PARK UNDER STRESS: MODELLING NDVI FOR DAYS WITH MISSING DATA

ABSTRACT

In Urban parks provide multiple ecosystem services, drought can lead to damage to vegetation and a reduction in the services. The NDVI can help to analyse the vitality of the vegetation but is often not available in fine resolution. In this project, a random forest model was trained to model the NDVI.

Paulina Harder

Advanced Environmental Modelling

Table of Contents

1. Introduction

1.1 Provide the motivation and justification for the chosen topic. *Why did you choose this topic?* (max. 300 words)

Urban parks provide a large number of ecosystem services and have a positive impact on public wellbeing (Larson et al., 2016). Among them are regulating services such as the improvement of the micro-climate conditions, air purification and mitigation of noises (Elderbrock et al., 2020, Nowak et al. 2018) but also cultural services like recreation and parks being a place for social interaction and physical activity (Dickinson & Hobbs 2017). For example, Berlin's largest urban park, Tempelhofer Feld serves as a social meeting point during all seasons, harbouring community gardens and promoting health, social participation and environmental justice. At the same time, it allows human nature interaction and the exploration of urban nature (Brenck et al., 2021). However, with increasing number of dry summer seasons e.g. 2018, 2020 and 2022 and periods of extreme heat that act as stressors to the vegetation, urban green areas may no longer be able to provide these ecosystem services (Kabisch & Haase, 2018). Yet especially during such extreme weather events which are often amplified in built-up and highly artificial urban environments, adaption is especially needed to mitigate the effects on the urban residents (Guerreiro et al., 2018). For future park vegetation planning and park management it is essential to account for extreme conditions when planning climate-resilient parks that continue to provide valuable ecosystem services. Therefore, it is important to assess how current urban park vegetation is performing under hot and water scarce conditions. It would be of interest know the dynamics of the vegetation, especially regarding when there is damage and recovery, how often it recovers and how much water would trigger it. Therefore, my aim is to investigate how the vegetation of an urban park performs under water stress. Although it seems unlikely that urban parks like Tempelhofer Feld or Tegeler Stadtheide would be irrigated as irrigation would be a strain on cities with already limited d urban water availability during drought periods (Choi et al., 2021)

1.2 Describe the objectives of the project. *What do you want to achieve with your model?* (max. 200 words)

For addressing the above-mentioned challenge, it is important to have data available on park conditions and the stresses that vegetation has to endure. To assess the vitality and condition of the park vegetation, the Normalized Difference Vegetation Index (NDVI) derived from remote sensing can be employed. However, when the NDVI data is usually not available in a daily resolution and it was only a scarce dataset, much of it was excluded in the quality filtering process and a continuous time series that would reflect the vegetation's response throughout the growing season and especially during water scarce periods is lacking. Such a timeseries would be beneficial for performing further analysis. Therefore, the aim of this project is to build a model that can predict accurately the NDVI for urban parks. Using this model, gaps in NDVI datasets can be filled. Through the model building process key environmental variables influencing vegetation's vitality can be identified. That could prove insightful when wanting to develop a park drought monitoring system to decide which variables should be considered. With a full timeseries and such insights, critical thresholds at which the vegetation begins to show stress could be identified.

1.3 What are the foreseeable challenges for the development of your project? (max. 200 words)

The main challenge of my model is the small amount of NDVI available data per year as it made the model prone to overfitting. Initially, there were on average 136 satellite images available per year, but after cutting and applying a quality filter only between 27 and 37 images remained. However, throughout a year, the images are not evenly distributed, sunny days are well covered, but in rainy periods, there no data due to clouds blocking remote sensing images. Therefore, there could be a bias in the training and testing data of my model towards certain weather conditions and some vegetation recovery patterns could be missing. Additionally, I only have six full seasons of data. For three of them the NDVI remained in a rather stable range, while three seemed to be extreme years. With only scarce data on both possibilities, it could be difficult to teach the model to capture peaks or to determine what would be an average year pattern.

2. Materials and Methods

2.1 Describe the problem and system boundaries (max.300 words)

The aim of the project is to model the NDVI for park vegetation. Therefore, I selected as study area for this project is the urban park Tempelhofer Feld (Figure 1) a former airfield located in the centre of Berlin, popular with the residents of the city (Brenck et al., 2021). It was selected because it is one of the largest parks in the city. The size of ca. 200 ha vegetated areas allows it to be clearly identified on NDVI images without having to include surrounding areas. Moreover, it harbours classic park vegetation, primarily grassland (Brenck et al., 2021). The vegetation coverage of the area is quite homogenous, therefore it should display similar responses to dryness and follow a similar vegetation period. That is advantageous as sometimes in remote sensing some pixels are missing and not included in the NDVI calculation because the they are shaded by a cloud or simply not in the image. The homogeneity should prevent a large bias, which would be possible if the area was partly covered by trees or shrub and would increase the variability of the NDVI.

The temporal extent are the vegetation seasons for the years 2017 to 2023, for those years we have necessary NDVI data available, starting 1.4. to 15.11. That time window covers the whole vegetation period with peaks in May, dryness effects in summer months and early autumn (QUELLE). The temporal resolution is daily. I used a random forest model (RF) to predict the NDVI. It has been already used to estimate NDVI (Mohite et al., 2020) reconstruct values (Sun et al., 2023) where it outperformed other models.

2.2 Describe your model concept (empirical, conceptual, process or data based, etc). What are your target and state variables, drivers and parameters?

For predicting the missing NDVI I started to test several models but quickly it became obvious that a RF would be most suitable. Additionally, it is robust to overfitting which important given the scarcity of my data. A random forest is a supervised machine learning algorithm. With a RF models predictions can be made based on the data by learning its patterns but it does not explain how different variables interact. Random forest models are empirical relying on observed data to be built and validated (Breiman, 2001). My target variable is the value I am aiming to predict: the NDVI. Drivers were selected from the soil moisture, evaporation and climate variables. To select the right drivers, I first created a correlation matrix to do a pre-check of which features could be important to model the NDVI. For the final drivers, I relied on the matrix but I also on the drivers that enhanced the performance of the model (for performance metrics check 3.2). I was careful not to add too many to prevent overfitting. My final drivers were mean potential evaporation of the last seven days, %nFK in the soil depth 10-20 cm, mean maximum temperature of the last seven days and mean %nFK in the soil depth 0-10 cm of the last seven days.

2.3 Provide the mathematical basis of your model

A Random Forest is an ensemble learning model based on decision tree. They consist of a root node, branches, internal nodes and leaf nodes. At every node, the training data is split based on feature values. In the splitting process, the aim is to decrease the prediction error, only at the last ones, the leaf nodes, the average of the target values (NDVI) are used as the prediction (Pal, 2005). A RF trains several different decision trees but it is not merely a group of individual trees as the model takes the average prediction across the decision trees to improve the predictions in accuracy and prevent overfitting. Generally, the combination of the trees makes a RF more robust then simple decision trees.

2.4 Describe the data used in your model implementation (source, resolution, extent, etc).

So far, there has been not a consecutive dryness monitor for urban parks in Germany measuring soil moisture or evapotranspiration. The DWD provides calculated historical daily values of real evaporation, potential evaporation and soil moisture among other values are derived from characteristic features of the soil and the plant population. The values are calculated for most of the running weather stations of the DWD. For the calculation of potential evaporation and soil moisture the model AMBAV is used, while the potential evaporation is calculated according to FAO. The model's output is checked with calculated soil moisture only on occasion of special measurement campaigns and it is very satisfactory (DWD, 2024). For my model building, the station of Berlin-Tempelhof (433) was chosen as it is closest to Tempelhofer Feld.

As an approximation for vegetation health Normalised Difference Vegetation Index (NDVI) was employed. It has been widely used for vegetation monitoring and as Jin & Eklundh (2014) state as well, the NDVI is popular because it displays a relative robustness to noise and sun-sensor geometry variations. Additionally, it is widely and easily accessible globally for a long time. Two limitations acknowledged by the authors are a strong reaction by the index to soil background (e.g. snow) and that the NDVI struggles with saturation at high vegetation density. As this study only focuses on spring summer and early autumn, snow will not be an issue and a thorough quality evaluation will be conducted. Moreover, the main interest is on periods without a high vegetation density.

The NDVI images were derived from Sentinel-2 satellites A and B providing a spatial resolution of 10 m allowing for a detailed analysis with a revisit time of 5 days (Smet et al., 2023). The NDVI images and quality filters (QFLAG) were derived from the free and opensource Copernicus HR-VPP dataset and were downloaded from the WEkEO platform (https://www.wekeo.eu/) platform. Data from this dataset has been successfully used before, also for similar applications (Charrière et al., 2024, Borgogno-Mondino & Fissore 2022 and Filgueiras et al., 2019). I have added an exact description on the data on GitHub. NDVI layers were downloaded for the period between 01.04. and 15.11. for the years 2017 to 2023. In the pre-processing the NDVI and the QFLAG images were first cropped to the extent of Tempelhofer Feld. Then, following the recommendation of Smet et al. (2023) a medium filter by masking the NDVI pixel for the QFLAG2 values ranging from 4 to 2048 was applied. With the filter, pixels influenced by snow, clouds and their shades their surrounding pixels were removed. Finally, the mean NDVI for each available day was calculated. After the filtering process between 27 and 37 images remained per year. Finally, I created dataframe containing the variables derived from DWD and the NDVI.

3. Results

3.1 Present here an overview of your model. What does it work? Present examples of its usage and results of the analysis you performed.
First of all, I modelled the NDVI for the days between the 1.4. and 15.11. of the years 2017 that lack data (Figure 2) and we have a complete dataset for this period. A full NDVI dataset could be used for threshold analysis. Daily NDVI data could allow to analyse vegetation in greater detail. Recovery could be evaluated e.g. how fast and how often vegetation recovers and how much precipitation would be needed to trigger a substantial recovery. The most important factors were revealed to be mean potential evaporation of the last seven days, %nFK in the soil depth 10-20 cm, mean maximum temperature of the last seven days and mean %nFK in the soil depth 0-10 cm of the last seven days. As this is not data that is currently measured in parks, an early warning system cannot be build, we can only assess in hindsight when the vegetation was stressed.

3.2 Model performance evaluation. How well does your model perform? Describe here how you assessed model performance and what it was.
I assessed my model on its general performance and as I have only scarce data, I gave attention not to overfit the model. The general performance was assessed with the metrics that were suggested in class: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Nash-Sutcliffe Efficiency (NSE) and RMSE-observations standard deviation ratio (RSR), results in Table 1. Drawing from the scores, the model seems to be performing overall well, however it can be also a sign of overfitting. To prevent it, I performed a ten-fold cross-validation analysis. With such analysis, the model is trained and tested on k splits of the data, making sure that every data point is used for training and testing. This method should provide a more robust estimate of the prediction error for new data (Fushiki 2009). Therefore, I performed 10-fold cross-validation on my RF model using the MSE as the scoring metric to assess the model's performance and stability with the different folds. As the MSE punishes larger errors, it seemed suitable because those would suggest that the model failed to capture important patterns in the data. The result of 0.0039 is slightly worse then in the original model. Still, it seems that my model makes reliable predictions even with the 10-fold which also indicates that it is not overfit. To further assess the performance of the model I plotted the residuals in a histogram (Figure 2). Although a random forest is a non-parametric model and thus does not require a normal distribution, histograms can reveal if my model fails to detect certain patterns or is unable to capture peaks. Additionally, strong skewness or extreme values can suggest that there

are certain conditions which my model cannot make solid predictions. The histogram revealed a moderate performance. It did produce a slight skewness but it was still in an acceptable range and with some weaknesses in certain areas.

Finally, learning curves were plotted (Figure 3). They are common tool for machine learning algorithms. Their learning process be assessed by visualising the train and test scores against the training set sizes in curves. The former reveals how well the model is learning while the latter shows how far the model is able to generalise. Their shapes indicate if the model provides generally a good fit or under-overfitting. My model certainly does not display an ideal learning curve but does not exhibit severe overfitting or underfitting.

3.3 Model communication. Have you prepared a manual for your model, How people can access it? Is there an interactive interface? Present those items here.

All the data and code I used will be made available as a GitHub repository. There will be a brief description of the model and a manual for everyone wanting to replicate the study or to apply it to another location. The scripts have comments making it easy to follow for users with basic programming knowledge. All the data that I used and that would be necessary for an own analysis are open source. Only four pieces of input are needed to replicate the study. Since the model has only been tested for one location and for limited years and it is yet to produce very reliable results, I cannot confirm that it will work well for other areas. Therefore, I did not create a dashboard for entering own soil moisture data and producing NDVI data as it could give results that are not as exact as they should be yet.

4. Discussion and self-reflection

4.1 What experiences did you gather during the project when it came to transferring theoretical course knowledge to application in practical work? (max. 300 words)

First, I found it challenging to start from having weekly sessions and running together through the scripts to implementing one’s own project with own that data that brings up its own challenges. During the classes in the computer lab, I sometimes struggled to fully understand all the scripts, especially because everything was prepared for us and running smoothly which made it hard to see the implementation details or the mechanisms behind some functions. When building my own model, it was at first difficult to implement the code with my own data. However, that process forced me to engage more with the course content for improving my model. Here the lectures and scripts that were provided proved to be valuable and were connecting theory and practice for me. Additionally, when building the model I felt sometimes insecure regarding for modelling rules and conventions, e.g. for some analysis data must be withhold or when a model would be considered overfitting. I believe that comes with certain experience but sometimes during the implementation I was unsure if the model was actually working well or whether there is an obvious mistake which makes the model useless.

4.2 What did you learn from the project? What has proven successful? Where do you see potential for improvement? (max. 300 words)

One of the most valuable outcomes was gaining hands-on experience with modelling. The model building processes involved a lot of drawbacks and mistakes I made but in the future, I will be better equipped for them. Overall, I was able to gain a better understanding on modelling techniques and evaluating outcome from which I am sure I will benefit from in my future studies. Potential for improvement lies in my time management. I spent most of the time coding, building and implementing the model as I considered it the most important task. However, I later realised that I neglected other things like the communication of the model and analysing the results that are as crucial as building the model. Therefore, I had to rush those parts which lowered their quality and at the same time it impacts the overall project.

4.3 What were the unforeseen challenges you faced? How did you solve them? (max. 300 words)

When I started building my model, I included all the NDVI measurements without performing an outlier performance in advance. When my model was performing extremely well (NSE = 0.9) with only few variables, I removed gradually some of them but the model kept performing well. Then I plotted a learning curve which revealed that the model was not performing as well as the metrics suggested. I did a test run without the outlier, then the model suddenly performed much worse (NSE = 0.55). Therefore, I had to start the model building and parameter selection anew. Additionally, I added the k-fold cross correlation validation. So far, the model fails to predict reliable NDVI values for days without data, thus it needs to be trained on natural continuity of the NDVI. So far it predicts values independently from the values of the previous day and thus the NDVI fluctuates between the days not accurately refecting reality. For example, for two consecutive days in July of 2018 the model predicted an NDVI of 0.34 and 0.64. In order to perform threshold or lag analysis the model would need to perform better to perform it with realistic assumtptions. Additionally detailed analysis of the NDVI is not possible, because so far, according to the model, the vegetation recovers within three days, therefore it should not be used yet. In conclusion, the model needs substantial improvement to actually be used. The model could be improved by modelling the daily change of NDVI instead as the is influenced by the previous day and is not in isolation. Moving forwards instead of a RF, time series models can be employed such as a LSTM.

FIGURES AND TABELS

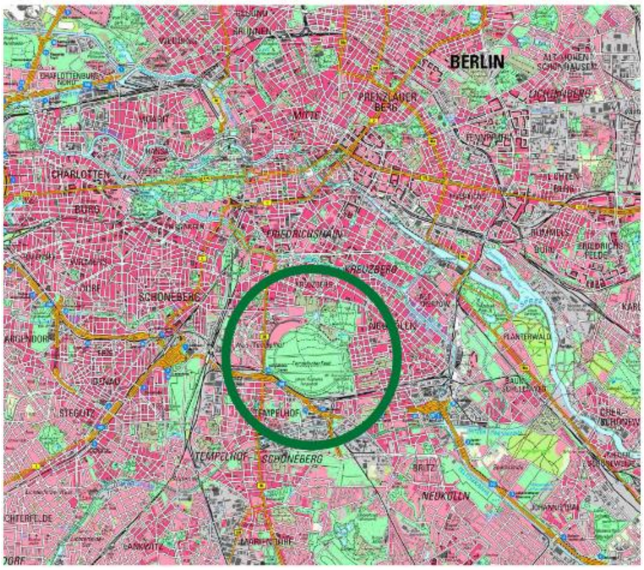

Figure 1 The location of the study site Tempelhofer Feld in Berlin, from the Geoportal of Berlin, figure produced by Brenck et al., 2021

Table 1: Scores of the evaluation metrics of the model

Metric	Score
MAE	0.0384
MSE	0.0024
RMSE	0.0489
NSE	0.8513
RSR	0.3856
k-fold MSE	0.0039

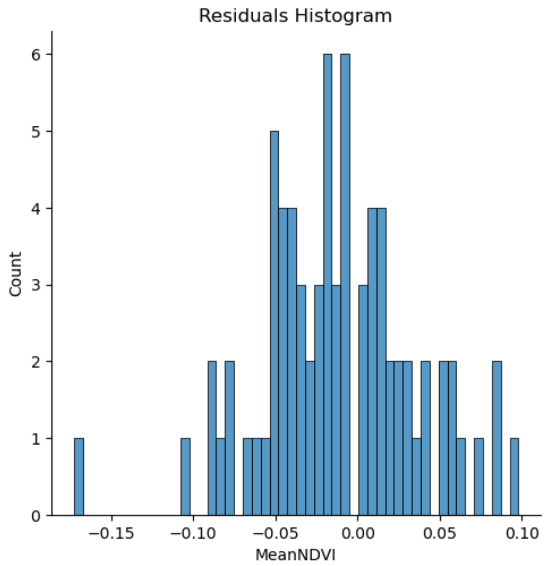

Figure 2 Residuals of the model in a histogram

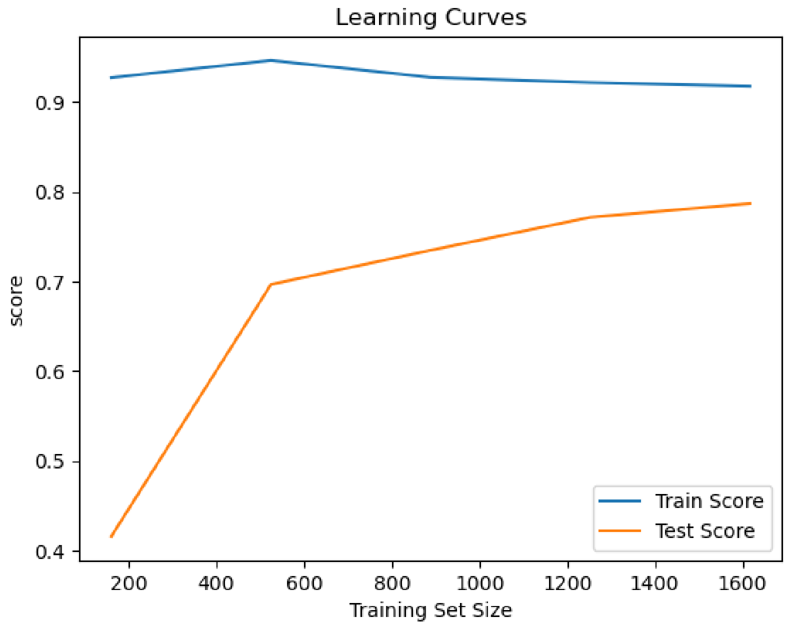

Figure 3 Learning curves of the model to check it on overfitting

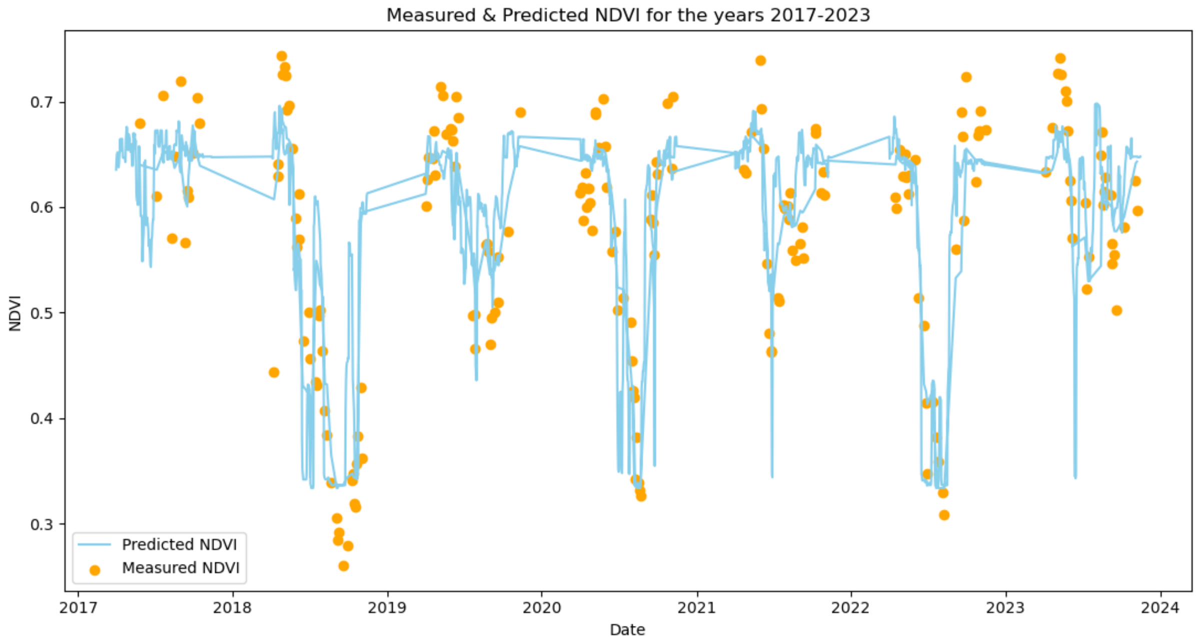

Figure 4 Modelled the NDVI for the days lacking NDVI data for Berlin-Tempelhof

REFERENCES

Borgogno-Mondino, E., & Fissore, V. (2022). Reading Greenness in Urban Areas: Possible Roles of Phenological Metrics from the Copernicus HR-VPP Dataset. *Remote Sensing*, *14*(18), 4517. https://doi.org/10.3390/rs14184517

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32, 2001

Brenck M., Hansjürgens B., Schröter-Schlaack C., Tröger U., Wessner A., Wittmer H. (2021). Gesellschaftliche Wertigkeit des Tempelhofer Feldes – Qualitäten erfassen und sichtbar machen. Helmholtz-Zentrum für Umweltforschung – UFZ, Leipzig. https://www.econstor.eu/handle/10419/246582

Charrière, D., Francon, L., & Giuliani, G. (2024). NDVI or PPI: A (Quick) Comparison for Vegetation Dynamics Monitoring in Mountainous Area. *Remote Sensing*, *16*(20), 3894. https://doi.org/10.3390/rs16203894

Choi, C., Berry, P., Smith, A. (2021). The climate benefits, co-benefits, and trade-offs of green infrastructure: a systematic literature review. *J. Environ. Manag*. 291, 112583. https://doi.org/10.1016/J.JENVMAN.2021.112583

Dickinson D.C., Hobbs R.J. (2017). Cultural ecosystem services: characteristics, challenges and lessons for urban green space research. *Ecosystem Services* 25:179–194. https://doi.org/10.1016/j.ecoser.2017.04.014

DWD (2024). Berechnete historische Tageswerte von charakteristischen Elementen aus dem Boden und dem Pflanzenbestand, Version v2. Link

Elderbrock E., Enright C., Lynch K.A., Rempel A.R. (2020). A guide to public greenspace planning for urban ecosystem services. *Land*. https://doi.org/10.3390/land9100391 , 9.

Filgueiras, R., Mantovani, E. C., Althoff, D., Fernandes Filho, E. I., & Cunha, F. F. d. (2019). Crop NDVI Monitoring Based on Sentinel 1. *Remote Sensing*, *11*(12), 1441. https://doi.org/10.3390/rs11121441

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Stat Comput* 21: 137–146 https://doi.org/10.1007/s11222-009-9153-8

Guerreiro, S.B., Dawson, R.J., Kilsby, C., Lewis, E., Ford, A. (2018). Future heat-waves, droughts and floods in 571 European cities. *Environ. Res. Lett*. 13 https://doi.org/10.1088/1748-9326/aaaad3

Hongxiao J., Eklundh, L. (2024). A physically based vegetation index for improved monitoring of plant phenology, *Remote Sensing of Environment*, Volume 152, Pages 512-525, https://doi.org/10.1016/j.rse.2014.07.010.

Kabisch N., Haase D. (2018). Urban nature benefits – opportunities for improvement of health and well-being in times of global change. *WHO. Newsl House Health*. 29:1–10.

Larson, L.R., Jennings, V., Cloutier, S.A. (2016). Public parks and wellbeing in urban areasof the United States. *PLoS One* 11, e0153211. https://doi.org/10.1371/journalpone.0153211

Mohite, J. D., Sawant, S. A., Pandit, A., and Pappula, S. (2020). Investigating the performance of random forest and support vector regression for estimation of cloud-free NDVI using Sentinel-1 sar data, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B3-2020, 1379–1383, https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-1379-2020 2020

Nowak D.J., Hirabayashi S., Doyle M., McGovern M., Pasher J. (2018). Air pollution removal by urban forests in Canada and its effect on air quality and human health. *Urban For Urban Green*. 29:40–48. https://doi.org/10.1016/j.ufug.2017.10.019

Pal, M. (2005). Random forest classifier for remote sensing classification. *Int. J. Remote Sens*. 26 (1), 217–222. https://doi.org/10.1080/01431160412331269698

Smets B., Cai, Z., Eklund, L., Tian, F., Bonte, K., Van Hoost, R., Van De Kerchove, R., Adriaensen, S., De Roo, B., Jacobs, T., Swinnen, E. (2023). High Resolution Vegetation Phenology and Productivity (HR-VPP), Daily Raw Vegetation Indices. Issue 1.1. Copernicus Land Monitoring Service (CLMS)

Sun, M., Gong, A., Zhao, X., Liu, N., Si, L., & Zhao, S. (2023). Reconstruction of a Monthly 1 km NDVI Time Series Product in China Using Random Forest Methodology. *Remote Sensing*, *15*(13), 3353. https://doi.org/10.3390/rs15133353

Figures and tables references

Figure 1:

Brenck M., Hansjürgens B., Schröter-Schlaack C., Tröger U., Wessner A., Wittmer H. (2021). Gesellschaftliche Wertigkeit des Tempelhofer Feldes – Qualitäten erfassen und sichtbar machen. Helmholtz-Zentrum für Umweltforschung – UFZ, Leipzig.

Figure 2-4: Produced by myself using Python and data from Sentinel 2 and data from DWD for Berlin-Tempelhof (433)

Table 1: Produced by myself using excel