UTRECHT UNIVERSITY
Department of Information and Computing Science

Applied Data Science Master thesis

The blooming period of trees in the Gargano National Park, Italy

FIRST EXAMINER:
WIEBE NIJLAND

CANDIDATE:
BO-LAN SCHOUTEN

SECOND EXAMINER:
MATHIEU GRAVEY

July 14, 2023

Abstract

This master's thesis focuses on the blooming period of trees in the Gargano National Park in Italy. The research question addressed in this study is whether there has been a shift in the period of trees turning green and if this change can be detected using satellite imagery, specifically the Normalized Difference Vegetation Index (NDVI). The aim is to analyze NDVI scores derived from Landsat 5 and 7 satellite imagery in the period from 1985 to 2020. The data preparation involves masking non-forest areas, applying cloud, cloud shadow, and snow masks, and rescaling the surface reflectance bands as well as normalizing them. An exploratory analysis of the data reveals an increasing trend in NDVI values over time. To determine the presence of a statistical trend, the Seasonal Mann-Kendall test is applied. The results indicate significant relationships between NDVI and time, demonstrating the potential for detecting changes in vegetation dynamics. Ethical considerations are addressed, highlighting the accessibility of the data and its freely available nature. Future research directions are suggested, including the incorporation of additional climate variables, the utilization of Landsat 4 and 8 data, and the exploration of leaf senescence dynamics. Overall, this study contributes to the understanding of long-term vegetation changes and their association with climate variations in the Gargano National Park.

Table of content

1. Introduction

1.1 Motivation and Context

Every spring, temperatures start to rise again, and the number of hours of daylight increases. The plants begin to bloom, the insects start to appear, and it marks the start for birds to begin nesting. But what would happen if the trees started blooming later or earlier? Changes in this period impact for example bird migration or bird nesting (Papes et al., 2012).

The study of vegetation dynamics and the timing of phenological events, such as the onset of leaf greening, is of great significance in understanding the impacts of change in ecosystems. One widely used metric in this situation is the Normalized Difference Vegetation Index (NDVI), which provides valuable insights into vegetation health and productivity. Analyzing the NDVI scores over the period from 1985 until 2020 and comparing the trend, it can be concluded whether there has been a shift in the period of trees turning green. This can contribute to the understanding of long-term vegetation changes and their potential association with climate variations.

1.2 Research question

The research question is as follows: Is there a shift in the period of trees turning green, and can this change be detected through the utilization of satellite imagery? The aim is to determine whether there has been a (significant) change in the period trees started to gain leaves. To achieve this, NDVI scores deprived of satellite imagery over the last 36 years will be used. Understanding these changes can provide valuable insights into the relationship between ecosystems and climate variability. If this change does exist, then this could potentially impact the functioning of the ecosystem. During the research, there will only be looked at leaf greening due to cloud cover. Too much cloud cover results in inaccurate satellite images.

1.3 Literature review

This report examines two additional studies to explore various methods employed for analyzing NDVI values in relation to vegetation changes. It delves into the datasets utilized and the specific methodologies applied in these studies. The aim of doing this literature review is to gain insights into techniques that could be used.

The first study looked at the change in vegetation based on NDVI values. Two different data sets were used, the first one was a biweekly NDVI dataset which was provided by Global Inventory Modeling and Mapping Studies. The second dataset used was a 10-day global land surface sensor dataset from SPOT-4's vegetation. They divided the NDVI into three groups: spring NDVI (April and MAY), Summer NDVI (June, July, and August), and autumn (September and October).

For detecting the variation trends in NDVI, temperature, and precipitation, a least-squares linear regression model was applied. The dependent value was the NDVI, or a climate variable (temperature or precipitation) and the year was the independent variable. This was followed by a piecewise regression to determine whether there was a change in the NDVI time series. (Zhang et al., 2013)

The second study looked at the vegetation change trend and the cause in the Mongolian Autonomous Region, China (Li et al., 2011). They used Advanced Very High-Resolution Radiometer data to determine the change trend. For determining the overall trend in vegetation change they used the monthly maximum NDVI values in the growing season. They did this by fitting a linear regression using least squares estimation where the dependent value was the monthly NDVI and the independent variable time. After this, the results were then converted into change rates and expressed as percentages relative to the linear trend at the start of the time series. For determining the significance of the linear regression models F-statistics were applied. If the regression model achieved significance at the 0.05 level, the NDVI was interpreted as indicating a trend of increase or decrease.

In both studies, they used the NDVI for calculating changes in vegetation. NDVI stands for Normalized Difference Vegetation Index and ranges between -1 and 1. Low NDVI values indicate that the vegetation is experiencing moisture stress, whereas higher values indicate a greater abundance of green vegetation. It indicates the state of the vegetation. Due to this, it is well suited for studying changes in vegetation (Li et al., 2011) .

The NDVI is calculated by measuring the difference between near-infrared (NIR) (which the vegetation strongly reflects) and red light (RED) (which the vegetation absorbs/has a low reflectance) and the formula is as follow (Gessesse & Melesse, 2019):

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

Multiple studies have shown that NDVI can be used to analyze the concentration of leaves as well as detect changes in vegetation (Pastor-Guzman et al., 2015) (Zhang et al., 2013) (Li et al., 2011). And it is especially useful for combining it with Landsat, Sentinel and MODIS imagery (Pettorelli et al. 2005). A further description of these satellites is described in 2.2. Description of the data.

2. The data

2.1 The research area

For finding a suitable research area, specific criteria were formulated. These criteria include the requirement for an increase in the green ratio throughout the plant lifecycle, the exclusion of evergreen trees due to their persistent green foliage, and the recommendation from an expert to focus on a Mediterranean country. Additionally, the presence of a significant number of trees in a national park was considered desirable. Based on these considerations and a selection of recommended national parks from the expert, Gargano National Park[1] in Italy was identified as a suitable candidate.

It is part of the Foggia province in Puglia and is considered part of south Italy. It is approximately on the same latitude as Rome and is therefore, from a geographical point of view, in a position of passage between southern and central Italy. The national park has a surface of about 2,100 square kilometers (Licht & Wagensommer, 2020).

This park has approximately 33% of the area covered by forest which primarily consists of deciduous trees that shed their leaves in autumn and regrow them in spring. Notably, Umbra Forest, which is one of the largest broad-leaved forests in Europe, is also located within the Gargano National Park. Broad-leaved woodlands are characterized by non-needle-bearing trees, predominantly composed of deciduous species, thus exhibiting a change of leaves per season. (Gargano National Park, sd)

2.2. Description of the data

As mentioned earlier, there were more satellite imagery sources usable for calculating the NDVI score. Landsat, Sentinel, and MODIS satellites could be used to obtain the NDVI values.

Landsat satellites have consistently provided high-quality imagery since the launch of the first satellite in 1972, offering the longest continuous space-based record of Earth's land (Wulder et al., 2019). The dataset allows for comprehensive long-term analyses, a crucial advantage for NDVI analysis, facilitating consistent monitoring of the temporal bands. Moreover, Landsat's spatial resolution of 30 meters is well-suited for local and regional scale studies, offering a good balance between detail and coverage area (Wulder et al., 2016). The fine resolution ensures that data isn't diluted by averaging over large areas, thus improving the precision of NDVI computation. With a 16-day revisit time, Landsat facilitates consistent monitoring, making it a good fit for NDVI analyses.

However, Sentinel-2 was also an option. Sentinel-2, a part of the Copernicus Program by the European Space Agency (ESA), offers multispectral imagery with a spatial resolution of up to 10 meters (Drusch et al., 2012). Sentinel-2's higher spatial resolution can provide finer detail in NDVI computation, potentially outperforming Landsat in detecting subtle changes in small vegetation patches. additionally, Sentinel-2 has a five-day revisit time at the equator, which improves temporal resolution (Drusch et al., 2012). However, despite these advantages, the Sentinel-2 program started in 2015, and thus has a significantly shorter data history than Landsat, limiting its ability for long-term change analysis. Another disadvantage is that it does not include thermal imaging, which typically is required for cloud detection (Wulder et al., 2016).

[1]https://www.google.com/maps/place/Nationaal+park+Gargano/@41.8121497,15.6470042,10.3z/data=!4m10!1m2!2m1!1sgargano+national+park!3m6!1s0x1337748eb8432525:0x1a2293822e2b9ca4!8m2!3d41.7817081!4d15.8516421!15sChVnYXJnYW5vIG5hdGlvbmFsIHBhcmtaFyIVZ2FyZ2FubyBuYXRpb25hbCBwYXJrkgENbmF0aW9uYWxfcGFya-ABAA!16s%2Fm%2F0gy15bz?entry=ttu

The Moderate Resolution Imaging Spectroradiometer (MODIS) was the third option. It provides daily global coverage, offering a broad perspective that is ideal for observing large-scale patterns and trends in vegetation (Justice et al., 2002). Despite MODIS coarse spatial resolution (250-1000m), its frequent revisit time (1-2 days) makes it a viable tool for NDVI analysis, allowing researchers to identify and analyze large-scale phenological trends. However, it does have a major downside when using it for calculating NDVI scores. It cannot detect and monitor small vegetation patches or individual trees, so the area should only include areas covered in primarily trees. Also, MODIS started in 2000, which has also a significantly shorter data history than Landsat (Pettorelli et al., 2005).

Of these 3 options, Landsat was the one chosen. The main reason for this is the historical data available in Landsat. It enables the observation of NDVI values over an extended period of time. However, Landsat has multiple satellites which are accessible to the public. Due to the time periods and the option to merge multiple Landsat the decision was to use Landsat 5 and Landsat 7. This resulted in data ranging from April 1984 to June 2020. According to the documentation on Landsat 7, it could range till June 2023, however, it also warned that after 2020 it cannot preserve its science capabilities. For this reason, it was decided to only look at the data from 1985 to 2020. 1984 was not included because the earliest data point is around April.

The reason to only include Landsat 5 and 7, and not Landsat 8 was due to the different wavelengths and their spectral ranges of the red and near-infrared sensors of the satellites. For Landsat 5 and 7, these were the same. However, Landsat 8 had different spectral ranges, which resulted in the NDVI scores being differently scaled, as shown in Figure 1.

Additionally, the reflectance collection was used because by utilizing reflectance, the impact of scattered radiation in the atmosphere is reduced (Huang et al., 2021).

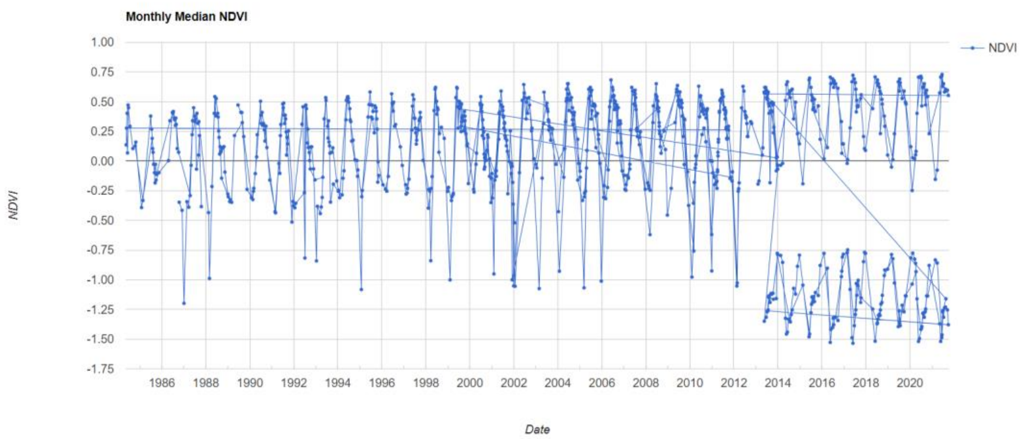

Figure 1: Reason why Landsat 8 was not included.

It was decided to ultimately apply a sampling technique instead of examining the entire area. The reason for this is that the entire area is 1,181 square kilometers, which ultimately amounts to more than 8,000,000,000 pixels. Since the intention was to calculate all NDVI values per pixel, per available date, and then aggregate them, this would result in an excessive amount of data. For this reason, it was decided to use a sample of the area. The chosen method was simple random sampling. The reason for this is that all pixels (containing forests) have an equal chance of being included in the sample, thereby minimizing bias (Sharma , 2017). Additionally, this method was easier to implement in Google Earth Engine, as the program randomly selects the pixels. Ultimately, a sample size of 3000 pixels was chosen. The main reason for this was limitations related to the available hardware.

2.3 Preparation of the data

Google Earth Engine was used to retrieve the data. As mentioned in 2.2. Description of the data it was decided to use a sample area instead of the whole study area.

When determining the sample area that could be used, a forest mask was employed to ensure that each sampled pixel indeed represents a forest or a tree. To create this mask, Hansen et al. Global Forest Change Data[2] dataset was utilized. This dataset provides information on forest changes and includes a band, the treecover2000, indicating the percentage of tree cover within a pixel. By utilizing this band, a mask was generated based on the value of the band, which is applied to the entire research area. This resulted in sampled pixels that are only within the forest mask, and thus containing some sort of tree cover. By using this map, it was ensured that only tree covers were randomly sampled and e.g., not cities. After establishing the sample area, the coordinates were retrieved and could be used for extracting the NDVI values for each pixel individual.

After retrieving the sample data, the Landsat 5 and 7 images were extracted. While extracting these images a cloud mask, shadow cloud mask and snow mask were used. For this, the QA_pixel band was used. This is part of Landsat 5 and 7 and has information about the quality of the pixel. These attributes were deprived of the CFMASK algorithm. These masks were combined into one mask and applied on the Landsat 5 and 7 Images.

There were several different techniques for handling cloud, cloud shadow, etc. such as CFMask, FMask and Automated Cloud Cover Assessment.

FMask algorithm can be used to detect clouds in Landsat imagery. By using spectral tests, it identifies cloud pixels. Besides clouds, it can also detect, cloud shadows, and snow for Landsat 4,7 and 8. It identifies per pixel with a high probability of being a cloud (Escobar-Silva et al., 2022).

Automated Cloud Cover Assessment is another cloud detection algorithm. It identifies cloud objects using a per-pixel decision tree and statistical aggregation of the decision tree outputs to reduce ambiguity. However, this algorithm does not contain thermal infrared bands (Foga et al., 2017).

Also, a possible option is the cloud cover function in Google Earth Engine. Based on a threshold it decides if the image is usable or not. However, when implementing this technique, it was noted that the amount of usable dates was greatly affected by this. Due to this, it was decided not to apply this technique.

CFMask is based upon FMaks version 3.2 and can be used as a cloud filter algorithm, and it is one of the most widely used cloud detection algorithms. It is the default cloud mask algorithm used by Landsat (Foga et al., 2017) (Xu et al., 2019). It is an algorithm that labels pixels by using a pre-defined set of threshold tests and with statistics, it further refines the method. A major downside of CFMask is that it can label snow as clouds, if it is necessary to differentiate between snow and cloud, another method would be more suitable (Fawcett et al., 2017). However, in this case that is not of relevance, for both snow and clouds should be filtered out.

Eventually, the choice was made to apply CFMask. A study comparing CFMask with ATACCA found that CFMask has an overall better performance (Foga et al., 2017). Furthermore, CFMask calculates it per pixel and is by default the cloud mask algorithm used by Landsat, which made it easy to implement.

[2] https://developers.google.com/earth-engine/tutorials/tutorial_forest_02

After applying a cloud, cloud shadow and snow mask, the bands used for calculating the NDVI (Band 3 the (red) surface reflectance and Band 4 the (near infrared) surface reflectance) were rescaled, in accordance with the documentation on the satellites[3][4]. After rescaling the value, the NDVI values were calculated and normalized to get the values between -1 and 1. Lastly, the NDVI values, date, and coordinates were saved as a CSV file. It was then further processed in python.

In Python, the data was so manipulated that for the Exploratory analysis results. This meant empty values were filtered out of the data, and the data was formatted to retain only the date, NDVI, and coordinates. Based on the coordinates, pixel IDs were determined to easily distinguish between the 3000 pixels.

To determine the amount of data per week, the number of unique week numbers in the total dataset was counted, resulting in Figure 2. It was concluded that many weeks were missing. Therefore, the data was ultimately aggregated into half-month periods.

2.4 Exploratory analysis results

To get an insight into the available data, the number of weeks in a year were analyzed. As shown in Figure 2 lots of weeks are missing. The reason for this is the cloud mask, as pixels with too much cloud cover were not included in the dataset.

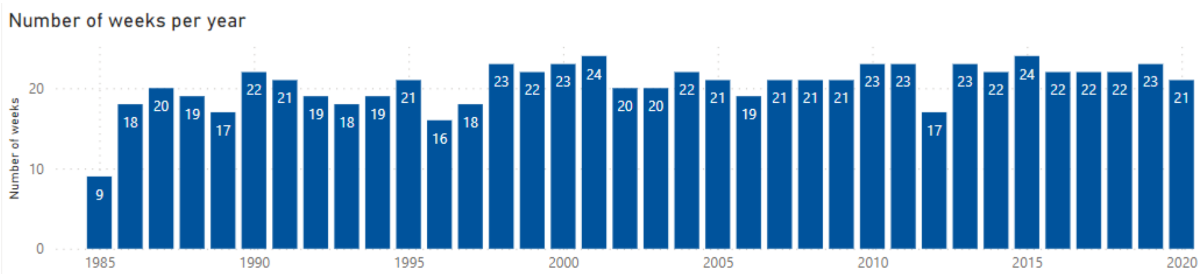

Figure 2: Number of weeks in the data per year

However, when aggregating the data per half month, there are a lot more data points available as shown in Figure 3. Due to this, it was decided to look at the NDVI per half month. To, on the one side have as much possible data available (each year has around 9 to 10 months, whereas when looking at the week level half the weeks in the year are missing). On the other side, to detect if there is a change in the month itself.

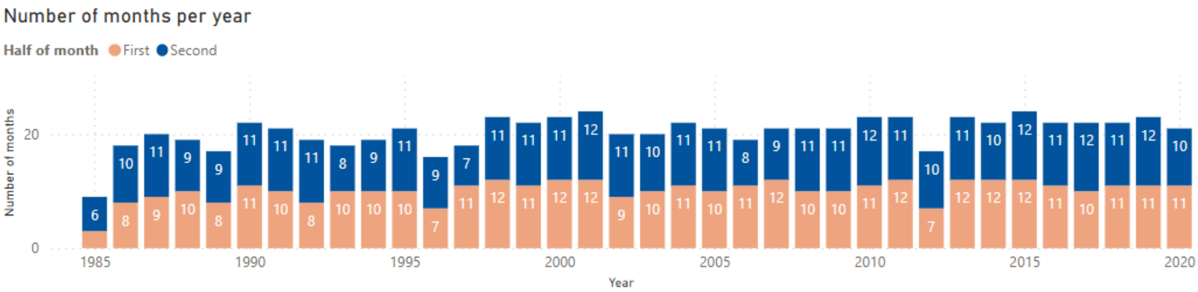

Figure 3: Number of half months per year

[3] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LE07_C02_T1_L2#bands
[4] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LT05_C02_T1_L2#bands

After preparing the data, an exploratory analysis has been done to see what the data (that ranges from 1985 until 2020) looks like. First, all the data points are aggregated by date so there is only one NDVI value per date, this is shown in Figure 4. At first glance, it looks like there is a trend present in the data.

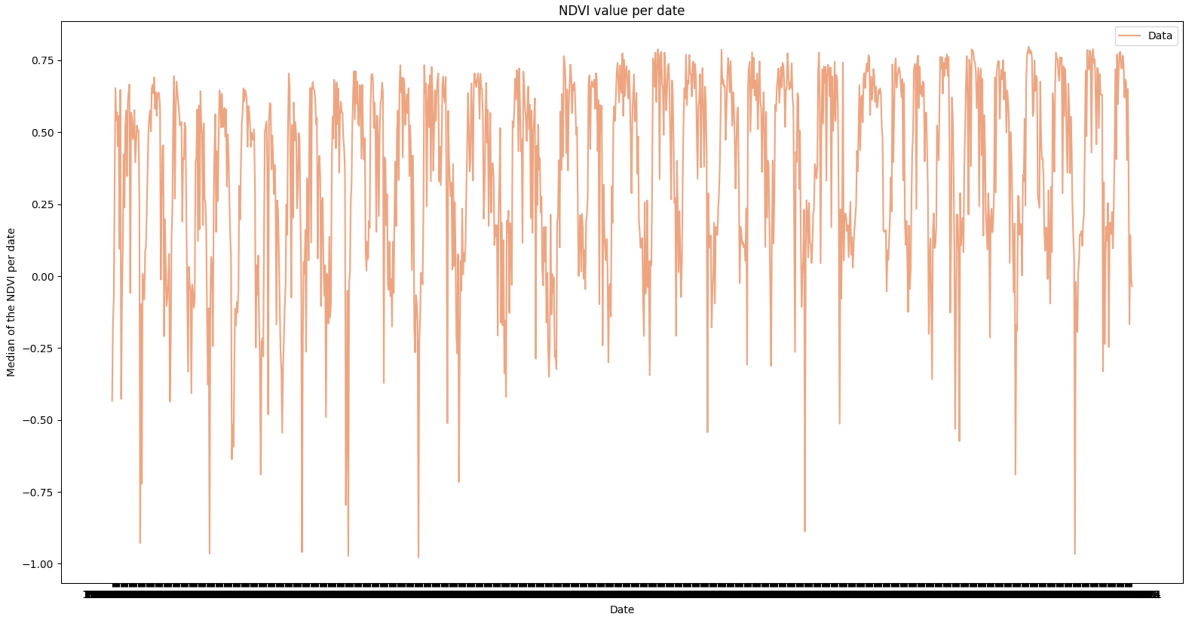

Figure 4: The median NDVI per available date

To find a trend in the data linear regression model could be applied on the data. Multiple studies used linear regression model to show a trend in the data (Li et al., 2011) (Zhang et al., 2013). Due to it being such a frequently used method it was decided to also apply it in this study to find a possible trend. For the dependent value the median was used and for the independent values, the week number with the year was used. That resulted in a trend shown in Figure 5.

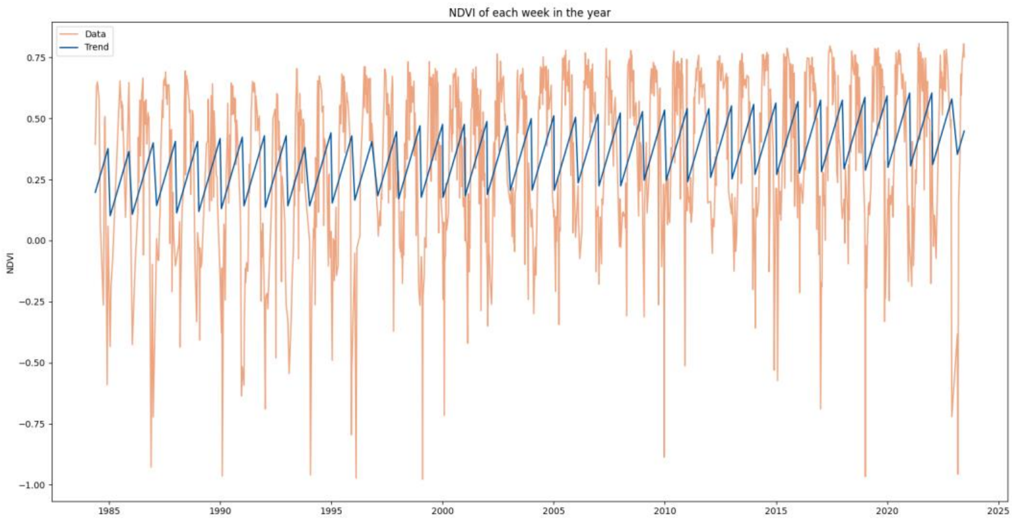

Figure 5: Trend in NDVI

The statistical analysis resulted in a F-value of approximately 61.88 and a p-value of approximately 2.67e-26, indicating a relationship between the dependent variable (median NDVI per week) and the independent variables (year with week numbers).

Furthermore, the NDVI values per half month were visualized in a matrix, revealing notable changes. Figure 6 illustrates the following observations:

In April, the median NDVI score for the years between 1984 and 2003 was 0.21. However, for the period of 2004 to 2023, the median NDVI score increased to 0.36, indicating a difference of 0.15 between the two time periods.

In May, there was a slightly greater difference. The median NDVI score for the years 1984 to 2003 was 0.51. Meanwhile, for the years 2004 to 2020, it was 0.67. This is a difference of 0.16 between the two time periods.

Overall, the NDVI scores increased over the years, with an exception for 2021. This suggests a general trend of higher NDVI values with the passage of time.

month	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	**Total**
⊟ **3**	**-0.07**	**-0.08**	**-0.08**	**-0.10**	**-0.09**	**-0.10**	**-0.11**		**-0.27**	**0.04**	**0.00**		**0.16**	**0.01**	**-0.03**	**0.04**	**0.06**	**-0.25**	**-0.05**
First	-0.07	-0.08	-0.08	0.05	-0.09	0.05	-0.09		-0.31	0.05	0.14		0.12	-0.01	-0.06	0.04	0.16		**-0.01**
Second			-0.08	-0.24		-0.24	-0.13		-0.23	0.04	-0.14		0.19	0.02	0.01		-0.03	-0.25	**-0.09**
⊟ **4**	**0.27**	**0.26**	**0.12**	**0.14**		**0.35**	**0.13**	**0.17**	**-0.03**	**0.31**	**-0.10**	**0.15**	**0.17**	**0.25**	**0.10**	**0.29**	**0.53**	**0.31**	**0.21**
First		0.12	0.09	0.14		0.25	-0.05		-0.11	0.34		0.02	0.17	0.32	-0.03	0.13	0.54	0.21	**0.15**
Second	0.27	0.41	0.16			0.44	0.30	0.17	0.06	0.29	-0.10	0.29		0.17	0.23	0.45	0.53	0.40	**0.27**
⊟ **5**		**0.53**	**0.50**	**0.45**	**0.40**	**0.36**	**0.47**	**0.50**	**0.33**	**0.61**	**0.47**	**0.53**	**0.65**	**0.53**	**0.73**	**0.48**	**0.68**	**0.59**	**0.51**
First			0.44	0.24		0.56	0.38		0.14	0.56	0.39		0.59	0.50		0.59	0.67	0.57	**0.47**
Second		0.53	0.56	0.67	0.40	0.15	0.56	0.50	0.51	0.65	0.55	0.53	0.70	0.56	0.73	0.36	0.68	0.60	**0.54**
⊟ **6**	**0.65**	**0.47**	**0.62**	**0.62**	**0.49**	**0.48**	**0.65**	**0.50**	**0.66**	**0.65**	**0.58**	**0.66**	**0.60**	**0.68**	**0.62**	**0.62**	**0.65**	**0.72**	**0.60**
First		0.35	0.57	0.57	0.41	0.43	0.65	0.52	0.70	0.65	0.48	0.65	0.68	0.69	0.66	0.58	0.65	0.76	**0.59**
Second	0.65	0.60	0.66	0.68	0.58	0.53		0.48	0.61	0.66	0.68	0.68	0.51	0.67	0.58	0.66	0.65	0.67	**0.62**
⊟ **7**	**0.55**	**0.61**	**0.62**	**0.61**	**0.12**	**0.57**	**0.62**	**0.25**	**0.51**	**0.60**	**0.53**	**0.62**	**0.53**	**0.62**	**0.67**	**0.64**	**0.59**	**0.62**	**0.56**
First		0.65	0.66		0.12	0.56	0.62	-0.10	0.50	0.65	0.49	0.66	0.60	0.63	0.64	0.67	0.60	0.65	**0.54**
Second	0.55	0.57	0.58	0.61		0.57	0.63	0.60	0.52	0.56	0.57	0.58	0.46	0.62	0.70	0.61	0.57	0.58	**0.58**
Total	**0.35**	**0.39**	**0.36**	**0.34**	**0.28**	**0.33**	**0.32**	**0.36**	**0.24**	**0.44**	**0.34**	**0.49**	**0.45**	**0.42**	**0.38**	**0.45**	**0.50**	**0.47**	**0.39**

month	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	**Total**
⊟ **3**	**0.07**	**-0.10**	**0.10**	**-0.07**	**0.15**	**0.17**	**0.11**	**0.12**	**0.11**	**0.11**	**-0.01**	**0.14**	**0.14**	**0.17**	**0.11**	**0.01**	**0.15**	**0.22**	**0.10**
First	0.14	-0.10	0.26	-0.07	0.15	0.11	0.11	0.06	0.08	0.06	-0.04	0.11	0.20	0.15	0.18	0.12	0.16		**0.10**
Second	0.00		-0.06		0.15	0.23		0.19	0.13	0.17	0.02	0.17	0.08	0.19	0.04	-0.10	0.15	0.22	**0.11**
⊟ **4**	**0.25**	**0.40**	**0.17**	**0.17**	**0.45**	**0.17**	**0.43**	**0.34**	**0.41**	**0.28**	**0.37**	**0.50**	**0.44**	**0.48**	**0.38**	**0.46**	**0.34**	**0.33**	**0.36**
First	0.21	0.27	0.04	0.17	0.33	-0.31	0.28	0.33	0.30	0.34	0.26	0.52	0.37	0.29	0.52	0.31	0.21	0.25	**0.26**
Second	0.29	0.53	0.30		0.58	0.65	0.59	0.34	0.52	0.22	0.47	0.47	0.51	0.68	0.24	0.61	0.47	0.42	**0.46**
⊟ **5**	**0.68**	**0.61**	**0.74**	**0.62**	**0.73**	**0.73**	**0.67**	**0.63**	**0.66**	**0.72**	**0.73**	**0.68**	**0.67**	**0.72**	**0.75**	**0.56**	**0.60**	**0.64**	**0.67**
First	0.67	0.54		0.59	0.79	0.72	0.66	0.48	0.64	0.73	0.72	0.62	0.72	0.69	0.71	0.36	0.71	0.50	**0.64**
Second	0.70	0.68	0.74	0.64	0.67	0.74	0.69	0.78	0.67	0.71	0.74	0.75	0.62	0.75	0.79	0.75	0.49	0.77	**0.70**
⊟ **6**	**0.65**	**0.67**	**0.71**	**0.67**	**0.59**	**0.66**	**0.70**	**0.64**	**0.73**	**0.76**	**0.73**	**0.70**	**0.77**	**0.74**	**0.78**	**0.73**	**0.76**	**0.72**	**0.70**
First		0.68	0.67	0.65	0.66	0.61	0.72	0.58	0.73		0.74	0.76	0.75		0.78	0.75	0.75	0.72	**0.70**
Second	0.65	0.66	0.74	0.69	0.51	0.70	0.69	0.70	0.72	0.76	0.72	0.63	0.79	0.74	0.78	0.70	0.76		**0.70**
⊟ **7**	**0.67**	**0.68**	**0.70**	**0.69**	**0.64**	**0.64**	**0.66**	**0.68**	**0.66**	**0.69**	**0.68**	**0.72**	**0.76**	**0.66**	**0.66**	**0.74**	**0.77**	**0.75**	**0.69**
First	0.67	0.65	0.74	0.69	0.71	0.64	0.63	0.71	0.76	0.67	0.73	0.71	0.78	0.61	0.76	0.73	0.76	0.78	**0.71**
Second	0.67	0.71	0.66	0.70	0.57	0.65	0.69	0.66	0.56	0.72	0.63	0.74	0.74	0.71	0.57	0.75	0.78	0.72	**0.68**
Total	**0.44**	**0.51**	**0.46**	**0.51**	**0.51**	**0.47**	**0.56**	**0.48**	**0.51**	**0.49**	**0.50**	**0.55**	**0.56**	**0.53**	**0.54**	**0.50**	**0.52**	**0.55**	**0.51**

Figure 6: Overview of NDVI scores for each half month for the years 1985 - 2020

2.5 Ethical Considerations of the data

The data was collected through Google Earth engine and originates from Landsat 5 and 7. The data is for everyone accessible and can be freely used. Furthermore, no personal data is used whatsoever, and the images obtained are concerning forest areas.

3. Method

3.1 Method selection

To analyze the data and find a statistical relation between de NDVI and the dates, the Seasonal-Mann Kendall test was utilized to find a trend in the data. First, the data was aggregated over 3 years. As was shown during the exploratory analysis there were not a lot of unique weeks in the years available. Even when a year has data for 26 weeks, that means for at least half of the instances there is no data. To try to “gain” more data in a year, the data was aggregated by 3 years, where each cluster was named the latest year in the cluster. So, for example the years 1985, 1986, and 1987 were aggregated into one cluster named 1987.

After that, the Seasonal-Mann Kendall test was applied to calculate the slope and intercept per pixel for each month and each cluster, resulting in different slopes and intercepts for the month and cluster per pixel. These data were then aggregated by month and cluster and the median was taken. Next all the intercepts and slopes for the different months and clusters were displayed in a scatterplot.

The same was also done for each year, where it only looked at the month from March till July. The intercept and slope are then only calculated for each year, instead of each year and month.

The seasonal Mann-Kendell test is a non-parametric test for finding out whether there is a monotonic trend in seasonal data. It takes into account the seasonal effects that might be present in the data. It allows for the identification of trends with specific time intervals. It examines the presence of a monotonic trend in time series data, considering seasonal effects. Monotonic refers to the pattern in data. If one variable increase and so does the other, of both degreases then there is a monotonic relation between the variables However, if this is not the case then there is no monotonic relation (Glen, sd)

It works by dividing the dataset into subgroups, each corresponding to a particular season. Then, it applies the Mann-Kendall test to each subgroup.

The seasonal Mann-Kendell test was used because it takes into account the seasonal effects and allows for the detection of trends with specific time periods.

4. Results

For visualizing the result, powerbi was used. The reason for this was because it is so easy to predefine the layout and visuals. The results were as follow:

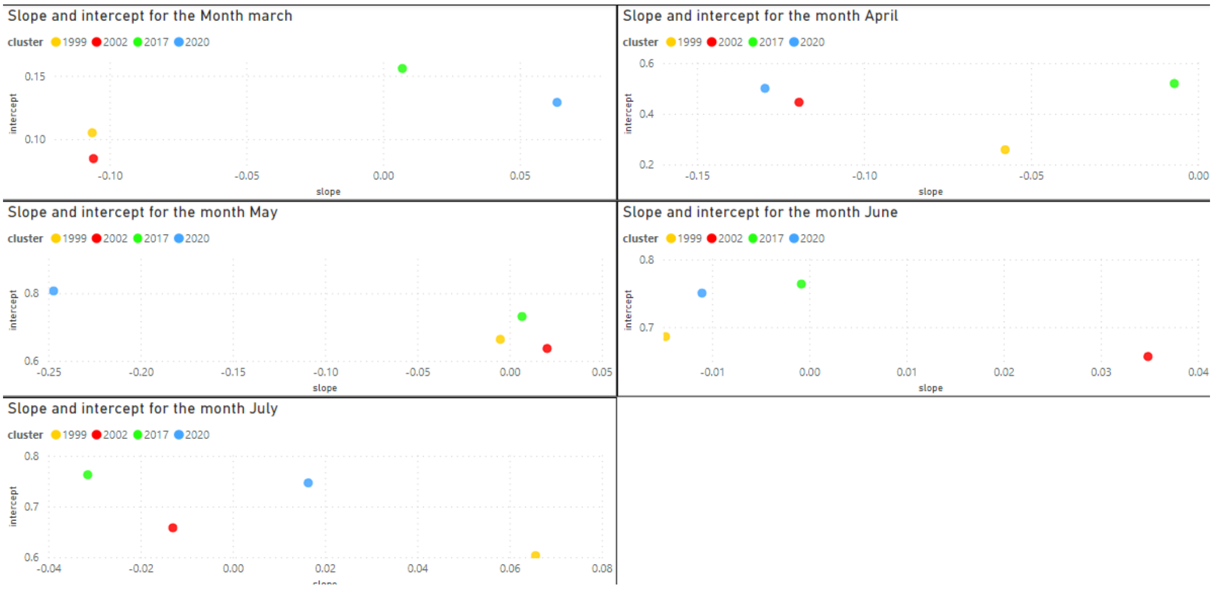

Figure 7: the intercept and slope for different months and different clustered years

Figure 7 provides an insight into the NDVI over time. The graphs contain the slope and intercept values for March to July for different clusters (1999, 2002, 2017, and 2020).

The slope reflects the rate of change in NDVI over time, meaning it illustrates whether vegetation cover is increasing (positive slope) or decreasing (negative slope) in a given month. However, the intercept provides a baseline of the NDVI value at the start of the period.

In 1999, there was a increase in the slope from March (-0.106) to July (0.0656). This indicates an increase in vegetation over the time period. The intercept values also rose significantly over the same period.

Contrarily, for the cluster 2002, the dataset shows a decrease in slope from March (-0.1059) to April (-0.1194) but an increase from April to July (-0.0130). This could signify a decrease in vegetation in the early spring followed by a subsequent recovery.

For the cluster 2017, the slope values remain relatively stable compared to the previous clusters, showing minor fluctuations across the months. It suggests a relatively stable vegetation cover in that year. Similarly, the intercept values also remain relatively high, which suggests an overall high amount of greenness during that period.

Finally, the cluster 2020, the dataset shows an increase in the slope from March (0.0636) to a decrease in May (-0.2471) followed by a slight increase in July (0.0163). This could be indicative of fluctuations in vegetation growth during the year.

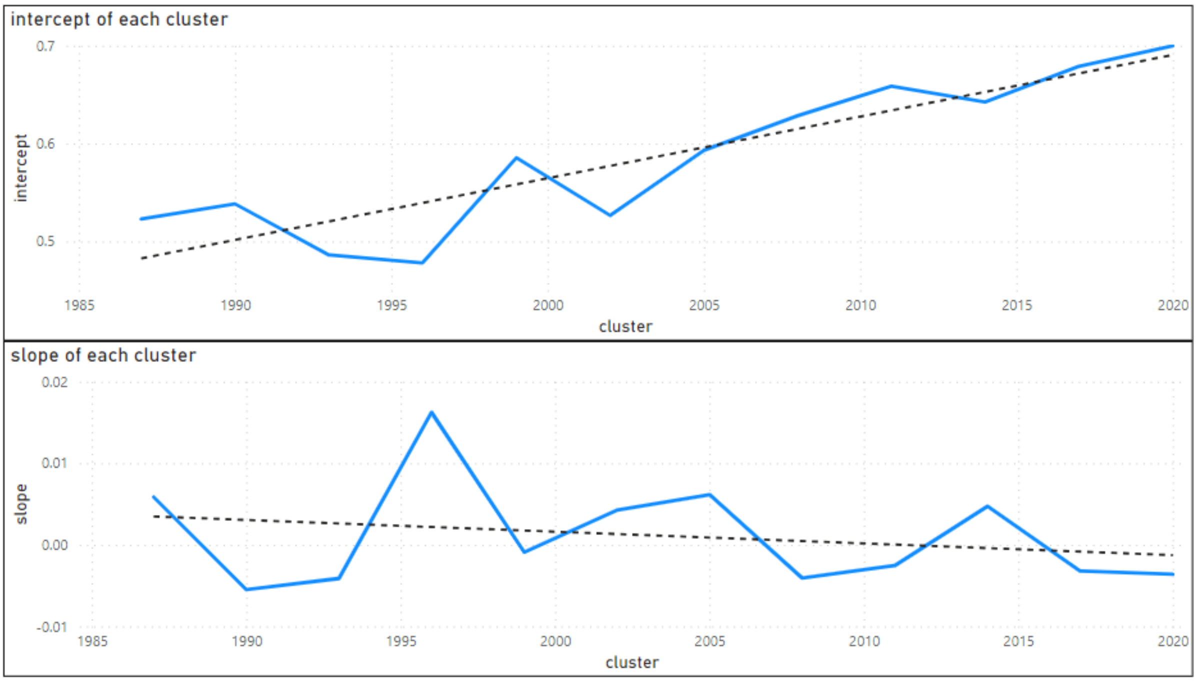

Figure 8: the intercept and slope for each cluster

In Figure 8Figure 8 the intercept and slope per cluster are visualized. The slope represents the trend of the NDVI for each cluster. A positive slope indicates an increase in vegetation over time, while a negative slope indicates a decrease. The intercept, on the other hand, represents the starting NDVI value for each year. This based on the months march till July.

The data shows varied trends over different years. In 1987, the positive slope (0.0059) suggests an increasing trend in vegetation. Similarly, in 1996 and 2005, the positive slopes suggest increasing trends in NDVI. However, the clusters 1990, 1993, 2008, 2011, 2017, and 2020, with negative slopes, hint at decreasing trends in NDVI for the specified months. The overall trend is decreasing.

Moreover, the intercept values fluctuate over time, ranging from around 0.47 to almost 0.70. The increasing trend in intercept values, especially after 1993, may imply an overall increase in NDVI values for the different clusters. This may suggest an increased greenness in the march of each respective cluster.

5. Conclusion

5.1 Answering the research question.

In conclusion, the aim of this research was to find whether there is a shift in the period of trees turning green, and if this shift be detected through the utilization of satellite imagery. Normalized Difference Vegetation Index (NDVI) scores from the last 36 years (1985 – 2020) were used to determine this change. The Seasonal-Mann Kendall test was used to see whether there was a statistical relationship and trend between NDVI scores and dates. This test helped to take into account the seasonal effects and detect trends within specific time periods.

Looking only at the exploratory analysis it would suggest that there is a shift in the moment trees turn green. The NDVI values for the first clusters were much lower compared to the later clusters in the same period. This may suggest that the blooming time is progressively getting earlier. However, when examining the results of the Seasonal Mann-Kendall analysis, it becomes somewhat challenging to draw definitive conclusions.

However, one thing is notable, when looking at the intercepts of the clusters a clear positive trend is seen. This could mean that the overall NDVI values are increasing and that could indicate that the date for the starting point is also changing. Because the period considered is from March to July, it could indicate that the blooming time of trees is getting earlier. This is suggested by the increasing trend of the intercept. This could also explain the decreasing slope, as each month has a progressively higher NDVI, causing the NDVI values to decrease relative to other months. Therefore, this would imply that the flowering time is earlier.

5.2 Future work

Several things for future research can add to this study. Firstly, incorporating other climate components such as temperature or rainfall into the analysis would enhance the understanding of the relation between plant lifecycle, years, and climate-chancing factors. Study already showed that there is a correlation between vegetation and temperature and rainfall (Kaufmann et al., 2003). Another study found that there was from 1975 till 2005 an increase of 0.2 °C per 10 years (Hansen et al., 2006). This could explain as to why there is a change in the period of leaf greening.

Secondly, utilizing Landsat 4 and 8 data as additional data sources could provide more data, which might make it possible to look at the years individually instead of aggregating the years, this could improve accuracy and enables finer-scale analyses.

Furthermore, exploring the timing of leaf senescence, or the period when leaves fall from trees, would offer valuable insights into the seasonal dynamics of the studied ecosystem. Investigating the factors that influence leaf abscission timing, such as temperature, photoperiod, and species-specific characteristics, would contribute to a more comprehensive understanding of vegetation phenology.

Incorporating these aspects into future research endeavors would help expand the current knowledge base and provide a better understanding of the dynamics and factors influencing vegetation patterns and processes in the study area.

Appendix A: References

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., . . . Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment, 120*, 25-36. https://doi.org/10.1016/j.rse.2011.11.026

Escobar-Silva, E. V., De Souza, A. R., & Ferreira, M. L. (2022). *FMASK algorithm applied to Landsat images: a study case.*

Fawcett, D., Leiterer, R., Heisig, H., Wulf, H., Kellenberger, T., & Joerg, P. C. (2017). *Google Earth Engine product prototypes.* Google Earth Engine.

Foga, S., Scaramuzza, P. L., Guo, S., Zhu, Z., Dilley Jr, R. D., Beckmann, T., . . . Laue, B. (2017). Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sensing of Environment, 194*, 379-390. https://doi.org/10.1016/j.rse.2017.03.026

Gargano National Park. (n.d.). Retrieved from PeakVisor: https://peakvisor.com/park/gargano-national-park.html#flora

Gessesse, A. A., & Melesse, A. M. (2019). *Extreme Hydrology and Climate Variability: Monitoring, Modelling, Adaptation and Mitigation* (1 ed.). Elsevier Inc. https://doi.org/10.1016/B978-0-12-815998-9.00008-7

Glen, S. (n.d.). *Monotonic Sequence, Series (Monotone): Definition*. Retrieved from Elementary Statistics for the rest of us!: https://www.statisticshowto.com/sequence-and-series/monotonic-sequence-series-function/#relationship

Hansen, J., Sato, M., Ruedy, R., & Medina-Elizade, M. (2006, July 31). Global temperature change. *PNAS, 103*(39), 14288 –14293. https://doi.org/10.1073/pnas.0606291103

Kaufmann, M. K., Zhou, L., Tucker, C. J., Slayback, D., Shabanov, N. V., & Pinzon, J. (2003). The effect of vegetation on surface temperature: A statistical analysis of NDVI and climate data. *Geophysical Research Letters, 30*(22). https://doi.org/10.1029/2003GL018251

Li, B., Yu, W., & Wang, J. (2011). An Analysis of Vegetation Change Trends and Their Causes in Inner Mongolia, China from 1982 to 2006. *Advances in Meteorology, 2011*, 8. https://doi.org/10.1155/2011/367854

Licht, W., & Wagensommer, R. P. (2020). *Flora vascolare del Gargano e delle Isole Tremiti. Chiavi analitiche per la determinazione.* Biblioteca Verde del Parco Nazionale del Gargano.

Papes, M., Peterson, A. T., & Powell, G. V. (2012). Vegetation dynamics and avian seasonal migration: clues from remotely sensed vegetation indices and ecological niche modelling. *Journal of Biogeography, 39*(4), 652. https://doi.org/10.1111/j.1365-2699.2011.02632.x

Pastor-Guzman, J., Atkinson, P. M., Dash, J., & Rioja-Nieto, R. (2015). Spatio temporal Variation in Mangrove Chlorophyll Concentration Using Landsat 8. *Remote Sensing, 7*(11), 14530-14558. https://doi.org/10.3390/rs71114530

Pettorelli, N., Vik, J. O., Mysterud, A., Gaillard, J., Tucker, C. J., & Stenseth, N. C. (2005, September). Using the satellite-derived NDVI to assess ecological responses to environmental changes. *TRENDS in Ecology and Evolution, 20*(9), 503-510. https://doi.org/0.1016/j.tree.2005.05.011

Sharma , G. (2017). Pros and cons of different sampling techniques. *International Journal of Applied Research, 3*(7), 749 - 752. Retrieved June 30, 2023

Wulder, M. A., White, J. C., Loveland, T. R., Woodcock, C. E., Belward, A. S., Cohen, W. B., . . . Roy, D. P. (2016). The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment, 185*, 271-283. https://doi.org/10.1016/j.rse.2015.11.032

Xu, K., Guan, K., Peng, J., Luo, Y., & Wang, S. (2019). *DeepMask: an algorithm for cloud and cloud shadow detection in optical satellite remote sensing images using deep residual network.* Cornell University. https://doi.org/10.48550/arXiv.1911.03607

Zhang, Y., Gao, J., Liu, L., Wang, M., & Yang, X. (2013). NDVI-based vegetation changes and their responses to climate change from 1982 to 2011: A case study in the Koshi River Basin in the middle Himalayas. *Global and Planetary Change, 108*, 139-148. https://doi.org/10.1016/j.gloplacha.2013.06.012

Appendix B: Links

Link to the code:
https://github.com/bootje2000/Thesis_ADS_2023

Link to the data:
https://drive.google.com/drive/folders/1KLb-TUgj9FVv82L2iKWtvwb5rgSDkIyq?usp=sharing

Link to a translation of the study done by Licht, W., & Wagensommer, R. P.:
https://docs.google.com/document/d/1RN5VyQfXz4DOfpI8zkcdhN1cUIu-fwT3GhwzvQ_NkY4/edit?usp=sharing