**위성 사진 분석을 통한 옥수수 산출량 추정**

|  |  |  |  |
| --- | --- | --- | --- |
| 제출일 | 2023.06.21 | 전공 | 경제학과 |
| 과목 | AI금융 | 학번 | 20170529 | 20180555 | 20180492 |
| 담당교수 | 정재식 | 이름 | 김유성 | 손재범 | 김범준 |

**목 차**

**Ⅰ. 서론 -------------------------------------------------------------------------------------- 1**

1. 연구 목적 및 배경 ------------------------------------------------------------------------- 1

2. 선행 연구 검토 ---------------------------------------------------------------------------- 2

3. 연구 과정 요약 -----------------------------------------------------------------------------3

**Ⅱ. 연구 과정 및 방법 ----------------------------------------------------------------------- 3**

1. Google earth engine 및 State 단위 위성사진 -------------------------------------------- 4

2. County 경계 데이터 ----------------------------------------------------------------------- 6

3.County 단위 NDVI 및 yield 데이터 -------------------------------------------------------7

4. 시도해봤으나 수집하지 못한 데이터들과 그 이유 ---------------------------------------8

5. 데이터 가공 -------------------------------------------------------------------------------10

6. 고려했으나 적용하지 않은 데이터 처리방식과 그 이유---------------------------------11

7. ConvLSTM --------------------------------------------------------------------------------- 12

8. 고려해보았으나 적용하지 못한 모델과 그 이유----------------------------------------- 13

**Ⅱ. 결론 및 개선점 ------------------------------------------------------------------------- 14**

1. 분석결과 및 해석 ------------------------------------------------------------------------ 14

2. 개선가능한 부분--------------------------------------------------------------------------- 15

3. 결론-----------------------------------------------------------------------------------------16

**Ⅰ. 서론**

**1. 연구 목적 및 배경**

본 연구는 세계적으로 옥수수 생산량이 가장 많은 콘벨트 지역의 옥수수 단수를 NDVI와 태평양 해수면 온도 지수를 분석해 추정하는 것을 목표로 한다.

옥수수의 [[1]](#footnote-1)공급요인은 생산량과 재고량으로 나누어져 있다. 그 중 생산량을 단순화시키면 생산면적 \* 단위 수확량 으로 단순화 시킬 수 있다. 생산면적과 재고량은 미국 농무부의 재배의향면적 조사, 재고량 조사 등으로 매년 체계적으로 파악된다. 하지만 단위 수확량은 작황에 대한 농민들의 주관적인 설문과, 최종적으로 수확된 양에 따라 추정된다. 우리는 식물의 생장에 대한 지표인 NDVI를 위성 사진을 통해 분석해서 단수를 조기에 추정하고자 시도했다.

일반적으로 라니냐(La Nina) 발생 시기에는 수확량이 저조하며 엘니뇨(El Nino) 발생 시기에는 수확량이 높은 경향성을 보인다. [[2]](#footnote-2)옥수수의 경우 1970~2015를 대상으로 분석한 결과 엘니뇨 발생 시기에 단수가 평균적으로 4.1% 증가하고 ENSO Neutral 시기에 단수가 평균적으로 0.3% 증가했으며 라니냐 시기에 단수가 평균적으로 9.2% 하락한 모습을 보였다. 타 작물에 비해 라니냐에 강하게 영향을 받는 모습을 보였다. 라니냐와 엘니뇨와 같은 이상기후는 태평양 해수면 온도지수를 통해 추정된다. 일반적으로 대표적 태평양 해수면 온도지수인 중앙 태평양의 온도를 지수화한 ONI 지수가 -0.5 이하면 라니냐(La Nina), 0.5 이상이면 (El Nino), 그 사이면 ENSO Neutral로 판단한다. 따라서 태평양 해수면 온도지수가 이상기후를 나타내는 지표가 될 수 있다는 가정하에 태평양 해수면 온도지수를 두 번째 변수로 사용해 단수를 추정하고자 시도했다.

마지막으로 단수를 조기에 추정해 국제 옥수수 선물시장의 불확실성을 감소시켜 Hedger와 Speculator 모두 의사결정을 보다 효율적으로 할 수 있을 것이라 생각했다. 일반적으로 옥수수와 같은 상품 선물 시장은 선물보다 현물 가격이 큰 Backwardation 양상이 주로 나타난다. 그 이유는 옥수수와 같은 곡물 시장은 작황에 대한 불확실성이 커 농민과, 옥수수를 주 원료로 사용하는 기업들이 선물을 통해 Hedging을 함으로써 불확실성에 대한 Risk Premium을 지불하기 때문이다. NDVI와 태평양 해수면 온도지수를 통해 작황을 추정하면 작황에 대한 불확실성이 감소 할 수 있을 것이라는 가정하에 연구를 진행했다.

**2. 선행 연구 검토**

Pragathi Tummala., [[3]](#footnote-3)은 DNN을 사용하여 Crop Year, Season, Crop, area, NDVI, backscatter values 등을 Input Layer 로 설정하여 Hidden layer를 발견하고 최종적으로 산출량을 예측하는 모델을 구현했다. 이 모델은 약 95.47%의 정확도를 보여줬다.

김나리., [[4]](#footnote-4)MODIS 위성영상, PRISM 기후자료, GLDAS 토양수분 자료 및 옥수수 수확량 통계자료를 수집하여 Enhanced Vegetation Index (EVI), Leaf Area Index (LAI), Gross Primary Production (GPP), 강수, 최대기온, 토양수분의 6개 변수들로 2012년 미국 중서부 지역의 극심한 가뭄일 때를 대상으로 분석을 수행하였다. Multivariate Adaptive Regression Splines(MARS), Support Vector Machine (SVM), Random Forest (RF), Extremely Randomized Trees (ERT), Artificial NeuralNetwork (ANN), Deep Neural Network (DNN)의 총 6개 인공지능기법을 활용하여 결과를 비교한 결과 DNN 모델이 옥수수 수확량 예측에 있어 가장 높은 정확도를 보여준 것으로 나타났다.

장원진., [[5]](#footnote-5)NDVI, Land Surface Temperature (LST) , 강수량, 일조시간과 토양수분 5개의 자료를 활용하여 Python TensorFlow 기반의 RNN-LSTM 모델을 구축하여 국내 78개 지점에 대해 토양수분을 산정하였다.

Zhonglin Ji., [[6]](#footnote-6) 가장 주요하게 참고한 논문이다. Corn belt 지역의 12개 State의 766개 Agriculture County를 먼저 선별하고, Corn belt 지역에서 NASS CDL 을 활용해 Corn, Soybean을 분류했다. 이후 County를 식물 생장 시기(greenup day)별로 분류하여 시계열 타임라인을 재조정한 뒤MOD09Q1 에서 2008년부터 2018년까지 10년간 County 단위 NDVI와 EV12지수를 수집해 해당 지수의 변화율과 지수 자체를 모두 사용하여 County 단위 단수를 추정했다. 연구결과는 EV12를 활용해 지수와 지수의 변화율을 모두 변수로 사용한 모델이 R-square 0.73으로 가장 높은 모습을 보여줬다.

이 외에 깃헙 등을 통해서 다양한 코드를 ChatGPT-4 를 사용해 조사 및 분석했다.

**3. 연구 과정 요약**

**<그림 1> 연구 과정 요약도**

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

위의 그림은 연구과정을 간략하게 요약한 그림이다. 먼저 미국 정부의 [웹사이트](https://www2.census.gov/geo/tiger/TIGER2022/COUNTY/)에서 County 경계 데이터가 담긴 shp 파일을 다운로드 받아 QGIS를 통해 shp 파일을 수정하여 Corn belt 12개 State의 각 County별 경계 데이터가 담긴 geojson 파일로 변환했다. 이를 구글 Earth Engine에서 다운로드 받은 State 단위 NDVI 데이터에 적용하여 카운티 단위 NDVI 데이터를 최종 획득했다. 이후 이를 ONI, NINO12 데이터와 GAF 이미지로 합성하고, 이 데이터를 Numpy Array에 저장하여 모델의 하이퍼 파라미터를 변경하고, 데이터 스케일링 및 필터링 방법을 변경해가며 최적의 추정치를 획득하고자 노력했다.

**Ⅱ. 연구 과정 및 방법**

**1. Google earth engine 및 State 단위 위성사진**

본 연구는 분석을 위해 콘벨트 지역과 주요 남아메리카의 옥수수 수확지대의 NDVI 지수를 필요로 한다. NDVI(Normalized Difference Vegetation Index, 정규화 차이 식생지수)는 근적외선의 반사율 차이를 이용하는 식생지수로, 격자 단위의 식물의 분포 또는 식생의 활력도를 나타낸다. NDVI의 값은 -1.0에서 1.0의 범위를 가지며 보통 물이나 눈은 음수, 식생이 없는 지역은 0, 식생이 풍부한 지역은 0.3 이상의 양수 값을 가진다(장원진 외, 2019). 본 연구에서 수집된 수확시기와 및 그 전후의 NDVI 데이터는 해당 시기의 옥수수 수확량을 예측하는데 사용된다.

본 연구의 NDVI 분석을 위한 데이터는 구글어스 엔진(Google Earth Engine)의 위성사진을 이용하려 했으나 county별 NDVI 위성사진 데이터를 추출하는 데 너무나 많은 시간이 소요되어 사용하지 못하였다. 또한 state별 NDVI가 추출된 이미지 데이터를 사용하고자 했으나 데이터의 수가 부족하여 다른 데이터로 대체하였으며 구체적인 내용은 후술하겠다.

Google earth engine을 이용한 데이터의 분석은 Google earth engine 콘솔에서의 javascript를 활용한 ee 클래스 및 라이브러리와 파이썬의 ee와 geemap 클래스, Tenserflow 및 keras 라이브러리를 통해 정규화하고 사용한다. 먼저, google colab을 통한 데이터 수집 전 google earth engine의 콘솔에서 javascript를 이용하여 수집되는 위성사진을 미리 확인할 수 있었다.

**<그림 2>Google Earth Engine**

텍스트, 지도, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

다음 위성사진은 Google earth engine 콘솔에서 추출한 미국 Kansas 주 일부의 NDVI 위성사진이다. Google earth engine에서의 NDVI 위성사진은 green-brown scale로 시각화되어 있지만 google colab을 통해 추출한 위성사진은 gray scale로 이루어져 있다.

**<그림 3 Google EarthEngine Code>**

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

**<그림 4 Kansas 주 NDVI 사진 >**

천문학, 우주, 어둠, 블랙이(가) 표시된 사진

자동 생성된 설명

위 코드와 위성사진은 google colab을 통해 추출한 Kansas 주 NDVI 위성사진 일부이며. Gray scale로 이루어진 것을 알 수 있다. 파이썬 코드에서 해당 이미지를 불러온 뒤 모델의 input으로 삽입하려 하였다. 사진을 확인하기 위한 알맞은 시각화가 이루어진 데이터가 아니기 때문에 문서에 업로드된, gray scale로 이루어진 사진에서 NDVI 지수의 차이를 육안으로 확인하기 쉽지 않다. 하지만 이미지 데이터를 google drive상에서 확인하면 위성사진에 NDVI 데이터가 알맞게 추출된 것을 알 수 있었다. 해당 사진에서 완전한 검정색으로 이루어진 부분은 식생지수가 아주 낮거나 해당 주의 경계 밖에 위치한 데이터이며, 완전한 흰색으로 이루어진 부분은 담수 지역이라고 볼 수 있다.

또한 태평양 해수면 온도와의 상관관계 분석을 위해 태평양 해수면 온도지수를 필요로 한다. 보통 태평양 지역의 해수면 온도지수는 엘니뇨 및 라니냐 현상을 분석, 예측하는데 사용된다. 엘니뇨의 경우 감시구역 해수면 온도의 3개월 이동평균 편차가 0.5도 이상, 라니냐의 경우 -0.5도 이하일 경우 시작되었다고 보는데, 본 연구에서 태평양 해수면 온도뿐만 아니라 엘니뇨 및 라니냐와 옥수수 단수와의 상관관계에 대한 분석도 추가적으로 진행하였다. 태평양 해수면 온도지수는 APCC에서 데이터를 추출하였으며 그 중 ONI(해양 니뇨 지수)와 NINO(극동 태평양 온도지수)를 사용하였다.

**2. County 경계 데이터**

County 경계 데이터를 획득하기 위해 QGIS를 활용하였다. 지오메트리 데이터가 포함되어있는 shp 파일을 전처리하여 활용 할 수 있게 만들기 위해 QGIS를 사용했다. QGIS에서 STATEFP(State Fips Code)를 조정하여 Corn belt 지역의 12개 State를 지정하여 총 1053개 County의 지정학적 경계 정보가 있는 geojson 파일을 획득 할 수 있었다. 아래 그림은 QGIS에서 17 ( 아이오와 ) 지역의 지정학적 경계를 획정하고 있는 실행 화면이다.

**<그림 5 QGIS IOWA주 지정학적 경계>**

**텍스트, 스크린샷, 소프트웨어, 컴퓨터 아이콘이(가) 표시된 사진

자동 생성된 설명**

**3. county 단위 NDVI 및 yield 데이터**

**텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명** 본 연구에서 더 많은 데이터를 활용하기 위하여 state별이 아닌 county별 NDVI의 데이터를 사용하였다. County별 NDVI는 QGIS(Quantum GIS)를 사용하였고 state 및 county별 NDVI 데이터를 추출할 수 있었다. **<그림 6 Kansas NDVI Data>**

위 사진은 QGIS에서 추출한 Kansas 주의 county별 NDVI 데이터의 일부이다. STATEFP는 식별 가능한 고유의 state ID number이며 COUNTYFP는 해당 state에 속한 county의 고유 ID number이다. 추후의 분석에서는 해당 데이터에 minmax scaler를 활용한 정규화를 적용하여 사용하였다.

또한 모델의 종속변수 Y 데이터로 들어가는 county별 단위수확량 데이터를 추출하였다. County별 단위수확량은 미국 농무부의 데이터를 활용하였다.

**<그림 7 County 단위 단위수확량 데이터>**

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

위 사진은 미국 농무부로부터 추출한 county별 단위수확량 데이터의 일부이다. 해당 데이터에서 column index의 XX-XX는 state ID number와 county ID number를 뜻한다. convLSTM 모델을 활용한 분석 과정에서 county별 단위수확량 데이터는 Y의 input 데이터로 활용되었다.

**4. 시도해봤으나 수집하지 못한 데이터들과 그 이유**

본 연구에서 고려했으나 수집하지 못한 데이터 중 가장 아쉬웠던 것은 county별 NDVI 위성사진이다. Google colab에서 Google earth engine의 API를 사용하여 county별 NDVI 위성사진 데이터를 추출하는 데에는 성공하였으나 추출해야 하는 총 county의 수가 1017개였고, 한 county당 1년치 위성사진 데이터를 추출하는 데 약 2시간 정도가 소요되었다. 처음에는 로컬에서 코드를 실행하는 방법으로 속도를 향상시키고자 하였으나 Google earth engine의 보안 정책으로 인해 로컬이 아닌 google drive의 디렉토리로만 다운로드를 받아야 했다. 다운로드 속도가 저하되었던 가장 큰 이유는 google earth engine으로부터 google drive로의 통신 속도가 지나치게 느렸던 것으로 판단된다.

추가적으로, AppEEARS를 이용하여 county별 위성사진을 추출하는 방법도 고려하였다.

**<그림 8 AppEEARS>**

**텍스트, 스크린샷, 소프트웨어, 웹 페이지이(가) 표시된 사진

자동 생성된 설명**

위 NASA에서 제공하는 AppEEARS(Application for Extracting and Exploring Analysis Ready Samples)에 앞서 언급한 Geojson 파일을 이용해 요청한 화면이다. 이를 통해 NDVI 데이터를 다운로드 받을 수 있지만 지나치게 오랜 시간이 소요되어 분석에 활용하지 못하였다. 약 1개 State 대상으로 약 4개월의 자료를 수집하는데 10시간 정도가 소요되어 12개 State의 22년의 자료를 수집하는 것이 목표였기 때문에 기간 내에 모든 데이터를 수집 할 수 없을 것이라 판단해 시도해봤으나 분석에 활용하지 못했다.

마지막 방법으로 county별 경계 위도 및 경도를 직접 구하여 추출했던 state별 위성사진에 layer를 씌워 county별로 나누는 것이 있었다. 위의 python 코드에서 ee 라이브러리 클래스의 FeatureCollection 메소드로부터 county별 경계를 구할 수 있는 방법이 존재하는 것은 인지하였으나 해당 방법을 찾지 못해 적용할 수 없었다.

다음으로 일별 태평양 온도지수를 추출하려 했으나 사용하지 못했다. 본 연구의 데이터 수집 과정에서 NOAA(National Oceanic and Atmospheric Administration, 미국 해양대기청)의 NCEI(National Centers for Environmental Information)에서 일별 해수면 온도지수 데이터를 추출하고자 하였다. 하지만 데이터 수집을 위한 API 요청 과정에서 알 수 없는 오류가 발생하였고, 결국 월별 해수면 온도지수를 사용하였다.

**5. 데이터 가공**

본 연구에서 사용한 데이터 처리방식은 GAF(Gramian Angular Field)이다. 농사에서 가장 중요하게 바라보는 변수인 기후와 식생에 대하여 각 고유 데이터 값을 온전히 유지함과 동시에 그 값이 어떻게 변화하는지 패턴을 파악하는 것이 옥수수 단위수확량의 추정에 있어 유의미할 것이라고 판단하였다.

GAF는 시계열 데이터를 시각화하는 방식 중 하나로 보통 Cartesian좌표계를 통해 표현되는 그래프 이미지와 달리 Polar좌표계를 사용한다. 이때 나타나는 각도의 합을 이용하는가, 차를 이용하는가에 따라 GASF(Summation), GADF(Difference)의 둘로 나뉘는데 본 연구에서는 위에서 설명한 바와 같이 시계열 패턴을 중심으로 분석을 진행하기에 이를 더 부각시키는 데에 특화된 GASF가 사용되었다.

**<그림 9 GAF image Sample> <그림 10 GASF Matrix>**

A picture containing text, font, handwriting, white

Description automatically generatedA picture containing colorfulness, screenshot, pattern, purple

Description automatically generated

위의 이미지는 2000년도 County 17-1의 2000년 11개 시간대별 NDVI데이터를 이미지화 한 것이다. 각각의 픽셀은 각각 i, j번째 시간쌍의 Polar좌표계 각도 합을 의미한다.

이는 좌상단부터 우하단까지의 대각선이 시간의 흐름을, 좌하단부터 우상단까지의 대각선이 데이터의 각 값을 우회적으로 표현함을 나타낸다.

**6. 고려했으나 적용하지 않은 데이터 처리방식과 그 이유**

본 연구에서 고려했으나 적용하지 않은 데이터 처리방식은 IQR box plot, Robust filter, SG filter가 있다.

먼저 IQR box plot, Robust filter, SG filter는 데이터 중 아웃라이어를 처리하기 위한 방법으로 고려되었다. 위성 데이터를 기반으로 한 NDVI 데이터에서 구름 등의 이유로 NDVI가 지나치게 낮은 값을 가지는 데이터가 있을 수 있다고 판단했기 때문이다.

IQR box plot은 데이터의 중앙값을 중심으로 25%만큼 낮은 데이터부터 25%만큼 높은 데이터까지 box에 포함시키고 아웃라이어라고 판단되는 데이터를 제외한 maximum과 minimum까지를 측정하고 이것을 upper quartile, lower quartile의 1.5배만큼 벗어난 데이터를 아웃라이어라고 판단한다. Robust filter와 SG filter는 데이터를 스무딩하여 아웃라이어 데이터를 완화시켜준다.

**<그림 11 IQR Box plot> <그림 12 SG Filter>**

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명텍스트, 폰트, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

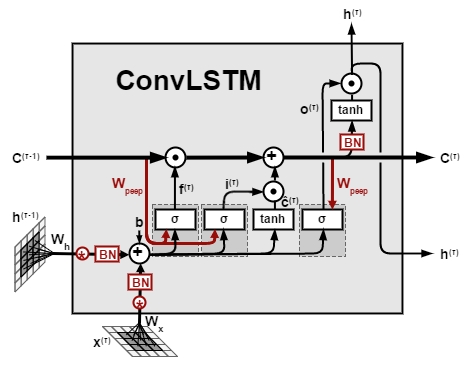
해당 방법들을 사용하여 아웃라이어를 잡아내지 않은 가장 큰 이유는 구름 등으로 지나치게 낮은 값을 가지는 NDVI 아웃라이어 데이터는 처리할 수 있지만, 다른 데이터보다 아웃라이어로 판단될 만큼 높다고 여겨지는 NDVI 데이터가 유의할 수 있다고 판단하였다. 즉, 해당 데이터 처리 방식에서 maximum 값을 벗어난 아웃라이어 데이터가 실제로는 유의한 데이터로 작용할 수 있다는 것이다. 실제로 본 연구에서 사용한 분석 방법에서 NDVI 값이 높은 옥수수 식생지대를 중점적으로 잡아내야 하는데 input 데이터에서 유의한 NDVI 데이터가 손상되는 경우가 생길 수 있는 것이다. 따라서 IQR box plot 및 Robust filter, SG filter를 이용한 아웃라이어 데이터 처리는 진행하지 않았다.

**Ⅲ. 모델 선정**

**A. ConvLSTM**

위에서 준비 및 가공된 모델을 기반으로 ConvLSTM모델을 사용하여 학습을 진행하였다. 앞의 데이터 가공과정에서 언급한 바와 같이 이번 분석에서 가장 중요한 변수는 매년 옥수수 생장기간에 나타나는 옥수수의 생장과 기후의 패턴이라고 보았다. 하지만 이 데이터가 고유하게 가지고 있는 시계열적 특성 또한 무시할 수 없는 중요한 지점이었다. 이에 선택하게 된 것이 ConvLSTM이다.

**<그림 13 ConvLSTM>**



ConvLSTM은 CNN을 통해 시계열로 주어지는 2차원 이미지파일(3D Tensor Input)의 Feature를 추출하고 이를 LSTM학습하는 모델이며 각각의 분야에서 특화된 두 모델을 결합하였다는 점에서 한 해에 나타나는 옥수수의 생장 및 태평양 온도지수라는 기후의 패턴과 이들이 20년간 보인 시계열적 특성을 모두 반영하여 좋은 학습이 가능할 것이라고 보았다.

**<그림 14 Keras ConvLSTM Model 구축>**

**A screen shot of a computer code

Description automatically generated with low confidence**

ConvLSTM의 코드는 Tensorflow에서 제공하는 Keras라이브러리를 통해 작성하였다. 입력 값인 GAF이미지가 시간별로 가로 세로 20픽셀도 넘어가지 않는 작은 크기를 가지고 있기에 하나의 이미지를 분석하는 데에는 크게 시간이 오래 걸리지 않는다는 점과 주어지는 시간대가 21개년으로 적다는 점을 감안하여 최대한 학습력을 높일 수 있도록 Layer의 수치를 CNN분석에서는 64, LSTM분석에서는 50으로 상당히 높은 값을 사용하였다. 또한 아래에 나타나는 Dense(1017)은 1017개의 County에 대하여 분석을 진행하였음을 나타낸다.

**<그림 15 Keras Model Training>**

**A screen shot of a computer code

Description automatically generated with low confidence**

Batch\_size 또한 일반적으로 많이 사용되는 수치인 32를 사용하였으며 Epoch은 100을 기준으로 Early Stopping을 사용하여 Overfitting을 방지하고 가장 Validation Set의 Loss(MSE)가 낮은 모델을 찾을 수 있도록 하였다.

**B. 고려해보았으나 적용하지 못한 모델과 그 이유**

분석 데이터의 시계열적 특성에 따라 추가로 Facebook에서 제공하는 Prophet, Amazon에서 제공하는 Autogluon과 같은 라이브러리의 사용을 고려해보았으나 이들의 경우 간단하게 코드 구현이 가능한 ConvLSTM과 달리 이미지 데이터를 CNN분석하여 얻은 Feature를 입력으로 제공하기 쉽지 않았기에 사용에 어려움을 겪다가 결국 중도 포기하게 되었다.

위의 Keras라이브러리에서 제공하는 방법론을 그대로 활용하되 CNN에서 Feature를 추출하는 부분에서 기존에 사용된 Conv2D보다 더 학습력이 좋다고 알려져있는 Residual Neural Network-50(ResNet50)을 활용하는 방안 또한 고려해보았다. 하지만 학습력이 좋아지는 만큼 메모리의 사용 또한 늘어나 Colab Pro에서 제공하는 메모리 용량을 넘어 Crash가 반복되었기에 이 또한 활용하지 못하였다.

**Ⅱ. 결론 및 한계점**

**1. 분석결과 및 해석**

분석은 GAF이미지에 사용된 데이터의 조합과 종류, Maxpooling여부에 따라 총 12번을 진행하였다. Maxpooling은 다운샘플링을 위한 기법으로 이 과정에서 특정 패턴들이 묻힐 수도 있겠다는 가정 하에 이 과정을 제외하고 학습을 진행하였으며 그 결과는 다음과 같다.

|  |  |
| --- | --- |
| **Maxpooling O (Validation set MSE)** | |
| NDVI+NINO12+ONI (RGB, 3채널) | 721.4478 |
| NDVI+NINO12 (2채널) | 716.4293 |
| NDVI+ONI (2채널) | 802.4728 |
| NDVI (1채널) | 722.2510 |
| NDVI+범주화 NINO12 (1채널+범주형 데이터) | 909.7753 |
| NDVI+범주화 ONI (1채널+범주형 데이터) | 945.7391 |

|  |  |
| --- | --- |
| **Maxpooling X (Validation set MSE)** | |
| NDVI+NINO12+ONI (RGB, 3채널) | 913.1200 |
| NDVI+NINO12 (2채널) | 805.9282 |
| NDVI+ONI (2채널) | 1362.2601 |
| NDVI (1채널) | 786 |
| NDVI+범주화 NINO12 (1채널+범주형 데이터) | 1153 |
| NDVI+범주화 ONI (1채널+범주형 데이터) | 954.9922 |

위의 결과를 보면 Maxpooling을 진행한 결과의 MSE값들이 전체적으로 더 낮은 모습을 보여준다. Maxpooling이 다운 샘플링을 통해 몇몇 특징이 감소하였을 수는 있으나 보통의 경우 학습의 일반화에 더 큰 강점을 지니는 기법이기에 이러한 결과가 나타난 것으로 보인다.

**<그림 16 Validation Set Loss (Maxpooling O)> <그림 17 Prediction 10%, 15% Error Graph>**

A picture containing text, screenshot, font, graphic design

Description automatically generated스크린샷, 텍스트, 디자인이(가) 표시된 사진

자동 생성된 설명

하지만 그럼에도 불구하고 여전히 MSE는 전체적으로 매우 높으며 이를 통해 우리의 가정을 통한 학습이 완전히 실패하였음을 알 수 있다. 가장 낮은 MSE인2채널 NDVI+NINO12데이터를 통한 분석에서도 MSE는 716.4293이라는 높은 값을 보여주고 있으며 좌측의 그래프에서 확인할 수 있듯 실제 값과 추정치가 각각 10%, 15% 내의 차이를 보이는 결과들이 2020년 이후 50%를 넘기지 못함을 볼 수 있다.

**2. 개선가능한 부분**

**A. 데이터 문제**

본 분석에서는 Corn Belt에 포함된 모든 State의 1017개 County 데이터를 모두 사용하였다. 연도별 데이터양이 풍부하지 못하여 최대한 데이터를 보장하기 위한 선택이었으나 오히려 이로 인하여 데이터의 학습능력이 하락했을 가능성이 있다. Corn Belt에 위치하지만 옥수수 밭이 아닌 County가 존재하며 이에 선행연구에서도 목적에 맞는 County를 선별하는 과정을 거쳤음을 고려하여 본 분석에서도 County 선별 과정이 수반된다면 데이터의 수는 적어질 수 있으나 그 퀄리티를 높여 학습에 더 도움을 줄 수 있을 것으로 판단된다.

이와 동시에 더 높은 퀄리티의 NDVI데이터를 구하는 것 또한 분석에 긍정적인 영향을 줄 수 있을 것이다. 본 연구에서는 시간적 한계로 인해 GEE에서 제공하는 State별 NDVI위성사진을 사용하였다. 하지만 앞서 언급한 NASA 제공의 AppEEARS서비스를 사용하여 보다 고해상도의, 노이즈 수준이 낮은 NDVI데이터를 수집한다면 실제 옥수수 식생과의 오차를 줄일 수 있어 학습에 도움이 될 것이다.

**B. 모델 문제**

앞서 설명한대로 본 연구에서는 데이터의 크기를 제어하지 못하여 Keras의 ResNet을 사용하지 못하였다. ResNet은 보통 Conv2D를 통해서 구현되는 CNN보다 더 높은 학습 능력을 가진다고 알려져 있기에 데이터가 갖는 특징에 손상이 가지 않는 선에서 데이터의 간소화가 충분히 이루어진다면 이를 통한 더 좋은 학습이 가능할 것으로 생각된다. 또한 Keras의 경우 ResNet-50을 최소치로 기본 제공하고 있기에 입력으로 주어지는 이미지의 픽셀밀도와 컴퓨팅 파워의 한계를 고려하여 ResNet-34를 적절히 코드로 구현하는 것 또한 Overfitting을 방지하면서 좋은 학습을 진행하는 방안으로 고려할 수 있다.

AutoGluon의 사용도 개선에 도움을 줄 수 있다. AutoGluon은 AutoML을 통하여 받은 입력에 대해 어느 모델, 어느 분석법이 가장 좋은 결과를 도출 할 수 있는지 스스로 판단하고 해당 모델을 통한 추정치를 도출한다. 본 분석에서는 ConvLSTM만을 사용하였고 이미지를 AutoGluon에 입력으로 제공하는 법을 찾지 못하여 이를 미처 사용하지 못하였으나 원본 Tabular데이터형태 혹은 Feature로 가공된 Image의 Tabular를 제공하여 분석을 진행한다면 기존의 방식보다 더 나은 모델을 찾아낼 수도 있다.

**3. 결론**

본래 이 연구의 최종적인 목적은 최소한의 정확도가 확보된 예측에 성공하여, 이를 옥수수 선물시장에서 의사결정 하는데 활용하는 것이었다. 하지만 예측치의 정확도가 매우 낮아 이는 현재로서는 불가능해 보인다. 다만, 데이터와 모델 부분이 개선되어 분석이 보다 정교해 진다면 의사결정을 이 지표만 가지고 온전히 내리지는 못하더라도 하나의 지표로 활용 할 수 있을 것으로 보인다. 또한 후속연구에서 분석 분야를 변경해 국내 농업지대의 식생지수를 대상으로 연구를 진행한다면 쌀 수급요인을 보다 정교하게 분석하여 최근 대두되고 있는 쌀 과잉공급과 같은 문제를 조기에 탐지 할 수 있을 것이라 생각한다. 드론을 활용하여 보다 정교한 식생지수를 획득하거나, County 단위가 아닌 농장 단위로 분석하여 모델을 구축한다면 불가능한 일은 아닐 것이라 생각한다.

마지막으로 이 프로젝트는 최종적으로 실패로 끝이 났지만, 이 프로젝트의 의의를 간략하게 언급하고 마무리하겠다. 이번 프로젝트는 옥수수 단수 추정에 대한 프로젝트였기 때문에 어떤 Feature를 사용해야 하는지, Raw Data를 어떤 식으로 가공해 데이터셋을 만들고 어떤 필터를 적용해 이상치나 결측치 문제를 해결할지, 시계열 데이터 분석이기 때문에 시계열적 Trend는 어떻게 해결할지 등 다양한 문제점에 대해 고민했다. 금융 데이터에서도 이와 같은 과정은 똑같을 것이라 생각한다. 어떤 Feature를 사용해야 하고, Raw Data를 어떻게 가공하여 분석에 용이하게 만들지에 대한 고민은 금융 데이터와 농업 데이터가 크게 다르지 않을 것이다. 이 프로젝트는 데이터를 어떻게 수집할지 , 데이터셋을 어떻게 구축할지, 어떻게 가공할지, 마지막으로 어떻게 분석 할 지의 전 과정을 기획하고 시도해봤다는 것에 의의를 둔다.

**참고문헌**

이병훈. (2015). 2015년 엘니뇨(El Nino) 발생과 세계 곡물시장 동향 및 전망. 해외곡물시장동향, 4(7), 89-100.

이병훈. (2016). 2016년 하반기 라니냐(La Nina) 발생과 국제 곡물시장 전망. 해외곡물시장동향, 5(4), 113-126.

윤병삼. (2020). 옥수수 선물시장의 Fundamental 분석. 해외곡물시장동향, 9(6), 81-105.

김나리, 이양원.(2019).특이기상 사례에 대한 인공지능기반의 곡물수확량 예측 연구.대한공간정보학회 학술대회,(),185-187.

장원진, 이용관, 이지완, 김성준. (2019). Terra MODIS NDVI 및 LST 자료와 RNN-LSTM을 활용한 토양수분 산정. 한국농공학회논문집, 61(6), 123-132. 10.5389/KSAE.2019.61.6.123

Zhonglin Ji, Yaozhong Pan, Xiufang Zhu, Dujuan Zhang, Jinyun Wang, A generalized model to predict large-scale crop yields integrating satellite-based vegetation index time series and phenology metrics, Ecological Indicators, Volume 137, 2022, 108759,

P. Tummala, M. Sobhana and S. Kakumani, "Predicting crop yield with NDVI and Backscatter values using Deep Neural Networks," 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 2022, pp. 390-394, doi: 10.1109/MECON53876.2022.9751969.

Kim N, Na S-I, Park C-W, Huh M, Oh J, Ha K-J, Cho J, Lee Y-W. An Artificial Intelligence Approach to Prediction of Corn Yields under Extreme Weather Conditions Using Satellite and Meteorological Data. Applied Sciences. 2020; 10(11):3785. <https://doi.org/10.3390/app10113785>

SaiaraSamiraSajid,”Corn-Yield\_Prediction\_US\_CornBelt”,Github,2022, <https://github.com/SaiaraSamiraSajid/Corn_Yield_Prediction_US_CornBelt/blob/main/Corn_Yield_Prediction_12States_Github.ipynb>

1. 윤병삼. (2020). 옥수수 선물시장의 Fundamental 분석. 해외곡물시장동향, 9(6), 94-100. [↑](#footnote-ref-1)
2. 이병훈. (2016). 2016년 하반기 라니냐(La Nina) 발생과 국제 곡물시장 전망. 해외곡물시장동향, 5(4), 113-126. [↑](#footnote-ref-2)
3. Pragathi Tummala, M Sobhana, Sruthi Kakumani. (2022) Predicting crop yield with NDVI and Backscatter values using Deep Neural Networks, 2022 International Mobile and Embedded Technology Conference (MECON), 390-394 [↑](#footnote-ref-3)
4. 김나리, 이양원.(2019).특이기상 사례에 대한 인공지능기반의 곡물수확량 예측 연구.대한공간정보학회 학술대회,(),185-187. [↑](#footnote-ref-4)
5. 장원진, 이용관, 이지완, & 김성준. (2019). Terra MODIS NDVI 및 LST 자료와 RNN-LSTM을 활용한 토양수분 산정. 한국농공학회논문집, 61(6), 123-132. 10.5389/KSAE.2019.61.6.123 [↑](#footnote-ref-5)
6. Zhonglin Ji, Yaozhong Pan, Xiufang Zhu, Dujuan Zhang, Jinyun Wang, A generalized model to predict large-scale crop yields integrating satellite-based vegetation index time series and phenology metrics, Ecological Indicators, Volume 137, 2022, 108759 [↑](#footnote-ref-6)