Technical University of Applied Sciences Würzburg-Schweinfurt (THWS)
Faculty of Computer Science and Business Information Systems

Master Thesis

Transfer Learning in Crop Remote Sensing

submitted to the University of Applied Sciences Würzburg-Schweinfurt in the Faculty of Computer Science and Business Information Systems to complete a course of studies in Master on Data Science

Ana Muñoz Gutiérrez

Submitted on: March 23, 2023.

Initial examiner: Prof. Dr. rer. Magda Gregorová
Secondary examiner: Prof. Dr. Edgar Román

Abstract

Wheat is one of the most important crops to feed the population. Regardless of its importance, obtaining global estimates of the total amount of wheat being cultivated each year before the harvest season is difficult due to lack of quality data. Recently, machine learning methods relying on satellite imagery have been explored for this purpose. This work focuses on the use of deep learning techniques and in particular on transfer learning methods for predicting winter and spring wheat production based on remote satellite data and agricultural nacional statistics on the US territory.

The results compared to the state of the art, show an overall better performance in various metrics. Additionally, our results demonstrate that transfer learning to different crops cannot be achieved with our approach.

This work was developed under the supervision of Prof. Dr. Magda Gregorová in collaboration with the green spin company.

Acknowledgment

I would like to thank the Green spin and CAIRO teams. Especially to Denise, Gunther and Joe. During the last six months they have not only teched me a lot about wheat agriculture and remote geodata, but also gave me a warm welcome to Germany.

I'm deeply indebted to Prof. Dr. Magda Gregorová for giving me the opportunity to work on this project, for her mentorship and her time.

I am extremely grateful to my friends and family for their support. Especially to my sister, mother and father for always supporting me in my life and career.

I would also like to thank the professors at ITAM for giving me the education and tools to grow in this field. Specially to Edgar Roman for sharing his passion on deep learning with students as me, and reaching me with Magda, making this collaboration possible. Also to my aunt Ana Lidia Franzoni for giving me advice and helping me find my way into data science.

I can not forget to thank the BAYLAY program for giving me the privilege of their scholarship that finiced big part of my stay in Germany.

Lastly I would like to thank my friends in Mexico and to my friends in Würzburg, that have shared love and support to me.

Contents

1 Introduction

1.1 Importance of wheat

Wheat is one of the most important crops by many reasons: Firstly, it is a staple food for a large portion of the world's population. Wheat is a major source of dietary energy and protein for people in many countries, especially in developing nations. It can be consumed in various forms, such as bread, pasta, noodles, and cereal.

Secondly, wheat is a versatile crop that can be grown in different types of soils and climates. It is widely cultivated across the world, with major producers including China, India, Russia, and the United States as show in fig.1.1.

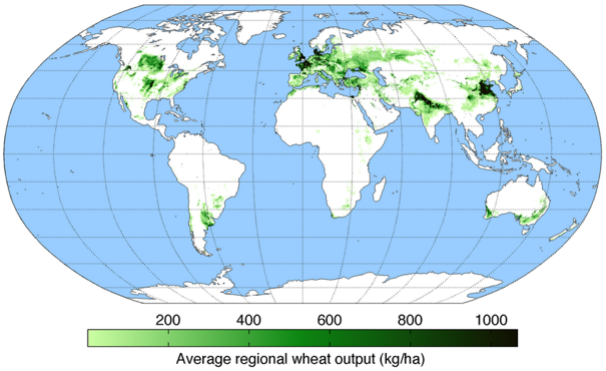


Figure 1.1: World areas dedicated to wheat production.

Thirdly, wheat is an important source of income for farmers and contributes significantly to the economy of many countries. The global trade in wheat is worth billions of dollars each year, and many countries depend on wheat exports for their economic growth.

Finally, wheat is also important for its nutritional value. It is a good source of essential nutrients, including fiber, vitamins, and minerals.

Overall, wheat plays a critical role in global food security, economic development, and human health, making it an essential crop in the world today.

1.2 Varieties of wheat

In the USA and Europe there are two varieties of wheat grown in two seasons: winter wheat and spring wheat.

Winter wheat

Winter wheat is a variety of wheat that is planted in the fall and harvested in the summer of the following year. It is called "winter" wheat because it is planted in the fall, before winter sets in. The cold temperatures and short days of winter allow the wheat to establish its root system and start growing slowly, even under snow cover.

Winter wheat is usually sown between September and November, depending on the location and climate, and is harvested in June or July. It is typically grown in regions with a cold winter and a moderate climate, such as the Great Plains of the United States, Canada, and Europe.

Spring wheat

Spring wheat on the other hand, is a type of wheat that is planted in the spring and harvested in the late summer or early fall. It is one of the major types of wheat grown in North America, particularly in the northern Great Plains region. Spring wheat is well-suited for cool, moist climates and is often grown in areas where winter wheat cannot be cultivated.

Spring wheat has a higher protein content than winter wheat, making it ideal for making bread and other baked goods that require a strong gluten structure. It also has a shorter growing season than winter wheat, which means that farmers can plant and harvest it in a shorter period of time. Some common varieties of spring wheat include hard red spring wheat, hard white spring wheat, and durum wheat, which is used to make pasta.

NDVI

NDVI stands for Normalized Difference Vegetation Index. It is a commonly used remote sensing index to measure vegetation growth and health. NDVI is calculated using the

reflectance of visible and near-infrared light, which is absorbed and reflected differently by healthy green vegetation and other surfaces like soil or water.

NDVI is used in various applications, including crop monitoring, land management, and climate studies, as it provides valuable information about vegetation growth, distribution, and stress.

NDVI curve for wheat

The seasonality growth of winter wheat can be captured by NDVI. In the early stages of growth, the NDVI values are low as the wheat is just emerging from the soil. As the plant grows and develops, the NDVI values increase, reaching a peak during the mid-season when the plant is at its healthiest and most productive. After the mid-season peak, the NDVI values gradually decrease as the plant matures and begins to senesce.

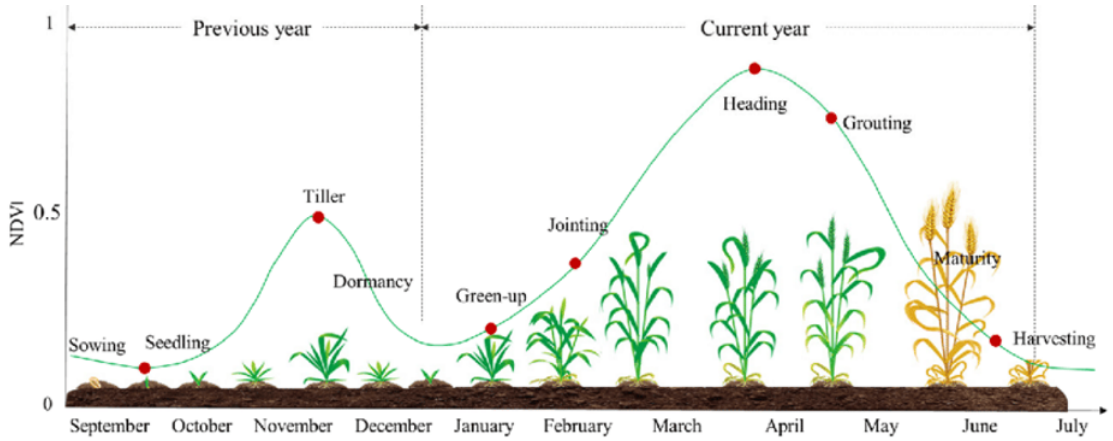


Figure 1.2: The growth cycle of winter wheat represented by the time series of remotely sensed normalized difference vegetation index (NDVI, green curve)

The same is true for spring wheat. With some differences in timing due to the differences in planting time and growing conditions as show in fig.1.3.

Overall, the NDVI curve for wheat provides valuable information about the crop's health and growth over time, and can be used to assess crop productivity, detect stress or disease, and guide management decisions [6].

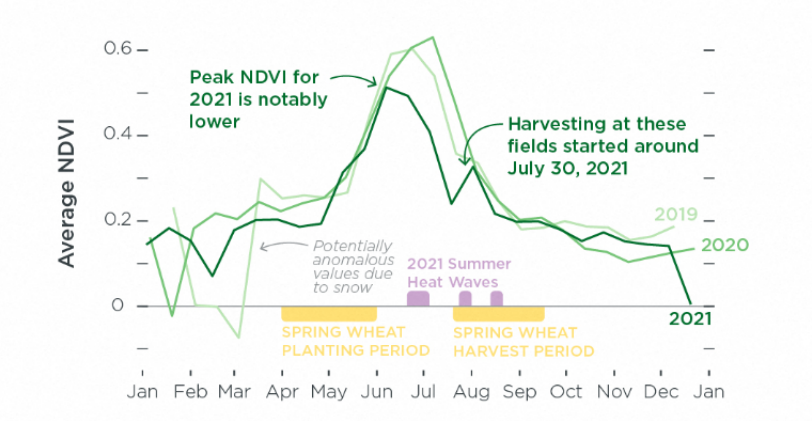


Figure 1.3: NDVI curve for some north countys of the US for years 2019-2021

1.3 Wheat remote detection

Regardless wheat's importance, obtaining global estimates of the total amount of wheat being cultivated each year before the harvest season is difficult due to lack of quality data.

Recent work by Liheng Zhong and collaborators (2019) has shown that it is possible to monitorate and create winter wheat maps (images that indicate the amount of wheat in an area) using satellite data with NDVI and agricultural statistics to train deep neural networks. The area of study of this work was the state of Kansas and north Texas during 2001-2017. Their models were able to detect the seasonality pattern of winter wheat and make reasonably good predictions.

Our work is strongly based on this previous publication and extends on it. This previous work has only focused on studying winter wheat, omitting spring wheat and has been limited to the States of Kansas and Northern Texas. In addition, it failed to address the discrepancies between their training and validation data. In Chapter 3 we explain more in depth the procedure and in Chapter 4 the limitations of their experiments and our proposal.

1.4 Motivation and proposal

The aim of this project is to extend the previous work to different geographical zones and also taking into account spring wheat. With the long term goal of applying this tecnics in Europe. The main challenge for this, is the lack of validation data outside the US to assess the quality of the predicted wheat maps.

Our work seeks to address this by exploring the viability of applying transfer learning from different geographical zones. However, before transferring to other zones where validation data is not available. We first did transfer learning from a neural network trained in the south region of the US to the north region to measure the performance of the transferred results. By the south region of the US we refer to some counties of the states of Kansas, Texas, Colorado, New Mexico and Oklahoma. By North region we refer to some counties in Washington, Oregon, Idaho, Montana, North Dakota and Minnesota. As shown in figure 1.4, where the North region is represented in green and the south region in blue. More details on data and methods is described in detail in Chapter 5.

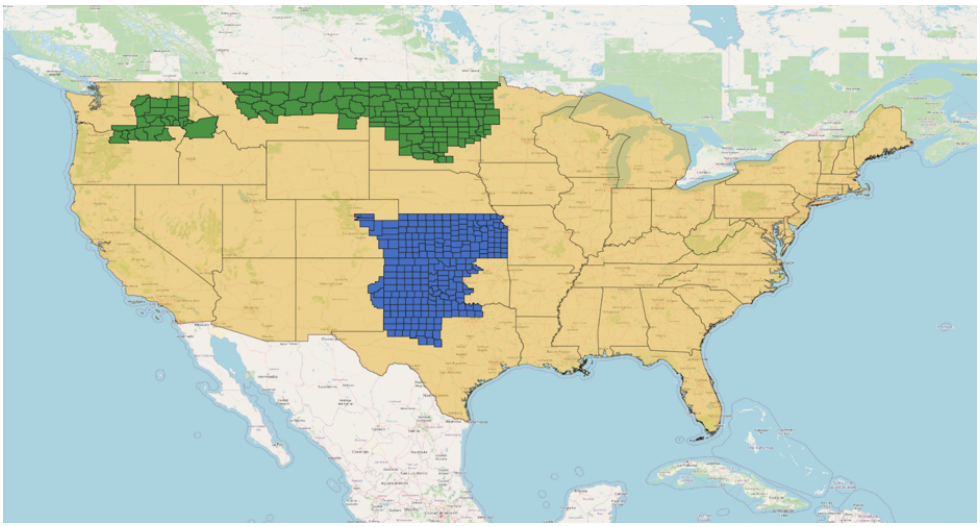


Figure 1.4: North (green) and south (blue) countys used for the project.

1.5 Structure of the rest of the script

The rest of the manuscript is organized as follows:

- Chapter 2 describes the data used in the previous work and on this project.
- Chapter 3 reviews the state of the art.
- Chapter 4 explains state of the art limitations and the areas of opportunity that our work aims to fulfill.
- Chapter 5 explores the data preprocessing and models architectures.
- Chapter 6 analyzes the results of the models.

- Chapter 7 opens discussion of results and shows room for future work.
- Chapter 8 gives conslusions concludes the script.

2 Data

Before reviewing the previous work, its limitations and our proposal. Is convenient to first take a look at the data used in both works.

We have three different sources of data that contribute to the models. First, we have satellite data, MODIS Vegetation Index Products (NDVI and EVI) provided by the National Aeronautics and Space Administration (NASA). Then, we have crop statistics provided by the USDA's National Agricultural Statistics Service (NASS). Finally, we have the geospatial data product called the Cropland Data Layer (CDL) also provided by NASS. Now, we proceed to explain each one of them.

2.1 MODIS

MODIS stands for Moderate Resolution Imaging Spectroradiometer. The MODIS vegetation indices offer a reliable way to make spatial and temporal comparisons of vegetation canopy greenness, which is a composite feature of leaf area, chlorophyll, and canopy structure.

For this project we are using the product MOD13Q1 Version 6. This data is created every 16 days with a resolution of 250 meters. This product consists of two main vegetation layers: the Normalized Difference Vegetation Index (NDVI), and the Enhanced Vegetation Index (EVI). Both indices are produced from the reflectance in the red, near-infrared, and blue wavebands that have been atmospherically-corrected. The NDVI maintains consistency with NOAA's AVHRR NDVI time series record for climate and historical applications, while the EVI minimizes variations in canopy-soil and improves sensitivity over dense vegetation conditions, which is more accurate for areas with high biomass. To select the best pixel value from all the acquisitions during the 16-day period, the algorithm considers low clouds, low view angles, and the highest NDVI/EVI value.

Figure 2.1 displays an example of a MODIS layer.

The MODIS pixel values have a range from $-20,000,000$ to $100,000,000$. Subsequently, we scaled these values by a factor of 0.0001. Leaving the final range from $-2,000$ to

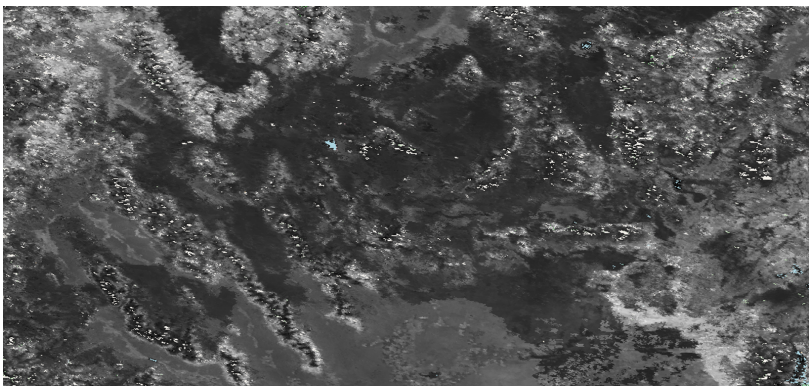


Figure 2.1: MODIS NDVI panel of north America.

10, 000.

2.2 NASS

The National Agricultural Statistics Service (NASS), carries out numerous surveys annually and generates reports on almost all aspects of American agriculture, such as the production and availability of food and fiber, farmer earnings, labor and wage data, financial data, chemical usage, and shifts in the demographics of American producers, among many others.

NASS is dedicated to delivering prompt, precise, and valuable statistical data that supports American agriculture. For example, they put effort in:

- Supply impartial and unprejudiced statistics at a scheduled time that is equitable and unbiased to all market contributors.
- Conduct the Census of Agriculture every five years, delivering the sole source of standardized, comparable, and comprehensive agricultural information for every county across the United States.

The data item that we used in this study is the survey of harvested acres of winter wheat plus spring wheat per county retrieved from the NASS. This data can be accessed in their portal Quick Stats. Figure 2.2 shows the portal of NASS and the product we selected for the study.

The NASS may merge counties that have relatively minor harvested areas in their statistical records, and these counties may not be valuable for analyzing data at the county

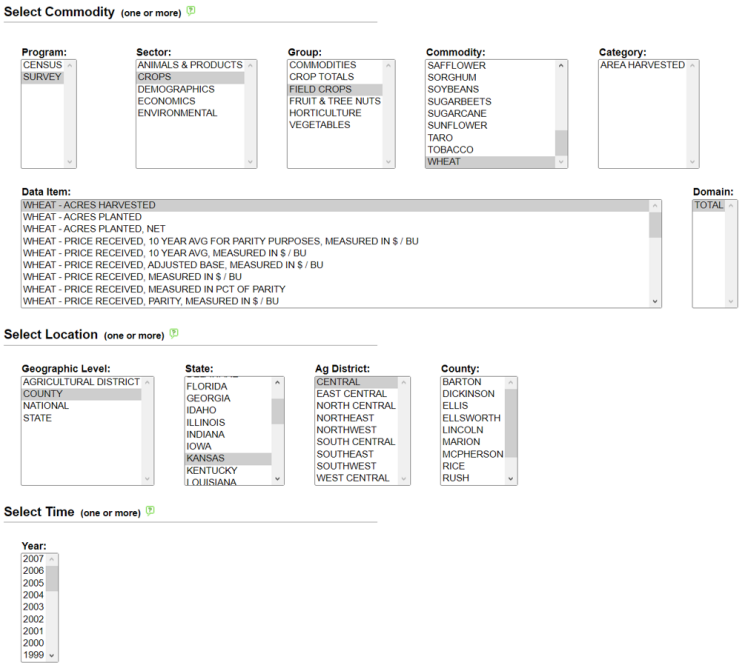


Figure 2.2: NASS portal of Quick Stats, wheat example of survey.

level. Consequently, county-level statistics may not be accessible for all counties each year.

The data was collected by green spin on their own private database. This database contains all the available NASS surveys for all US counties from 2000 to 2021.

2.3 CDL

The CDL is a raster, geo-referenced, crop-specific land cover data layer created annually for the continental United States. Is hosted on CropScape and was created by the USDA, National Agricultural Statistics Service, Research and Development Division, Geospatial Information Branch, Spatial Analysis Research Section. Figure 2.3 shows an example of the CDL layer.

The CDL provides a resolution of 30m. Each pixel of the CDL is identified as one of the 255 crop-specific land-cover classes. Each one is represented by a different color. This resolution allows us to even distinguish the individual fields as 2.4 shows.

The first layers of CDL have been available since 1997, but for the state of Texas (which is one of our areas of study) data is available since 2008.

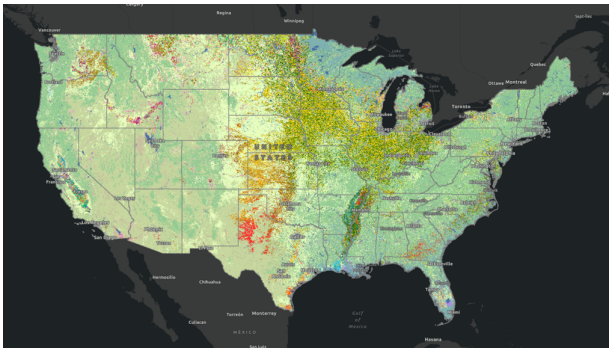


Figure 2.3: CDL layer of the year 2020.

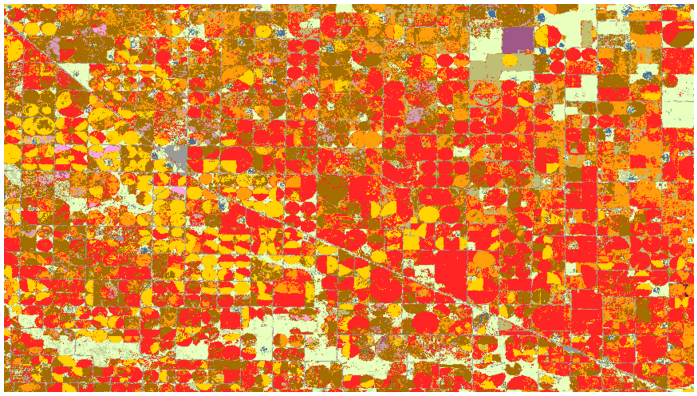


Figure 2.4: Corn and wheat fields on CDL 2020 Kansas layer.

3 State of the art

As mentioned our work is strongly based on the publication "Deep learning based winter wheat mapping using statistical data as ground references in Kansas and northern Texas, US" published by Liheng Zhong and collaborators (2019) [9]. In their study, they utilized a deep learning based approach by using in conjunction agricultural statistics (NASS) and satellite images with NDVI (MODIS). To predict winter wheat maps. They validate their predictions by comparing them to the CDL map.

The area of study was composed of the whole state of Kansas and some north counties of northern Texas. These areas were selected for two main reasons. First, they are intensively cultivated by winter wheat. Second, they are representative of two extremely different cases. In Kansas the CDL and NASS differ by 3.6% on average in the area covered by winter wheat. In contrast, the CDL of Texas overestimates the amount of wheat area by 55.2% on average. The figure X shows the areas just mentioned. The figure 3.1 by Zhong (2019) shows the area of study.

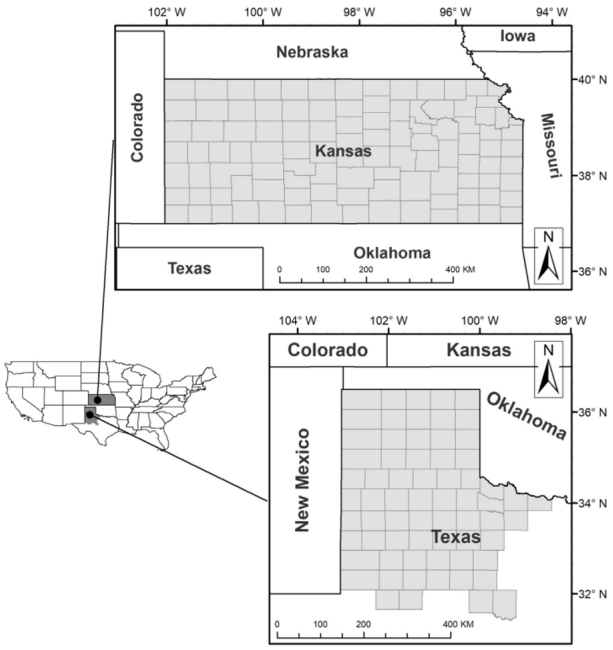


Fig. 1. One study area is the state of Kansas (105 counties), and the other is the 67 counties in four agricultural districts in northern Texas.

Figure 3.1: Study area of prev work.

The primary input data for their models is the MODIS product, which contains the vegetation index (NDVI) at a resolution of 250m. The MODIS product provides 23 images by year with an interval of 16 days.

The data is processed as follows (a more detailed explanation in chapter 5): MODIS images are clipped (cut) using the boundaries of counties in the study areas. For each county, each year a data cube is created with a certain width, height and depth of 23, depth corresponding to the temporal dimension. This data cube is the input for training the models. Each data cube of each county in a certain year is associated with the corresponding value of the winter wheat percentage provided by NASS statistics for that county in that year. As a pair of independent/dependent variables.

The data was splitted in training and validation (used for hyperparameter search) in a temporal split. For training 2001-2016 and for validation 2017. They did not use any test set.

The model input is a NDVI cube with the shape width by height by 23, and the output is a single number that tries to match the NASS wheat percentage.

The general architecture of the models is:

- Two convolutional 3d layers (Conv3D) with ReLU activation function
- Max-pooling layer along the temporal dimension
- Another two Conv3D layers with ReLU activation function
- Dropout regularization with probability of 20%
- One Conv3D layer with sigmoid activation
- Followed by a global average layer which reduces the output matrix of the Conv3d layers to a single number by taking the average.

The kernel of the models is what differentiates them. One has a kernel size of 3-by-3-by-3, while the other has a kernel size of 1-by-1-by-3. They are called “Space-temporal” and “Temporal-only” respectively, because the Space-temoral takes information from the neighbor pixels in both spatial and temporal dimensión , while the Temporal-only as the name implies only takes into account information of the same pixel in different times. The model’s architecture is represented in figure 3.2 [9].

Again, a more detailed explanation of the models is addressed in chapter 5. Because we kept practically the same architecture of the Temporal-only model.

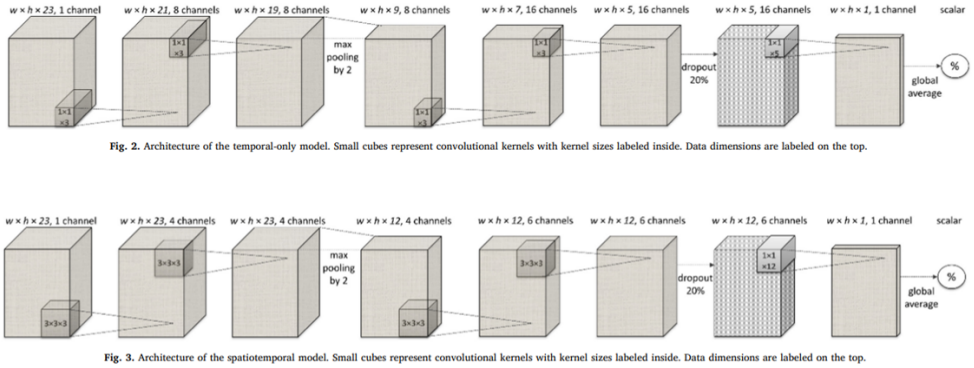


Figure 3.2: Models architecture for Space-temporal and Temporal-only models.

The output of the last Conv3D layer with sigmoid activation function can be interpreted as a wheat map, because the values of the pixels end up between 0 and 1 (thanks to sigmoid function's mapping) and can be interpreted as the percentage of wheat in that pixel.

The maps are obtained by taking the output of the last Conv3D layer. Winter wheat maps were produced for the years 2001-2017. They used their same train and validation data to make predictions without an independent test set. They compared their predictions to the CDL, and argued that the CDL was "an independent test set". Figure 3.3 [9] visualy compares their predictions to the CDL layer rescaled to MODIS resolution.

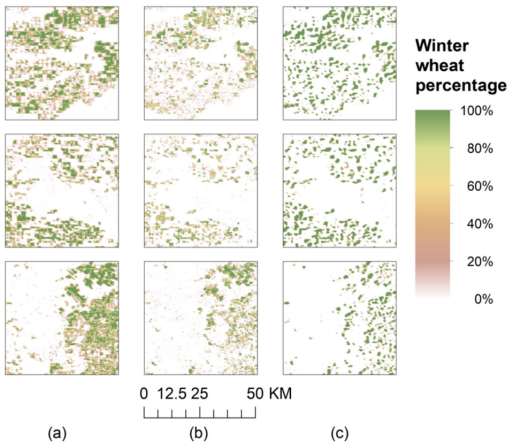


Figure 3.3: Comparison between the CDL and the resultant maps of Kansas in 2016 in three sub-areas. Column (a) includes the CDL winter wheat percentage. Columns (b) and (c) are winter wheat percentage maps by the temporal-only and the spatiotemporal models, respectively.

They made a wall-to-wall assessment of their predictions, using the CDL to measure accuracy and other metrics as F1 score. We have the intuition that they assigned labels

to pixels as "wheat" or "not wheat" by a threshold of 0.5. We assumed this because there are no details given in the paper and is also a common practice in the field of remote sensing and agriculture [2]. Their results are presented in the figures 3.4a and 3.4b [9].

Evaluation metrics of winter wheat maps in Kansas.

	Wall-to-wall assessment							
	Temporal-only model				Spatiotemporal model			
	Overall accuracy	Producer's accuracy	User's accuracy	F1 score	Overall accuracy	Producer's accuracy	User's accuracy	F1 score
2006	92.2%	65.6%	81.9%	0.729	87.8%	61.0%	61.9%	0.614
2007	91.7%	61.6%	81.9%	0.703	86.3%	55.7%	57.2%	0.564
2008	90.9%	54.2%	84.9%	0.661	87.0%	57.0%	61.1%	0.590
2009	91.3%	61.0%	79.8%	0.691	87.1%	58.8%	60.0%	0.594
2010	93.2%	63.9%	86.6%	0.735	88.6%	59.8%	62.0%	0.609
2011	92.2%	59.1%	84.7%	0.696	87.4%	54.2%	59.5%	0.567
2012	91.6%	65.4%	80.2%	0.720	86.0%	58.8%	57.5%	0.581
2013	90.8%	56.1%	83.4%	0.671	84.9%	50.5%	55.0%	0.527
2014	91.1%	54.0%	80.5%	0.646	87.4%	56.7%	58.6%	0.576
2015	90.9%	58.7%	79.9%	0.677	86.7%	55.8%	59.5%	0.576
2016	93.0%	67.6%	82.6%	0.743	87.4%	58.0%	58.0%	0.580
2017	93.9%	70.0%	80.6%	0.749	88.5%	59.3%	55.3%	0.572
Average	91.9%	61.4%	82.3%	0.702	87.1%	57.1%	58.8%	0.579

(a) Metrics of predictions in Kansas.

Evaluation metrics of winter wheat maps in Northern Texas.

	Wall-to-wall assessment							
	Temporal-only model				Spatiotemporal model			
	Overall accuracy	Producer's accuracy	User's accuracy	F1 score	Overall accuracy	Producer's accuracy	User's accuracy	F1 score
2008	91.7%	29.1%	88.3%	0.437	91.6%	35.8%	75.6%	0.486
2009	89.8%	23.4%	88.3%	0.370	89.7%	27.6%	75.8%	0.404
2010	93.4%	46.9%	83.7%	0.601	92.6%	53.8%	70.2%	0.609
2011	92.7%	20.1%	80.0%	0.322	92.6%	28.1%	66.4%	0.395
2012	91.4%	26.4%	70.3%	0.384	91.3%	35.9%	62.7%	0.457
2013	90.7%	22.0%	83.4%	0.348	90.4%	25.9%	70.8%	0.379
2014	90.7%	17.4%	82.9%	0.287	90.6%	22.8%	68.7%	0.342
2015	91.9%	35.2%	77.7%	0.484	91.4%	40.9%	67.0%	0.508
2016	93.1%	39.4%	80.9%	0.530	92.3%	45.0%	67.1%	0.539
2017	91.8%	28.3%	72.6%	0.407	91.5%	31.4%	64.5%	0.423
Average	91.7%	28.8%	80.8%	0.417	91.4%	34.7%	68.9%	0.454

(b) Metrics of predictions in Texas.

They get better results with the "Temporal only model" and their maps have better accuracy in Texas as expected because the CDL is more reliable.

They also verify that their model is truly capturing the winter wheat pattern on seasonality, and is not relaying to make predictions on just temporal patterns across years. They did this by inspecting the activation patterns of the intermediate convolutional layers with different inputs.

Overall, they showed that it is possible to train deep neural networks that produce winter wheat maps without using pixel level reference data (which is expensive to generate and more scarce) and relying only on historical statistics. Additionally, their models successfully identify the seasonal patterns of winter wheat in the temporal dimension.

4 State of the art limitations and proposal

4.1 State of the art limitations

The work made by Zhong, et al. is remarkable. Is a novel new approach that shows promising results for extending this technique, and apply it in regions without reference pixel data (as the CDL). Nevertheless, there is room for improvement points:

1. The most obvious and concerning one from the data science point of view, is that they didn't use proper training and testing sets to evaluate their predictions. They made predictions with the same training data and argued that the CDL was an independent test set.

2. The way they did the split between validation and training set was only a temporal one, without taking in consideration a spatial one.

3. Their data points excludes lots of nearby counties that are also rich in winter wheat, as in the states of Colorado, New Mexico.

4.2 Proposal

Our work looks to address this issues by:

1. Having a proper split of the data for training and testing. With 80% and 20% of the data points respectively.

2. Have a temporal and spatial split of the data for train and test sets. This allows us to verify that learning can be achieved through different times and different zones.

3. Extend the area of study to make use of nearby counties that are also rich in winter wheat.

Furthermore, our work seeks to:

1. Apply this technique to different geographical zones. The long term goal is to apply these algorithms on European land with transfer learning but before trying to move to a different continent, we first wanted to try it inside US territory where we can validate our predictions with the CDL. The equivalent of the CDL in other countries is not always available. For this, we worked in two different areas: the south region of the US is conformed by some counties of the states of Kansas, Texas, Colorado, New Mexico and Oklahoma. The North region includes some counties in Washington, Oregon, Idaho, Montana, North Dakota and Minnesota. As shown in figure 4.1, where the North region is represented in green and the south region in blue. We propose to train a neural network with the south data

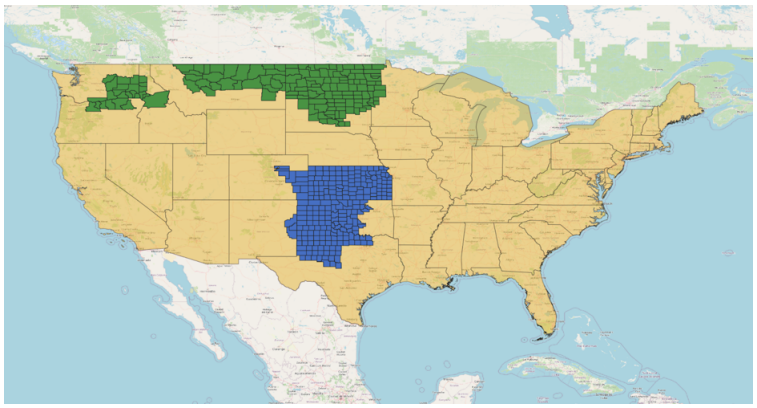


Figure 4.1: North (green) and south (blue) countys used for the project.

   and then transfer it to the north region, which is composed mostly of spring wheat and can be considered a geographical zone with different conditions.

2. Contemplate both winter and spring wheat into the data to generalize the models and address these two important crops.

3. Train a new model that targets CDL and NASS at the same time. Even though the paper argues that they want to rely only on historical statistics, at the end they end up comparing their predictions to the CDL (pixel reference data). The CDL and NASS statistics have big inconsistencies about the amount of wheat. NASS statistics has been widely used in agricultural applications and is considered a reliable reference [9]. Still, we see potential on the CDL as a target for training because of the spatial information it contains about exactly which pixels correspond to wheat. We address these issues by training a model that targets the CDL instead of a percentage, but a CDL that is "calibrated" to match the NASS numbers. Further detail of this in Chapter 5.

5 Methods

5.1 Data preprocessing

We have many sources of data, CDL, NASS statistics and MODIS satellite images. CDL and MODIS sources require preprocessing to feed the models. For this processing we used the software tool QGIS [7] and python with some specialized libraries such as GDAL [3].

Is worth mentioning that most of the pipeline to preprocess the MODIS data was developed by Lutz Ackermann during his internship (winter semester 2021-2022) in green spin. Thanks to his work we were able to do this project.

5.1.1 MODIS preprocessing

The overall preprocessing is the following:

- Download
- Merge
- Clipp
- Pad
- Stack

Download

As mentioned in chapter 2. MODIS data is a series of images that contain the vegetation index of a selected area. Our area of interest is composed of two different panels in the south and three in the north. As shown in fig. 5.1.

We download the data and proceed to treat it separately for north and south but the pipeline is the same. The data is open source and can be downloaded from NASA portal `https://search.earthdata.nasa.gov/search` [5].

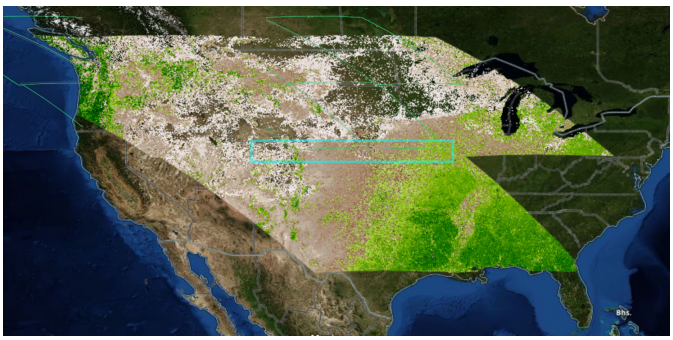


Figure 5.1: MODIS panels used to cover area of study.

Merge

Then, we need to combine the MODIS panels into a single one. We merge the corresponding panels that correspond to the same point in time. We end up having panels that look like this:

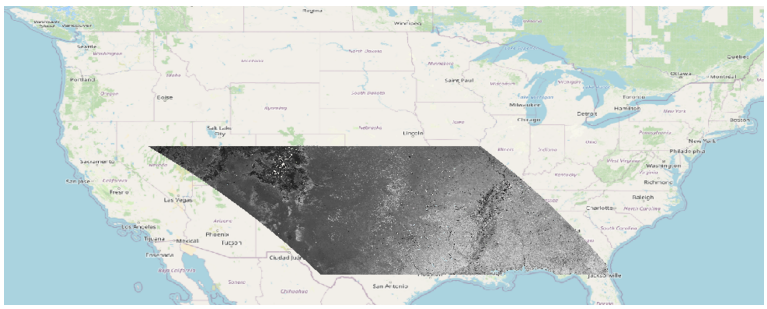


Figure 5.2: Merge Modis layer for south region.

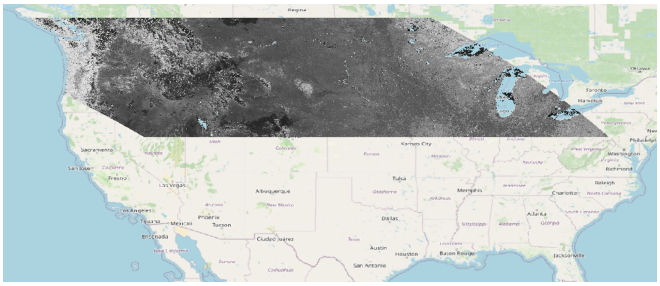
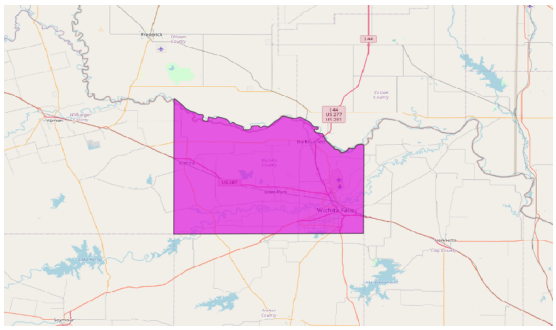


Figure 5.3: Merge Modis layer for north region.

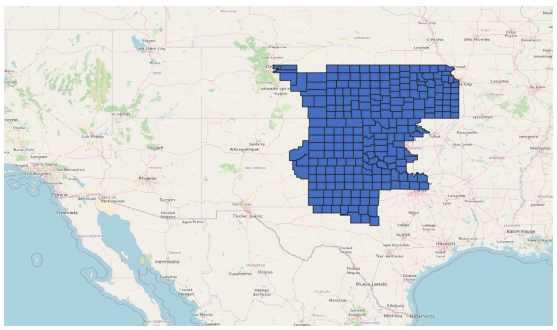
From now on we proceed by explaining only the south pipeline to avoid redundancy. The process is the same for both areas. The reason they were done separately was for computational reasons (data can become really heavy to process all at once) and to have the data organized for the different regions.

Clipp

After merging, the MODIS layer gets clipped with the shape of the counties of interest. We do this by using shape files. A shape file is an object that contains the geographical information of the borders and coordinates of a specific area. Shape files for the US counties can be downloaded `https://gadm.org/data.html` [4]. Figure 5.4a shows a shape file as an example for the county of Wichita, Texas. Figure 5.4b shows the counties of interest for the south region.



(a) Shape file of a county in Texas.



(b) Shapefiles used for south data.

These counties were selected because every year lots of wheat is produced in them. As shown in fig. 5.5 which is an overlap of the CDL of 2010 with the shape files. The black pixels represent the ones labeled as wheat pixels by the CDL.

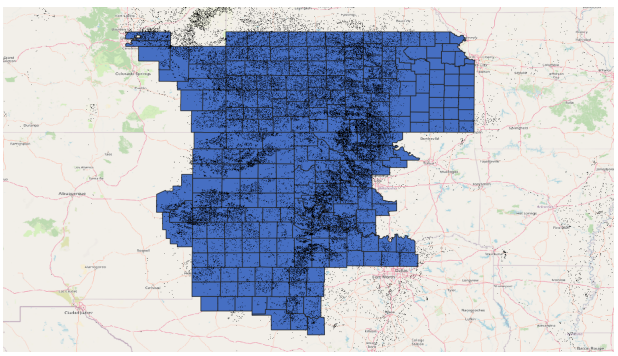


Figure 5.5: Overlap of the CDL wheat pixels (black) of 2010 with the shape files of south region.

After clipping the MODIS layer with the individual shapefiles we obtain layers that look like this:

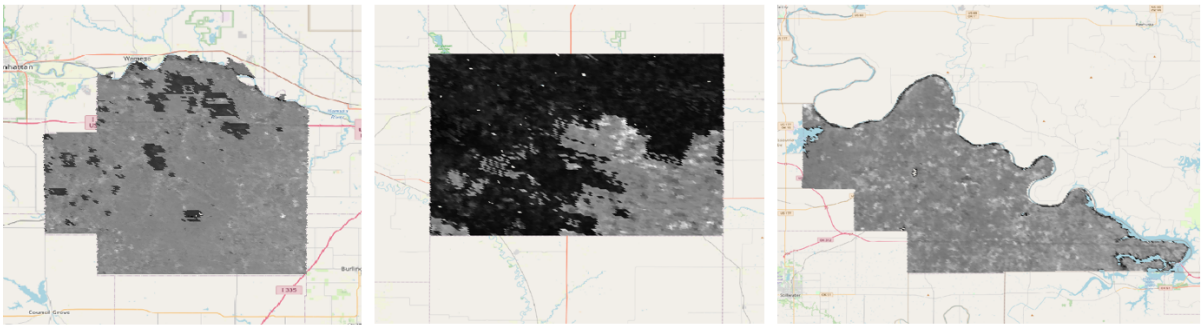


Figure 5.6: MODIS NDVI data clipped for some counties in the south region.

Pad

As figure 5.6 shows, the clipped images have many different shapes with different heights and widths. To standardize the input for the neural nets we looked for the max width and height for all the clipped files. In the south case this was max. height: 544 max. width: 993. We proceeded to pad the images so they all have the same dimensions. The padding assigns values of -1 to the filling values. The result can be visualized in figure 5.7.

Figure 5.7: MODIS data clipped and padded for some counties in the south region.

Stack

Finally, the last step is to stack the corresponding files into a data cube that becomes the input for our models. The cubes consist of the 23 padded images of the county during a year of harvest. A year of harvest x for this concern begins on the 29th of August of year $x-1$ and ends on the 28th of August of the year x. A visual representation of a data cube can be appreciated on figure 5.8.

The hole pipeline can be visualized in figure 5.9.

Figure 5.8: MODIS data cube example.

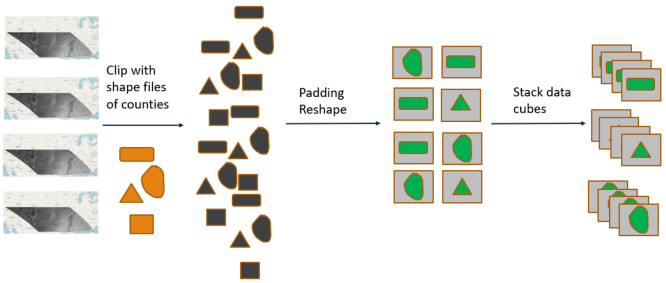


Figure 5.9: MODIS data cube example.

5.1.2 CDL rescaling

The CDL layer has a resolution of 30m and the MODIS data one of 250m. Our models take as input MODIS data and the wheat maps generated have the same resolution as MODIS. This means that we can not compare them directly, we need to rescale the CDL into the MODIS resolution. The whole rescaling process was made with QGIS.

The overall process goes as follows:

- Select pixels of interest
- Re-project
- Fill no data values
- Calculate proportion

Select pixels of interest

The CDL layer is composed of pixels with labels 255 different labels in total. For our purposes we are interested in all the labels that refer to winter wheat and spring wheat.

We selected the pixels with labels and ignore the rest:

- Label 23: Spring Wheat
- Label 24: Winter Wheat
- Label 236: Double Crop Winter Wheat/Sorghum
- Label 238: Double Crop Winter Wheat/Cotton

Figure 5.10 shows an area of the 2010 CDL layer before and after pixel selection.

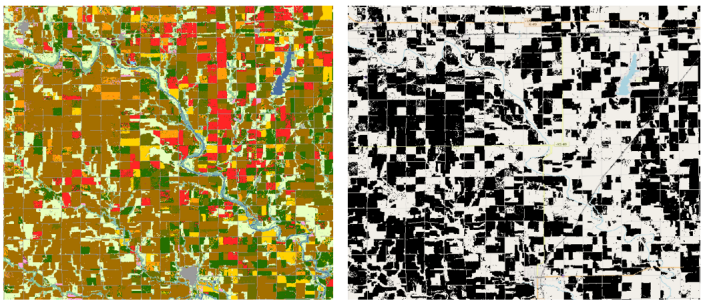


Figure 5.10: 2010 CDL layer before and after wheat pixel selection.

Re-project

The CDL and MODIS images have not only different resolutions but also a different projection over the earth surface.This means that the pixels do not align with each other. So, we send the CDL pixels to the MODIS projection. Figure 5.11 shows an example.

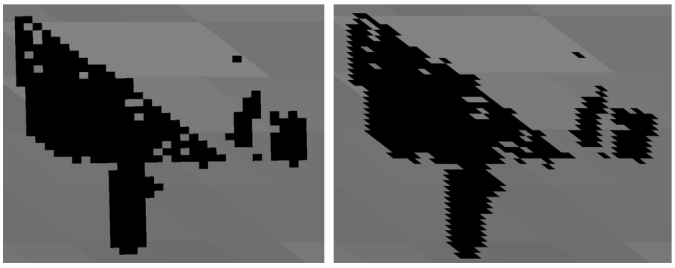


Figure 5.11: Overlap of CDL pixels with MODIS pixels before (left panel) and after (right panel) reprojection.

Fill no data values

Until now in our process layer we only have pixels with value 1. Indicating where is wheat but we also need values that indicate where there is no wheat, we assign to these

pixels the value 0.

Calculate proportion

Finally, we calculate the proportion of CDL pixels with wheat inside the bigger pixel of the CDL. This returns us an image in the MODIS resolution (250m) with pixel values between 0 and 1. With the pixels values indicating the percentage of wheat. Figure 5.12 shows the complete CDL rescaled for the year 2020. Figure 5.13 shows a selected area of the same layer where wheat fields can be appreciated.

Figure 5.12: CDL rescaled for the year 2020

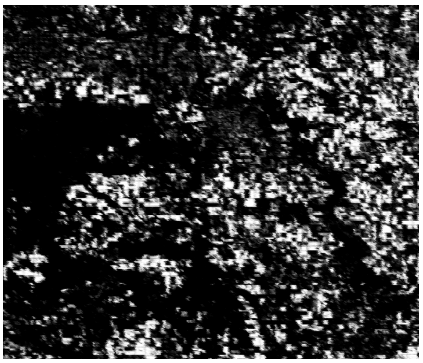


Figure 5.13: CDL rescaled for the year 2020

This process used to be done manually step by step in QGIS but we managed to develop an automated graphical processor. The processor takes as input a CDL layer and returns the CDL rescaled with percentages of wheat. Frigure 5.14 visualizes the processor diagram flow implemented in QGIS.

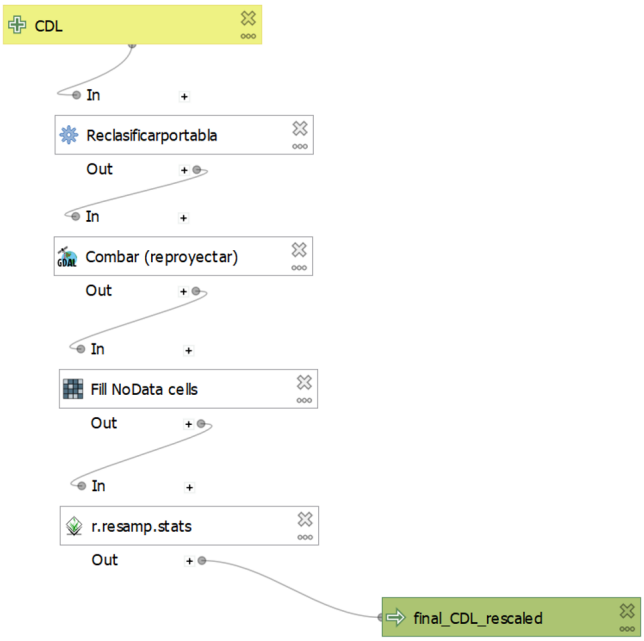


Figure 5.14: QGIS processor procedure for rescaling a CDL layer.

5.1.3 CDL calibrated target

As we previously mentioned. We see an opportunity in using the CDL as a target to generate wheat maps. The problem is that the CDL is less reliable than the NASS statistics. We trust the NASS numbers because of the way the data is collected in comparison to the CDL. The NASS numbers come directly from the farmers, after harvest season they are asked how much land they dedicated to which crops [2]. In comparison, the CDL is labeled by a decision tree-supervised classification trained on ground truth, with acurrcys ranging from 85% to 95% for the major crop categories [1]. Additionally, NASS statistics has been widely used in agricultural applications and is considered a reliable reference [9].

Our proposal is to combine the spatial information of the CDL pixels with the correct area percentage of the NASS statistics. We do this by clipping the CDL layers with the counties shapes and multiplying each one of them by a constant x to match the average of the pixels (which can be interpreted as wheat percentage) to the NASS percentage. We call this constant x "constant of calibration".

Mathematically we can say that the CDL average and the NASS percentage of wheat coverage are proportional eq. 5.1. To make them equal we add the constant x eq.5.2. Consequently, the constant x gets defined as the division of NASS percentage over CDL average eq.5.3.

$$\langle \text{CDL} \rangle \propto \text{NASS} \tag{5.1}$$

$$\langle \text{CDL} \rangle x = \text{NASS} \tag{5.2}$$

$$x = \frac{\text{NASS}}{\langle \text{CDL} \rangle} \tag{5.3}$$

Where $\langle . \rangle$ denote average.

For example, imagine we have a county A. The NASS statistics says that for county A in the year 2010 the area covered by wheat was 20%. Now, we rescale the CDL and get the clipped layer corresponding to the county A. We take the average of this CDL clipped and we find that the area covered by wheat is 40%. So, to make the numbers equal we have to multiply the CDL clipped by 0.2/0.4 or 0.5. After doing so the pixels of the CDL get "calibrated", in this case the pixels reduce their intensity by 50% to match the NASS percentages.

We can argue that each individual pixel of the CDL layer gets scaled by the constant of calibration, because of the propriety of matrices of multiplication by a scalar 5.4.

$$(kA)_{i,j} = k \cdot A_{i,j} \qquad \text{for all i,j} \tag{5.4}$$

We call this process of multiplying by the constant to match NASS numbers "calibration". After calibrating the CDL, our expectation was that we now have an image that contains the same information of the NASS statistics plus spatial information about the distribution of the wheat pixels. Therefore, we hope that by training the neural networks with this new CDL calibrated target would give us better results at predicting wheat maps.

The calibration process has the following steps:

1. Clip the CDL data with the shapefiles of the counties
2. Calibrate the clip data to match the NASS statistics for each county
3. Reshape the data to match shape of MODIS cube

Clip the CDL data with the shapefiles of the counties

Similarly to the clipping of the MODIS data. We take our CDL rescaled layers and clip them with the shape files of our counties of interest and obtain layers with different widths and heights. Figure 5.15 shows some examples:

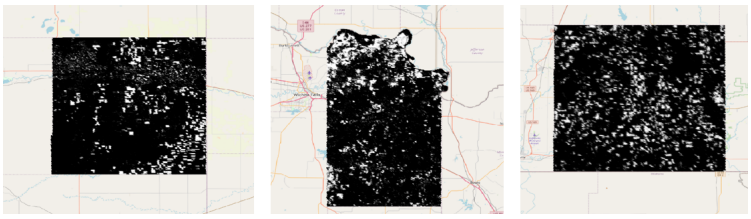


Figure 5.15: CDL recaled clipped with shape of counties.

We proceed to find the corresponding NASS percentage of wheat coverage for that county and corresponding year. With this NASS percentage we calculate the constant of calibration and multiply the CDL layer by this constant. We do this for all the counties and all years of data.

We registered all the constants of calibration to give us a look of the magnitude of this intervention. Figure 5.16 shows the histogram of these constants of calibration.

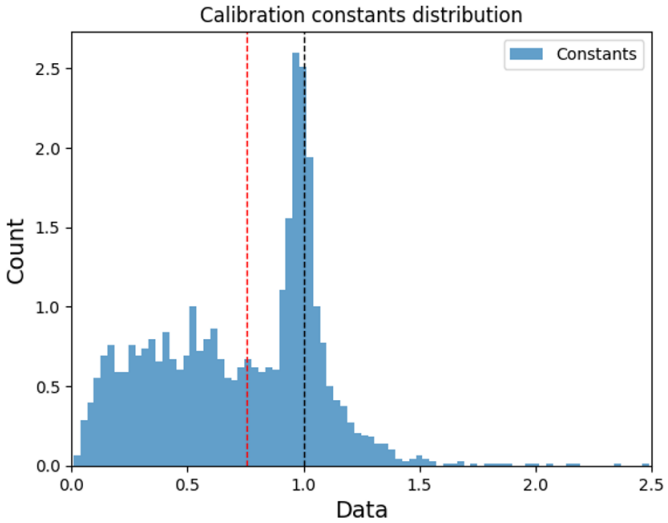


Figure 5.16: Histogram of constants of calibration in south data.

If the calibration constant was equal to 1, this would mean that CDL and NASS percentages are the same. So, ideally we would expect to find the data centered around 1.

$$\langle \text{CDL} \rangle x = \text{NASS} \tag{5.5}$$

$$\langle \text{CDL} \rangle 1 = \text{NASS} \tag{5.6}$$

$$\langle \text{CDL} \rangle = \text{NASS} \tag{5.7}$$

For lots of counties this is the case, the black doted line in figure (NASS and CDL numbers agree) 5.16 marks the value 1. But for many other counties we find that the constant of calibration takes values below 1. This means that the CDL is overestimating the amount of wheat and their numbers get reduced by the constant of calibration. This is what we expected due to the analysis of the previous paper. Additionally figure 5.17 shows the distribution of wheat percentage for all counties in all years according to NASS and CDL.

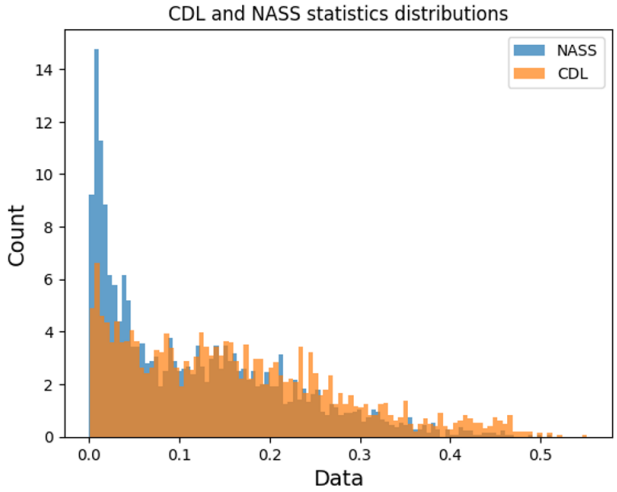


Figure 5.17: Histogram of both NASS and CDL coverage of wheat for all south data.

After calibrating the two distributions match perfectly (NASS and CDL calibrated). As shown in figure 5.18. A small shift to the CDL histogram was added to visualize both distributions.

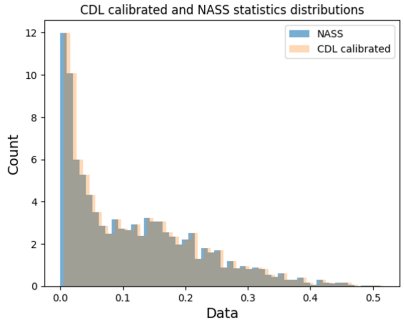


Figure 5.18: Histogram of both NASS and CDL coverage of wheat for all south data.

Now that the data is calibrated we proceed to pad the layers to standardize the shape of the targets. Similarly to MODIS preprocessing we fill the padding with -1 values. Fig 5.19 shows some examples of CDL calibrated layers after padding.

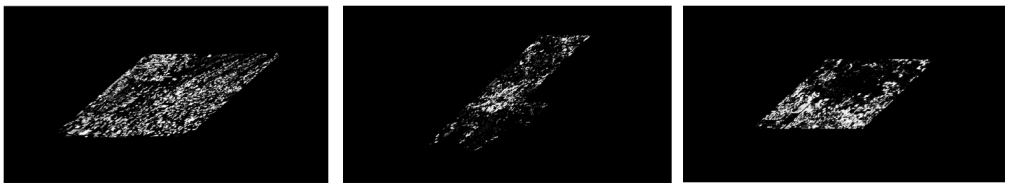


Figure 5.19: CDL layers rescaled, clipped and padded.

5.2 Available data and train-test split

Now that we have all the data ready to feed the algorithms we have to decide which data is going to be used for training and which for testing.

The CDL layer for Kansas and Texas is available since 2008 and the most recent available layer by the time of this project is 2021. So, our study period is composed of the years 2008-2021, 14 years in total.

The south region is composed of 241 counties. This means that the amount of data points is equal to $241 * 14 = 3374$. This is not the case due to the fact that NASS statistics are not always available to counties where the production of wheat was really low. In our case we ended up having a total of 2633 data points in the south region.

We dedicate 80% of the data for training and 20% for testing. Which is roughly 2106 points for training and 526.6 for testing. We propose two different ways of doing the train- test split, temporal split and spatial split. From now on we will often refer to the train set as group A and the test set as group B.

The temporal split consists of using all the counties of the years 2008-2018 for A and the rest of years for B. The figure 5.20 visualizes this split.

On the other hand, the spatial split consists of dedicating certain counties during all years to group A and the rest to group B. We randomly assigned the counties to A and B. Figure 5.21 visualizes this process.

The same approach was applied to the north data.

In this case the north counties are bigger that the south ones, in consequence we have less but bigger counties. 149 counties in total as shown in figure 5.22:

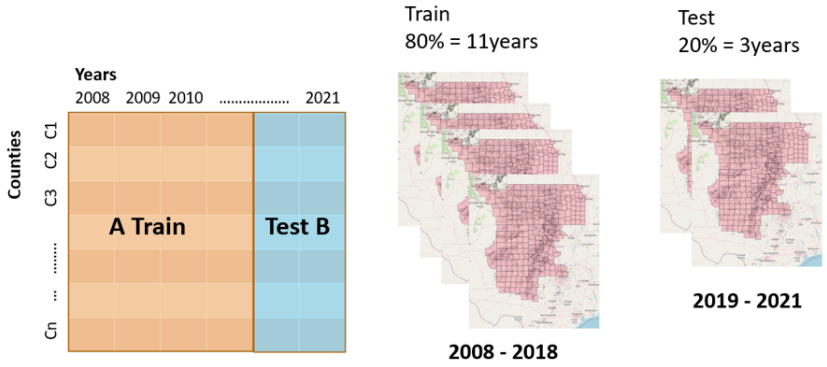


Figure 5.20: Visualization of temporal split. On the left there is a matrix with counties as rows and years as columns, colors indicate the split.

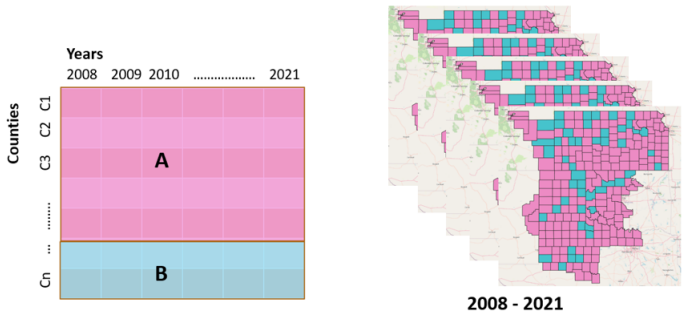


Figure 5.21: Visualization of spatial split. On the left there is a matrix with counties as rows and years as columns, colors indicate the split.

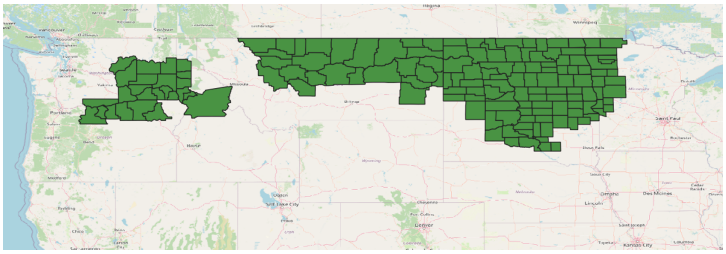


Figure 5.22: North counties used for this study.

After finding the available NASS percentages for these counties, we ended up having 1692 data points in the north. We splitted these counties also in groups A north and B north. In this case the point of having this split is to use B north to test the transfer learning directly and use A north to retrain the south models and then test on B north. We also used the proportions of 80% and 20% which left us with 1353 counties for A north and 338 for B north.

5.3 Models

The previous work showed overall good performance with their "Temporal-only" model, which outperformed their "Space temporal" model. So, we decided to keep the same temporal model design and almost all hyper-parameters.

We implemented two models in Pytorch [pytorch quote]. They have the same architecture and hyper-parameters. The only difference between the two is the target. The first model is the same as the one described in the paper. The input (MODIS data cube corresponding to a county) goes through the Conv3D layers, from the last layer a global average is computed and the result gets compared to the NASS target (the percentage of wheat corresponding to that county). We call this model "NASS model", the name coming from the target that the model uses.

The second model takes the same input and proccest it with the same layers but instead of taking the average of the last layer output, it directly compares it to the corresponding layer of the CDL calibrated. Both models are lustrated in the figure 5.23.

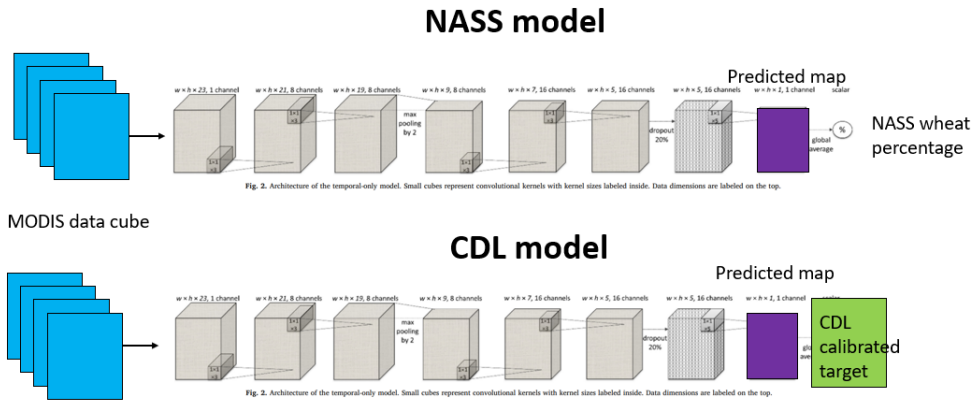


Figure 5.23: NASS model and CDL model architectures.

As mentioned both models only differ in the target, so let's address first the layers.

The general architecture of the models is:

- Convolutional 3d layer (Conv3D) with ReLU activation function, 8 filters kernel with size (3,1,1)
- Conv3D with ReLU, 8 filters with kernel size (3,1,1)
- Max-pooling layer size (2,1,1) along the temporal dimension
- Conv3D with ReLU, 8 filters with kernel size (3,1,1)
- Conv3D with ReLU, 16 filters with kernel size (3,1,1)
- Conv3D with ReLU, 16 filters with kernel size (3,1,1)

- Dropout regularization with probability of 20%
- One Conv3D layer with sigmoid activation function,1 filter with kernel size (5,1,1)

The Conv3D layers only take information of neighboring pixels in the time dimension. Therefore, the network learns to recognize the evolution on the NDVI curve of wheat during time, as the results of Zhong, et al (2019) [9] show.

The Max pool layer is used for reducing the sensitivity of small shifts in crop phenology [9]. The Dropout layer is used to force the network to not relay in only some neurons. Finally, the last Conv3D layer maps the incoming values to a 2D matrix with entry values on the range [0,1]. This generates our winter maps, due to the fact that the values of the pixels can be interpreted as wheat percentages.

As we didn't really play or fine tune for new hyperparameters, we didn't use a validation set in our training data.

The other hyperparameters used for both models were:

- ADAM optimizer with lr=0.001, betas=(0.9, 0.999)
- Number of epochs = 20
- Scheduler OneCycleLR-cos applied in the first 8 epochs with a maximum learning rate of 0.01
- Batch size = 3

We used the OneCycleLR scheduler because it showed to speedup the convergence of the models during training. Using a non constant learning rate has shown in many applications to help in this matter [8]. We applied the scheduler only to the first 8 epochs and then we kept training with the default learning rate of 0.001. Figure 5.24 displays the evolution of the learning rate during training.

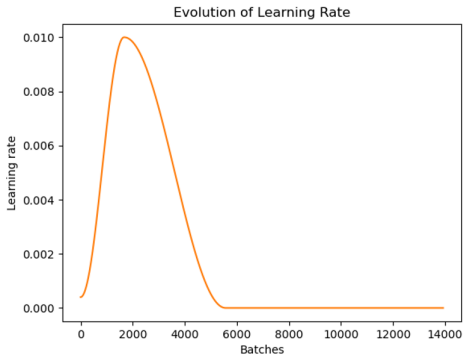


Figure 5.24: Lerning rate evolution over batches (20 epochs).

Going back to the map generated after the sigmoid layer. Now we have to compare this generated map to the NASS target in the case of the NASS model and to the CDL

calibrated target in the CDL model case.

5.3.1 NASS target comparison

Our intermediate layer after sigmoid still has the information from the input mask which can be visualized as blue pixels in the wheat map of figure 5.25. We only want to take into account the orange pixels which correspond to the MODIS NDVI pixels processed.

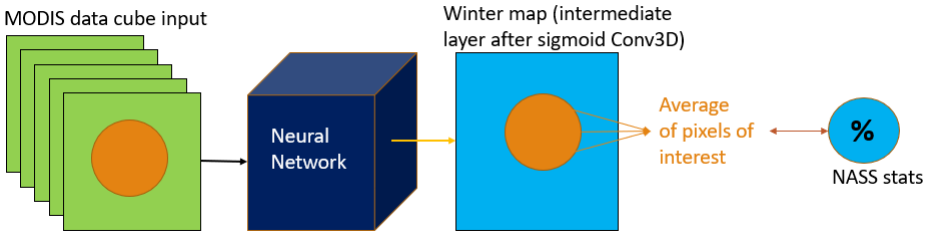


Figure 5.25: Visualization of NASS model, output and target.

To do this we generated a mask with the same shape of the input count and padding. This mask contains 0 values outside the shape of the county and 1 values inside. We multiplied this mask by the output, sending the padding (blue pixels) to 0 and preserving the same values inside the shape of the county (multiplication by one). Afterwards, we just took the sum of the resulting matrix and divided by the sum of the mask(number of pixels inside the shape of the county) to get the average of the pixels inside the shape of the county (pixels of interest), as shown in figure 5.26.

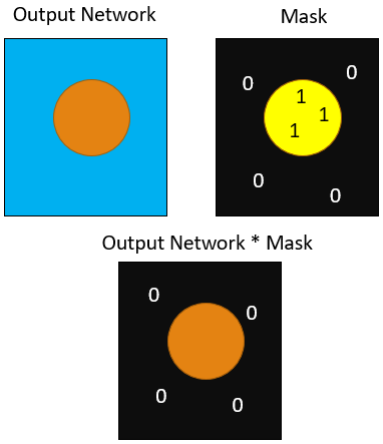


Figure 5.26: Visualization of NASS mask aplied to output.

Now that we finally have the average of the pixels of interest, we can compare this number to the percentage of wheat coverage that the NASS statistics has for the input

county in the respective year. We do this comparison with squared error (SE) and use this loss for training as shown in the equation 5.8.

$$SE = (x - \hat{x})^2 \tag{5.8}$$

Where $\hat{x}$ is the prediction and x is the target.

CDL target comparison

Dealing with the CDL comparison we face a similar problem as with the NASS target. We only want to compare the pixels that are inside the shape of the county. Comparing the mask makes no sense.

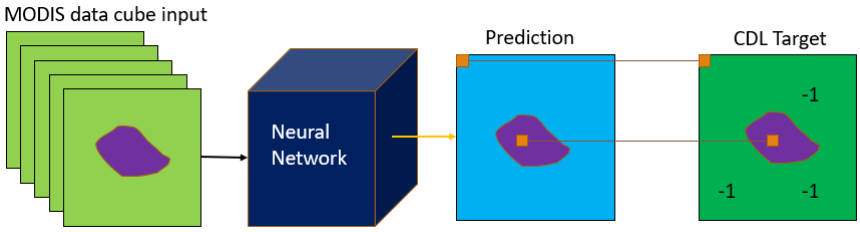


Figure 5.27: Visualization of CDL model, output and target.

Therefore, we applied a similar technique as with NASS. We created a mask for each MODIS input with the shape of the county and the same padding. Again this mask has 0 values outside the county and 1 values inside. We multiply this mask with the Prediction and with the Target as shown in figure 5.28.

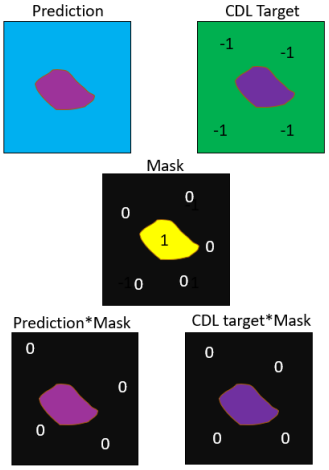


Figure 5.28: Visualization of CDL mask aplied to output.

We proceed to tho MSE between the two images. In this case the MSE compares pixel by pixel and takes the average. But at the moment of making the average, we divide by the sum of the mask which correspond to the number of pixels inside the county shape. With this we avoid the comparison between padding pixels which carry no information. Equation 5.9 display this loss.

$$MSE = \frac{1}{\mathrm{m}} \sum_{i}^{m} (x_m - \hat{x}_m)^2 \tag{5.9}$$

Where x is the true pixel, $\hat{x}$ the predicted pixel and m is the sum of the mask pixels which denote the pixels of interest.

6 Results

We have 4 models to analyze:

1. NASS temporal: NASS model trained with the train set of south temporal split. From now on represented in all graphs with the color blue.

2. NASS spatial: NASS model trained with the train set of south spatial split. From now on represented with the color green .

3. CDL temporal: CDL model trained with the train set of south temporal split. From now on represented with the color orange.

4. CDL spatial: CDL model trained with the train set of south spatial split. From now on represented with the color red.

Let's start with the training loss over epochs and then go on with the target comparison metrics.

In figure 6.1 we can look at the loss of the models over epochs. We can appreciate that the CDL spatial and temporal models converge faster than the NASS models. Additionally, CDL models have a lower loss. They can be compared because both losses are in the same domain and refer to similar concepts. The MSE from NASS models is just the comparation between two percentages, and the MSE from CDL models is the average of pixel comparation which also represent percentages.

1. We can compare against the NASS percentage.

2. Compare against original CDL rescaled (without calibration) as in paper.

3. Compare against CDL rescaled calibrated.

First, we will focus on the results obtained in the south region for train and test performances, we will later on address the results of the transfer learning in the north region.

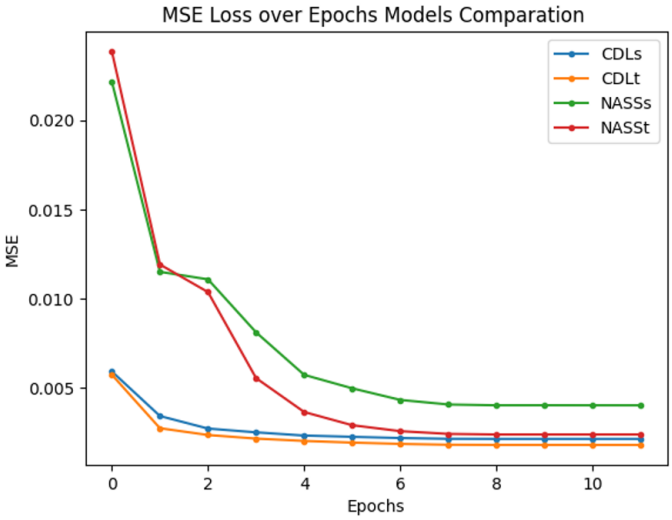


Figure 6.1: Trainig losses over epochs during trainig.

6.1 Comparison against the NASS wheat percentage - south

All models generate wheat maps. We can take the average of the generated map pixels and compare it to the corresponding NASS wheat percentage of coverage.

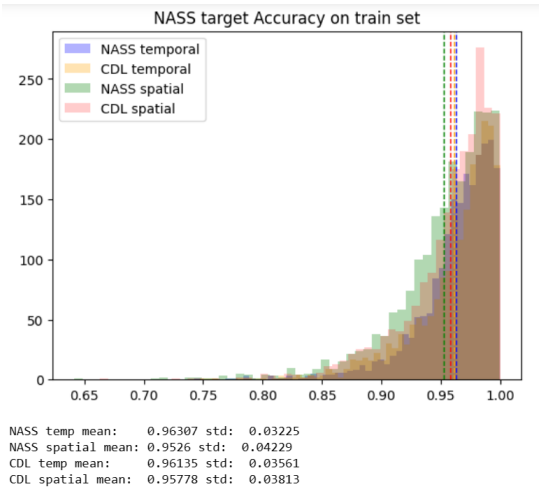
We measured the accuracy which we define in equation 6.1. In our case, do to the fact that we are dealing with numbers between 0 and 1, doing the MSE and then taking the square root returns us the percentage points of difference between the prediction and the truth percentage. Then we subtract this quantity to 1 and we obtain the accuracy of the prediction.

$$Accuracy = 1 - \sqrt[2]{(x - \hat{x})^2} \tag{6.1}$$

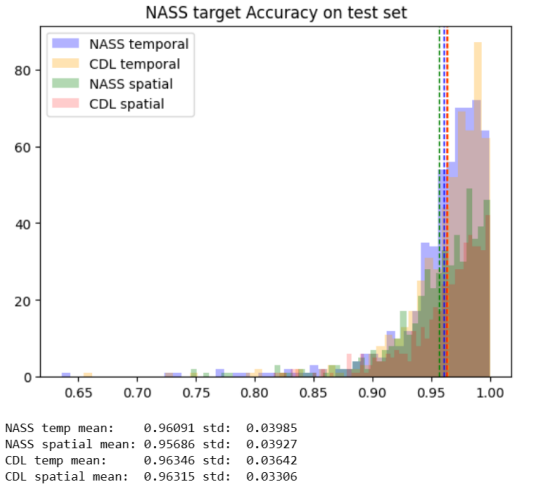
Where $\hat{x}$ is the prediction and x is the target.

We did this on train and test south sets for all models and obtain the following histograms with the corresponding means and standard deviations on the bottom, shoed in figure 6.2a for train and 6.2b for test performances.

We take a look at both train and test performances to assess that the network truly learned, and it did not overfit to the training data. From previous figures (6.2a,6.2b). We can see that the Accuracies for all models are really close and their distributions overlap. Overall, all 4 models are good at predicting the area covered by wheat, with

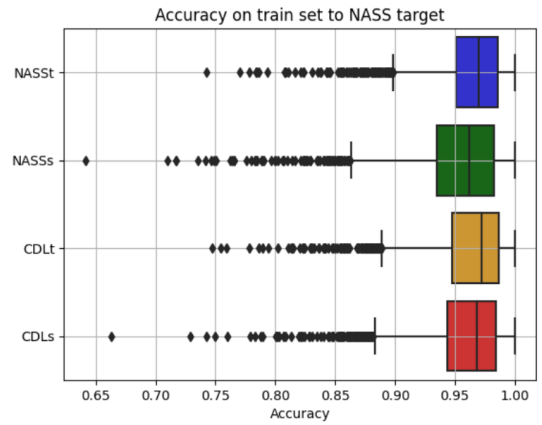


(a) Histogram of accuracies to NASS target on train set with means and standard deviations.

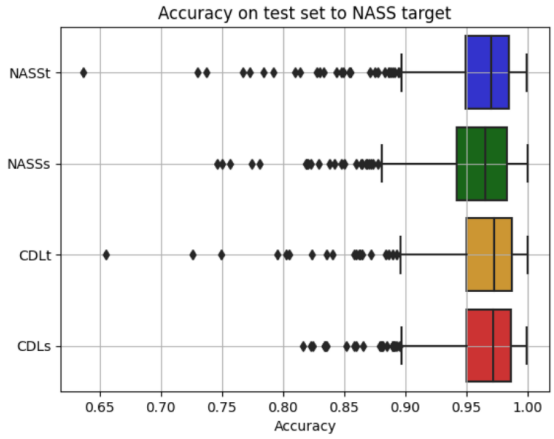


(b) Histogram of accuracies to NASS target on train set with means and standard deviations.

mean accuracy around 95%. Figures 6.3b, 6.3b show the same accuracies but in boxplot format



(a) Boxplot of accuracies to NASS target on train set.



(b) Boxplot of accuracies to NASS target on train set.

Here (6.3a,6.3b) we can aprecite that 75% porcent of mesurments are above 95% of accuracy on test set. And we can even apreccite a better performance on test set than train set. Overall, all models perform equally well.

6.2 Comparison against CDL target

Now we will compare the generated maps to the CDL rescaled without calibration. Even though we trained the network to predict the CDL calibrated, we wanted to compare our models predictions to the CDL without calibration, because the paper by Zhong et. al[9] did the same thing and we wanted to compare our results with theirs.

First let's take a look at one of the generated maps 6.4. As our maps still have the padding surrounding the wheat pixels, we applied the same technique of multiplying by a mask of 0-1 values to analyze only the pixels inside the county shape. As we did for computing the loss during training.

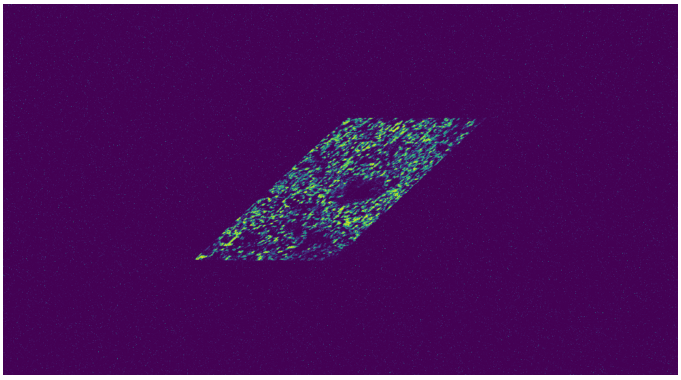


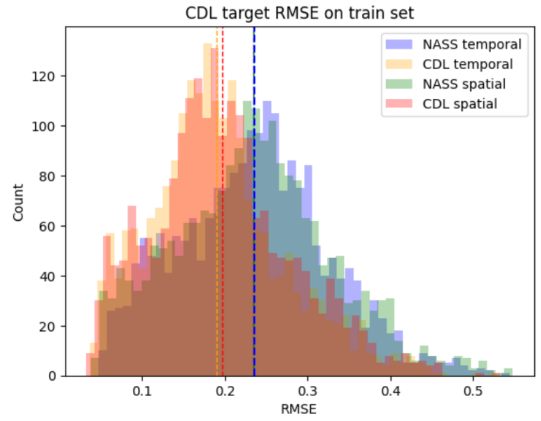
Figure 6.4: Wheat map generated by CDL spatial network.

6.2.1 RMSE comparison

The first metric we applied is RMSE. RMSE let us interpret the results as how big is the percentage difference on average pixel by pixel between the two images. Due to the fact that we discussed earlier in the accuracy section with NASS target. Doing the MSE and then taking the square root, returns us the percentage points of difference between the prediction and the truth percentage. Now this concept applies to all pixels and then we take the average.

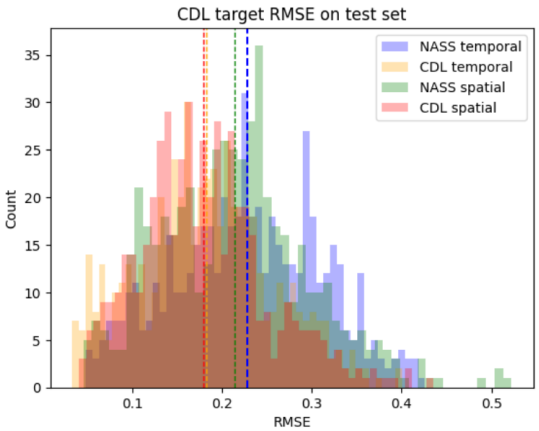
Figures 6.5a and 6.5b show the distribution of RMSE for train and test sets respectively. Here we can see a difference in performance between CDL and NASS models. CDL models have a lower mean MSE which means a better approximation to the CDL rescaled. This makes sense since the network was trained to match the CDL calibrated target.

In figures 6.5c, 6.5d we can also appreciate a lower variation for CDL models.



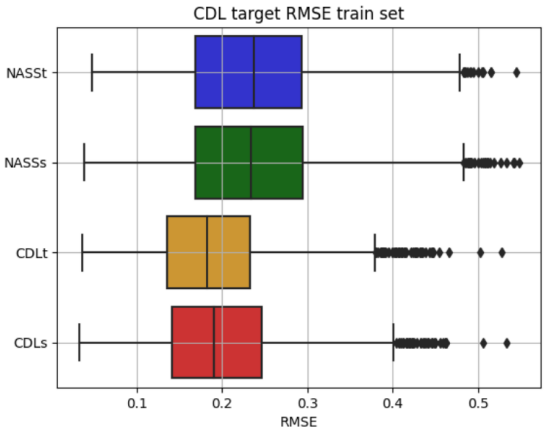

```
NASS temp mean:    0.2355  std:  0.09099
NASS spatial mean: 0.23473 std:  0.09623
CDL temp mean:     0.19016 std:  0.08159
CDL spatial mean:  0.19664 std:  0.0847
```

(a) RMSE to CDL on train set with means and standard deviations.

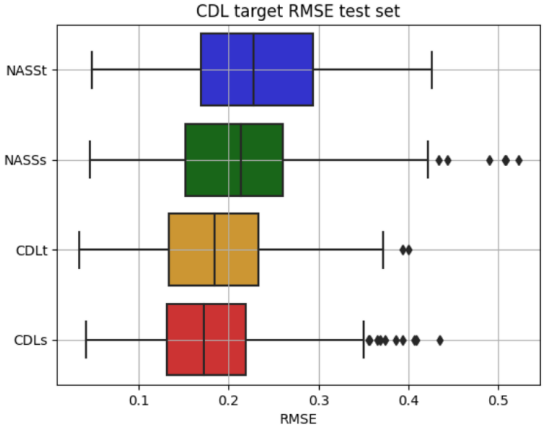



```
NASS temp mean:    0.22798 std:  0.08266
NASS spatial mean: 0.21382 std:  0.0838
CDL temp mean:     0.18203 std:  0.07495
CDL spatial mean:  0.17933 std:  0.0703
```

(b) RMSE to CDL on test set with means and standard deviations



(c) Boxplot of RMSE to CDL target on train set.



(d) Boxplot of RMSE to CDL target on test set.

Again, we find that the results are very similar for training and testing performances, even a bit better in testing. Is weird to get better performances on testing. This could mean that the test set is not representative enough of the training data, and it's easier to predict that the training data.

Still, the distributions overlap significantly. Further analysis with a test of statistical significance is needed to confirm the statement that our CDL models are significantly better than the previous work.

6.2.2 F1 metrics

The RMSE gives us a general idea of how good the predicted wheat maps are. Still we pressure a more detailed analysis of how good the models are actually at predicting wheat pixels. For this we transformed the predictions and the target pixels from a continuous range of [0,1](which indicated wheat coverage in the pixel) to just 0 or 1. Indicating wheat or not wheat. We did this by applying a threshold of 0.5. If the value of the pixel is greater than 0.5 then it gets assigned the value 1, in the other case 0.

This transformation allow us to find the number of:

- True positives: number of pixels correctly assigned as positive
- True negatives: number of pixels correctly assigned as negative
- False positives: number of pixels incorrectly assigned as positive
- False negatives: number of pixels incorrectly assigned as negative

And with this information we can calculate metrics as accuracy, precision, recall and F1 score.

Accuracy

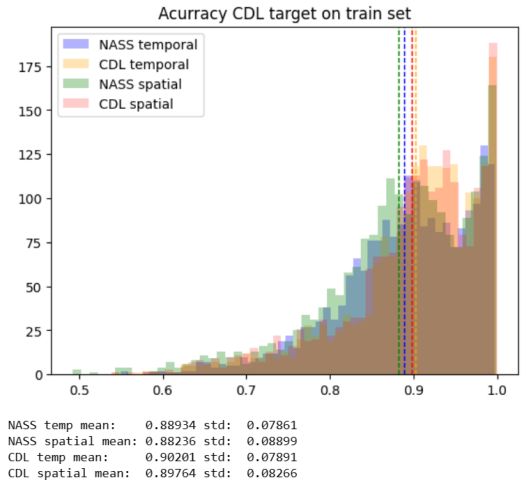
Accuracy is define as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{6.2}$$

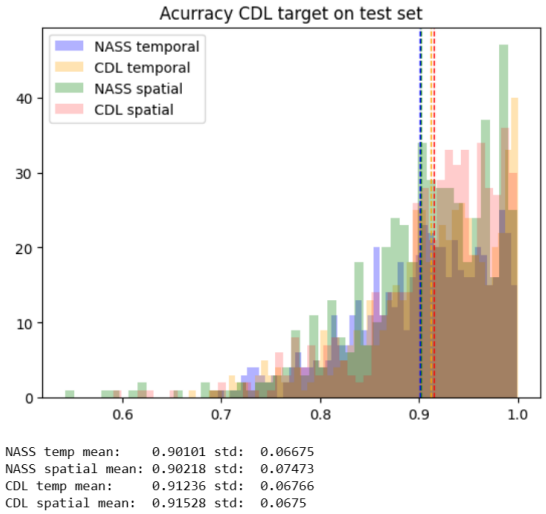
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.3}$$

This metric tells us the number of correct predictions over the total number of predictions. Figure 6.6a and 6.6b show the distribution of accuracies in train and test sets and figures 6.6c,6.6d the boxplots. We find the mean for all models close to 90%. Which means 90% percent of the pixels are correctly classified. Once again, we see a better performance with CDL models.

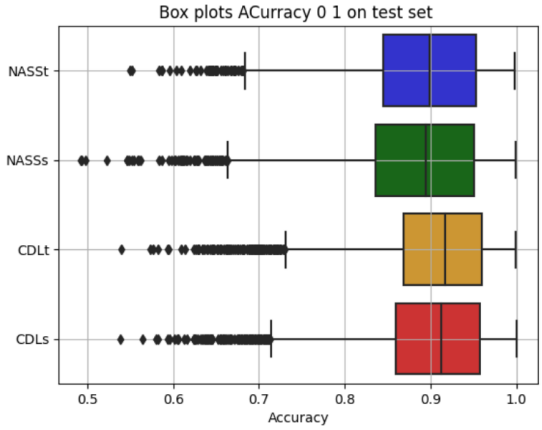
Accuracy can be misleading when the data is unbalanced. This is our case, we have much more pixels with the label 0 (no wheat) than 1. Therefore, the algorithm can have big accuracies by just assigning everything as 0. We address this by using other metrics as precision, recall and F1 score.



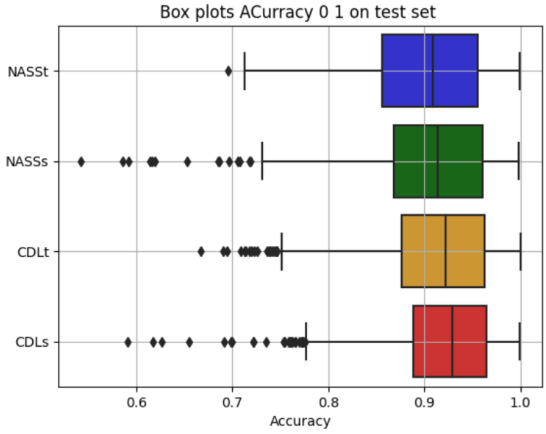
(a) Histogram of accuracies to CDL target on train set with means and standard deviations.



(b) Histogram of accuracies to CDL target on test set with means and standard deviations.



(c) Boxplot of accuracies to CDL target on train set.



(d) Boxplot of accuracies to CDL target on test set.

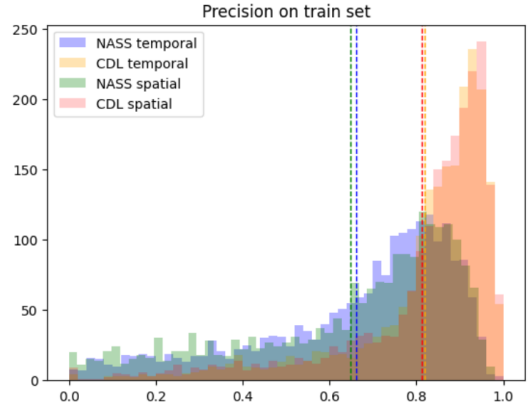
Precision

Precision is defined as:

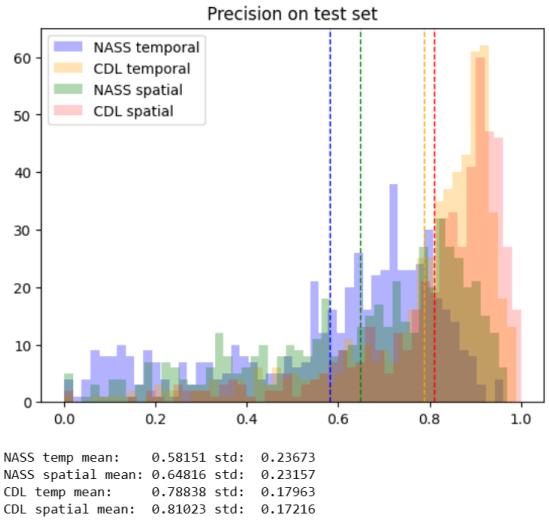
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6.4}$$

Precision can be interpreted as: within everything that has been predicted as a positive, what percentage is correct. A precise model might not find all the positives (might be a lot of false negatives), but the ones that it predicts as positive are correctly classified.

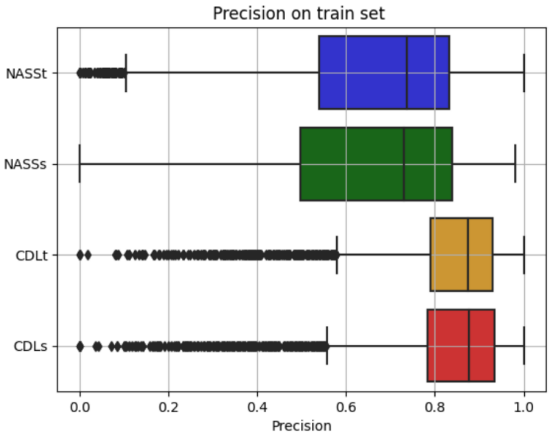
The most remarkable result from our proposal is performance over precision. Figures 6.7a,6.7b show the histograms and figures 6.7c,6.7d the boxplots for precision in train and test sets. Our models significantly outperform the NASS models (previous work). The difference between their means in precision is around 20%. And from figures **??**, 6.7d we can also appreciate a significant distribution towards 1 for CDL models.



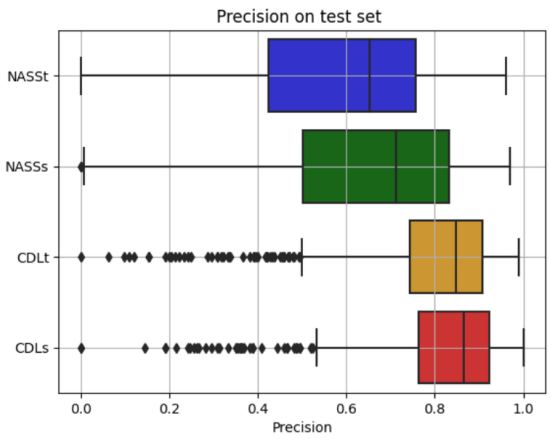
(a) Histogram of Precisions on train set with means and standard deviations.



(b) Histogram of Precisions on test set with means and standard deviations.



(c) Boxplot of precisions train set.



(d) Boxplot of precisions test set.

This means that our proposed models are better at correctly classifying positive values.

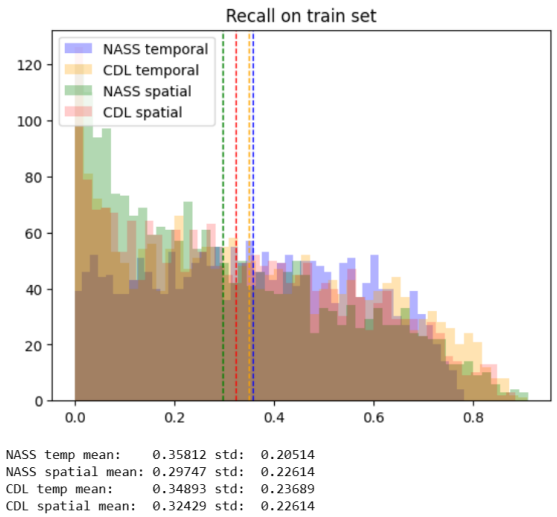
Recall

Recall is defined as:

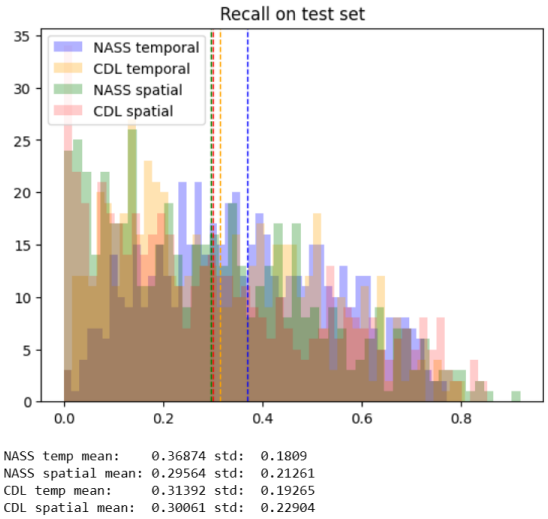
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6.5}$$

Can be interpreted as: from everything that is truly positive how much of it the model detects.

The results for recall can be observed in graphs 6.8a, 6.8b for histograms for train and test sets.



(a) Histogram of Recalls on train set with means and standard deviations.



(b) Histogram of Recalls on test set with means and standard deviations.

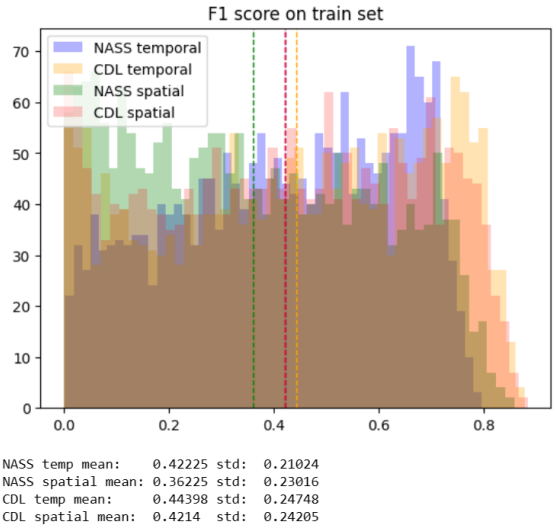
Unfortunately the recall metric is terrible for all models. This means that the models have a deficit of detecting positive pixels, we have a lot of pixels incorrectly classified as negative (false negatives).

F1 score

F1 score is defined as:

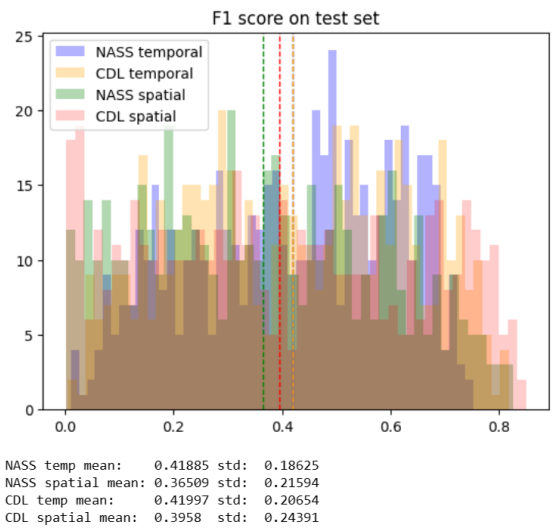
$$\text{F1 score} = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}} \tag{6.6}$$

F1 score is the harmonic mean between precision and recall. Is a metric that balances these two. A high score in F1 means that the model is good at finding all the positives and classifying them correctly as positives. Disappointingly, as the recall is low the F1 score also has low values as show in figures 6.9a and 6.9b.




```
NASS temp mean:    0.42225 std:  0.21024
NASS spatial mean: 0.36225 std:  0.23016
CDL temp mean:     0.44398 std:  0.24748
CDL spatial mean:  0.4214  std:  0.24205
```

(a) Histogram of F1 sccore on train set with means and standard deviations.




```
NASS temp mean:    0.41885 std:  0.18625
NASS spatial mean: 0.36509 std:  0.21594
CDL temp mean:     0.41997 std:  0.20654
CDL spatial mean:  0.3958  std:  0.24391
```

(b) Histogram of F1 sccore on test set with means and standard deviations.

This poor performance in recall can be explained due to the fact that the CDL is overestimating the amount of wheat. As we know from our analisis in chapter 5(5.16) and from previous work. The CDL layer says that there is more wheat than NASS numbers, and our models were trained with NASS target and CDL calibrated to match NASS numbers. Therefore, our models have big numbers in false positives that the CDL overestimates.

On a first look it doesn't make sense to train with NASS information and then compare with CDL information, but this was the approach from the previous work and we wanted to compare our results. We do this in chapter 7.

Now we proceed to compare the wheat maps to the CDL calibrated. where we would expect to get rid of this problem of overestimation of positives and false negatives.

Transfer learning results

We trained our four models with south data. This south data consists mostly of winter wheat. In contrast, the northern data is mostly spring wheat. With this experiment we seek to find if it is possible to transfer the knowledge of a network trained in one geographical area with one type of wheat, to another region with another kind of wheat.

For this we have different approaches in mind:

- Transfer directly without retraining the network with north data.

- Retrain/fine tune the hole network.
- Retrain just the head (last layer of the model).

We also have in mind training with the north data and comparing them to the transferred models. For all of these approaches we divided the northern data also into a training set and a test set.

Again, unfortunately, because of time constraints we left most of these experiments for future work. We only manage to do transfer without retaining. We transferred the 4 south models and applied them to the test sets of north data. The results are the following.

Comparing against NASS percentage of coverage gave us the results shown in figure 6.10. For south algorithms the accuracy was around 95%, now it is around 90% which on a first look seem to be promising results.

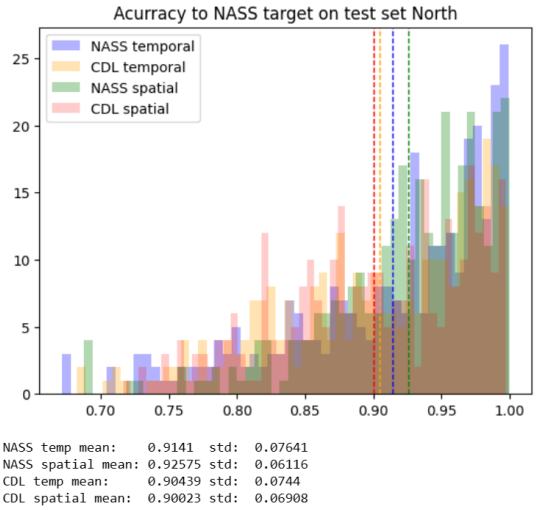


Figure 6.10: Accuracy agains NASS target on north test set with means and standard deviations.

Comparing against the CDL with RMSE we obtain the results shown in figure 6.11.

Again, we get a decrease in performance around 10% in comparison to test results in the south region. Nevertheless the algorithms seem to not do so bad.

These previous metrics (RMSE to CDL and NASS comparison) can be misleading. Due to imbalance in the data. The algorithms can be predicting just "no wheat" and still get good metrics. Now we proceed to again apply the threshold of 0.5 to the predictions and targets to apply the metrics of accuracy, precision, recall and F1 score.

For accuracy we obtain the following results (figure 6.12). This metric can also be

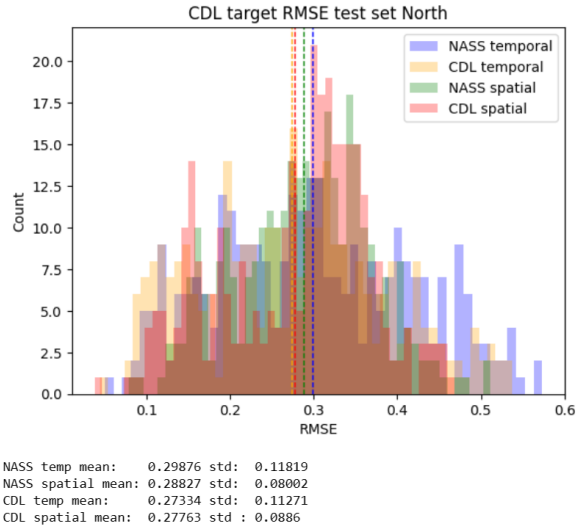


Figure 6.11: RMSE agains CDL target on north test set with means and standard deviations.

misleading by the reasons already mentioned.

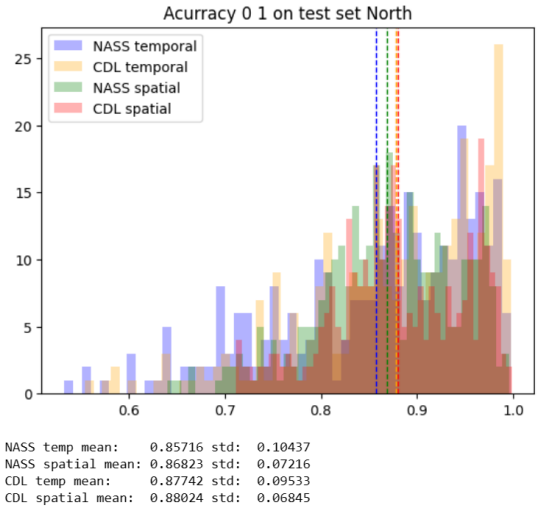


Figure 6.12: Accuracy agains CDL target on north test set with means and standard deviations.

For precision we obtain the results shown in figure 6.13.

For recall we obtain the results shown in figure 6.14.

For F1 score we obtain the results shown in figure 6.15.

The results in precision, recall and F1 score are terrible. Which means that the al-

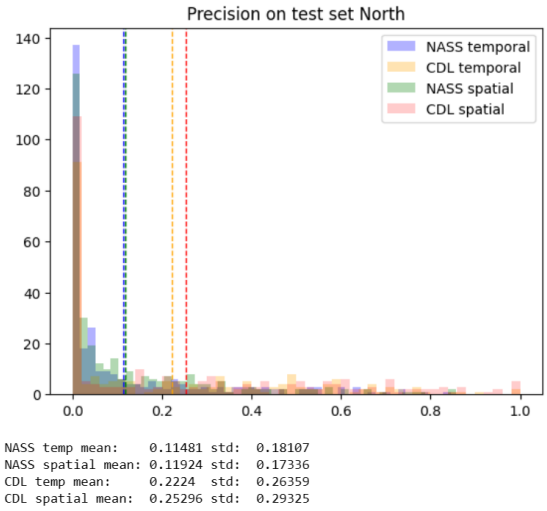


Figure 6.13: Precision agains CDL target on north test set with means and standard deviations.

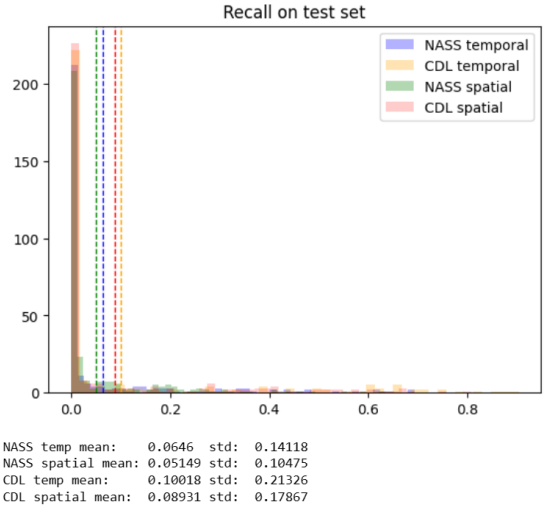


Figure 6.14: Accuracy agains CDL target on north test set with means and standard deviations.

gorithms are not able to detect True positive pixels of wheat. Therefore, the transfer learning did not work at this stage of just reapplying the models without retraining.

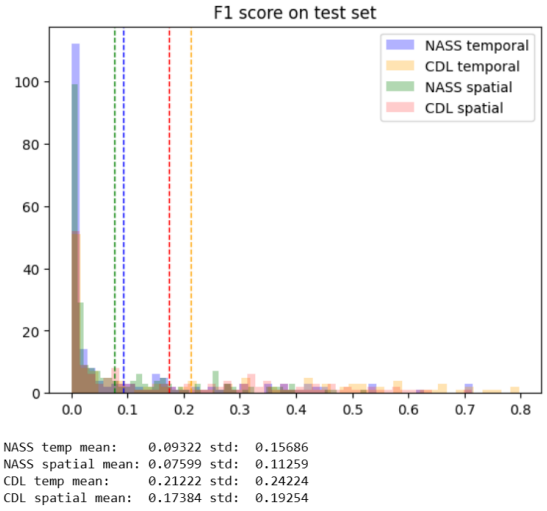


Figure 6.15: F1 score agains CDL target on north test set with means and standard deviations.

7 Discussion

From the results shown in the previous chapter we can say many things.

First, our proposed models have better or equal performances in all the metrics applied. Especially in the precision. This means that our models agree more with the CDL layer, but does this mean that our model is better at detecting wheat or is just better at matching the CDL? We have no way to answer this, because there is no ground truth. NASS and CDL data are not perfect. We can trust more the NASS data, but still we face uncertainties.

Another significant result which can be appreciated in almost all metrics and graphs is that the difference between temporal and spatial models performances is minimal. Which indicates that the network can learn in one area and then apply that learning into a new one (spatial models), and can learn in some timeline and then apply that learning to new years (temporal models).

Another interesting result is that we generated wheat maps by two approaches (NASS target, CDL calibrated target) and arrived at similar results. Figure 7.1 shows the prediction for the same input by the CDL temporal and NASS temporal models. Even though the NASS model was trained to match a single number, and the CDL model was trained to match an image, they make similar wheat maps. They both generally agree into a lot of regions. Further analisis between these intersections of agreement between the models can be really interesting, because if both models agree that there is wheat in the same pixel then we can state with more confidence that that pixel must have wheat.

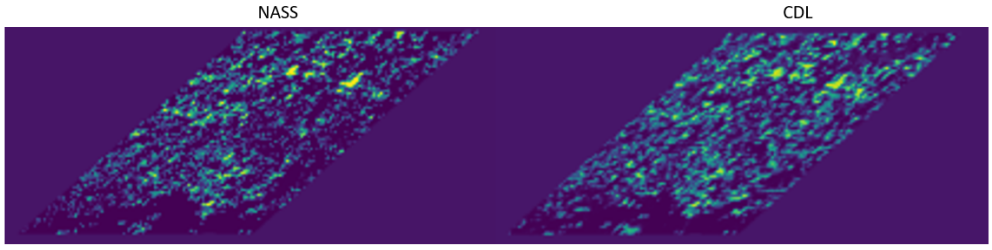


Figure 7.1: Wheat map predicted by NASS and CDL spatial models for the same input.

Regarding the previous work. Our results are not completely comparable to the ones in the paper because we used more data than them and we did not do the analysis by region as they did for Kansas in northern Texas. Figure 7.2 shows the counties used in the paper overlaped with our counties of study. Still we find in general similar trends in

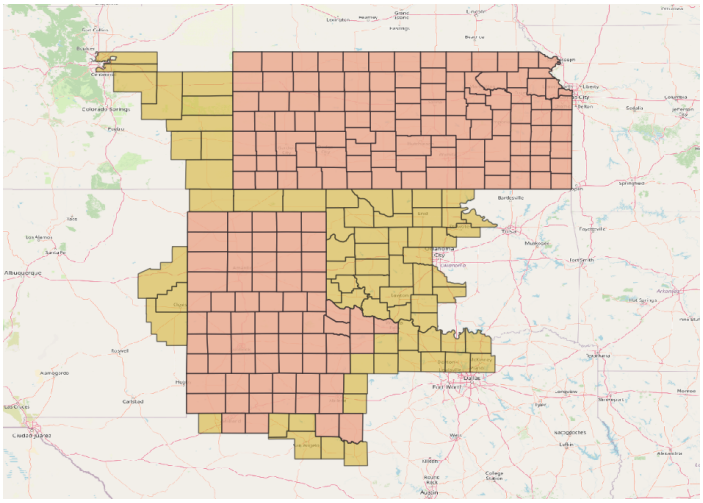


Figure 7.2: Counties shapefiles of previous work (pink) and our work (yellow).

our and their results of f1 score, recall and precision (see figures: 3.4a, 3.4b, 6.6b, 6.9b, 6.8b, 6.7b). So, we were able to partially replicate and reimplement their results (Their code was not open source, so we rebuilt it from scratch).

Regarding the transfer learning results. We can explain the terrible results due to the fact that our southern models were trained to find the season pattern of winter wheat. This might seem obvious at this point and that this transfer learning was not going to work because spring wheat and winter wheat have different seasonal patterns, and the network learns these patterns. But at the beginning of the project, when we set goals this was not obvious. As the philosopher Soren Kierkegaard once said "Life must be lived forwards, but can only be understood backwards.", that phrase truly applies to the story of this project.

Still, we gain a lot of insight from our results. And now we have many ideas of how to improve and move forward with these problem of detecting wheat with remote sensing data.

7.1 Limitations of our work

We are aware that our research may have some limitations:

- We can not trust the CDL always as a ground truth to compare our results, and as training data. CDL is not perfect. There are some areas where the agreement between CDL and NASS is great, but there are also regions of big incongruence. Our study makes no distinction between these two cases as the previous work did. Because of this, we might be using low quality data in some cases.

- The calibration process is restricted to the CDL distribution of pixels. If the CDL says that there is less wheat than NASS, the calibration process can send some pixels to values higher than 1 (100% wheat coverage) and that makes no sense. Also, the calibration process can only work with the previous existing pixels from the CDL, so it is biased to the CDL pixel distribution.

- We have mixed statistics for winter and spring wheat. Green spin combined into a single table the harvested area of winter wheat and spring wheat. We have no way of distinguishing the two crops to make independent analysis of the crops.

- Statistical significance still needs to be added for all metrics. Even though our proposed models outperform the previous work in metrics, we still need to check the statistical significance of these improvements. Because in many cases the distributions overlap between NASS and CDL models.

- validate

- validate

7.2 Future work

As we mentioned during the whole script, there are many more things to do with this project. Some ideas are:

- Try the other approaches of transfer learning: retrain the whole network with north train data, retrain just the head.

- Add metrics against CDL calibrated.

- Create an independent north network train to detect spring wheat.

- Analyze the areas of congruence and incongruence between both models. This could help us claim with more certainty the existence of wheat.

- Analyze the cases of the biggest error of the models. Maybe the model truly detects

wheat, but the CDL does a terrible job of representing the harvest area of that county/year.

- Transfer the south models to France. As we mentioned, the long term goal is to apply this techniques in Europe. France is a good candidate to start because it has good quality data on statistics similar to NASS and has also a product similar to the CDL to validate the predictions.

- Create and add new metrics to the comparison of the wheat maps. RMSE has limitations as well as the application of the threshold of 0.5 for f1 metrics.

7.3 Contributions to green spin

As a side product of our work, some contributions have been done to the green spin company.

- Functional implementation for generating wheat maps as the previous work described.

- Code and documentation of the whole data pipeline.

- Improvement into their processing of data (rescaling), automation in QGIS.

- New functional approach to generate wheat maps.

8 Conclusion

Detecting wheat with satellite data is still an open problem.

In this work we proposed a new approach of generating wheat maps with agricultural statistics and ground reference data. Our approach took advantage of the available CDL data to create models that outperform the previous work in various metrics regarding the same geographical zone.

Additionally, our work takes a first approach of applying transfer learning to different geographical zones and different crops. The results of this transfer learning are terrible at this stage.

Our work gives us inside of what can be achieved and what not. Giving us guidance in how to keep tackling this open problem for future work.

List of Figures

Appendix

Bibliography

[1] Claire Boryan et al. "Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program". In: *Geocarto International* 26.5 (2011), pp. 341–358.

[2] Gunther and Denise Dejon. *Greenspin agricultural knowledge and advice.* Spoken words. recompiled during 2022-2023.

[3] GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation. 2022.

[4] Global Administrative Areas. *GADM database of Global Administrative Areas, version 2.0*. `http://www.gadm.org`. [Online; accessed 16/03/2023. 2012.

[5] *NASA Earthdata Search*. `https://search.earthdata.nasa.gov/search`. Accessed: 16/03/2023.

[6] Planet. *Wheat in the Pacific Northwest: Understanding the Impacts of Droughts and Heat Waves*. 2021. URL: `https://www.planet.com/pulse/wheat-in-the-pacific-northwest-understanding-the-impacts-of-droughts-and-heat-waves/` (visited on 03/15/2023).

[7] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation. 2009.

[8] Leslie N Smith and Nicholay Topin. "Super-convergence: Very fast training of neural networks using large learning rates". In: *Artificial intelligence and machine learning for multi-domain operations applications*. Vol. 11006. SPIE. 2019, pp. 369–386.

[9] Liheng Zhong et al. "Deep learning based winter wheat mapping using statistical data as ground references in Kansas and northern Texas, US". In: *Remote Sensing of Environment* 233 (2019), p. 111411.

Declaration on oath

I hereby certify that I have written my master thesis independently and have not yet submitted it for examination purposes elsewhere. All sources and aids used are listed, literal and meaningful quotations have been marked as such.

Ana Muñoz Gutiérrez, March 23, 2023.

Consent to plagiarism check

I hereby agree that my submitted work may be sent to PlagScan (www.plagscan.com) in digital form for the purpose of checking for plagiarism and that it may be temporarily (max. 5 years) stored in the database maintained by PlagScan as well as personal data which are part of this work may be stored there.

Consent is voluntary. Without this consent, the plagiarism check cannot be prevented by removing all personal data and protecting the copyright requirements. Consent to the storage and use of personal data may be revoked at any time by notifying the faculty.

Ana Muñoz Gutiérrez, March 23, 2023.