriculture, and it is necessary to continuously verify through subsequent research. The semi-supervised regression method can provide yield estimates using historical weather information for a specific city even when yield data of the city are not observed. This makes it possible to determine if the crop is suitable for the region. The empirical analysis of the early prediction model suggests that prediction model using future weather variables that are not measured at the baseline are recommended for improving predictive performance. In this context, it is necessary to expand research focused on agricultural weather forecasts for early crop yield prediction. As the uncertainty of the crop production system increased due to the recent climate change, the importance of early prediction of crops is expected to increase gradually. Therefore, before the volatility of the production system due to climate change increases, analysis and data accumulation on the effects of climate factors should be preceded. In addition, considering the analysis of various crops and regions, there is a lack of analysis experts in the domestic agricultural sector. To solve this problem, Research on artificial intelligence-based technologies that are specialized for agricultural data analysis should continue to be developed.

Keywords : Rice Yield, Prediction Model, Machine Learning, Bayesian Model Averaging, Semi-supervised Regression
Student Number : 2012-30990