DETECTREE

The impact of trees in cities on the environment and society

REPORT

Lucie Eugenie Trützschler von Falkenstein
Matrikel-Nr.: 1182924

Benjamin Schwarz
Matrikel-Nr.: 11920840

Vincent Betz
Matrikel-Nr.: 11914104

DISC DATA LAB

Introduction

Nowadays, earth systems are altered to the needs of humankind, as our planet is dominated by human influence. At the same time, however, while the world seeks to slow down the pace of climate change, tree cover in cities is shrinking rapidly, leading to an increase in urban temperatures and air pollution.

It is certain that climate change is currently one of the greatest ecological, political, economic, and sociological problems facing humanity. As of today, 193 countries have committed to help stop climate change through the Paris Agreement, the treaty calling for measures to reduce emissions [1]. These measurements are to be increased over time [1]. Among all various actions that have the possibility of contributing to emission reduction, the expansion of forest areas has shown to be an effective means[2]. Especially in cities, trees and natural ecosystems help to tackle problems caused by pollutant emissions [3].

In addition, as air pollution is linked to a variety of diseases in humans and the majority of the world's population lives in cities, trees are essential for healthy communities. On top of that, it is shown that trees encourage physical activity, promote social ties and are believed to reduce stress among people [4].

In general, pollution removal (O_3, PM_{10}, NO_2, SO_2, CO) is only one of many positive effects of trees on air quality. Trees in cities also help to reduce air temperatures through transpiration, and reduce building energy and consequent power plant emissions [5]. Moreover, trees in urban spaces provide biodiversity and are thus integral to the environmental quality of cities all around the world [6].

Since, from an economic and political point of view, there are different models on how the CO2 emissions of organizations can be fairly compensated, a foundation for measuring the efficiency of different compensation methods was established. One of these efficient CO2 compensation measures is the planting of new trees, or the maintenance of existing ones[7].

In short, there is no doubt about the indispensable necessity of trees in cities, be it for the people, the environment or biodiversity. However, measurements for increasing tree canopies in cities are rarely performed [8]. Conversely, lots of trees in cities are being cut down to make space for new buildings or parking areas. In the United States, approximately 36 million trees per year were cut down, up until 2018 [9], and it has not gotten much better since that time [10].

To be able to determine the tree share of a city, usually, LIDAR (Light Detection and Ranging) data is used. LIDAR is a technology used to create high resolution models, which include elevation. However, to obtain such a dataset, it takes expensive equipment, and a good amount of workload. In addition, LIDAR data are rarely made openly available. Consequently, LIDAR data is hard to come by, which complicates the study of trees in cities even more.

In this study, we were able to find a more accessible way for calculating the tree shares of a city, called DetecTree [11]. DetecTree is an open source pythonic library that can perform a binary classification of tree/ non-tree pixels from aerial imagery. Using the DetecTree library, we were able to calculate city's tree shares which does not involve too high costs in terms of money. `

As the argument of climate change does not seem to be motivating enough for taking serious action against the decline of tree canopy in cities, and human health is in the interest of us all, we aimed to assess the benefits of trees in cities on people's health and social well-being, with the use of the DetecTree library.

2. Data

2. 1. Satellite Imagery – data description

To analyze tree shares of cities, complete aerial imagery from each city was needed, offering cloudless coverage of all cities to be analyzed. The imagery was obliged to suit the needs of our Pythonic tree-detecting library, DetecTree [11]. The functions from the Detectree library take small tiles as an argument (approximately 875x600 pixels), ideally with the highest resolution possible. This end, two options were given: Either rescale larger tiles by dividing them into smaller pieces, as was done in our first try-out binary tree/non-tree classification for the city Innsbruck (demonstrated in the notebook ibk_tiles on GitHub), for which there is a complete set of GeoTIFF satellite images openly available [12]. TIFF is an image file format that stores raster graphics. GeoTIFF-files allow for additional georeferencing information within a TIFF-file. Unfortunately, however, GeoTIFF imagery for cities other than Innsbruck is scarce and hard to come by. Considering that a minimum of 10 samples (in our case 10 cities) is needed to get significant results in a subsequent regression analysis, and life quality data is limited to a certain number of cities as well, an approach following such a method was not feasible.

The second option was to download small subtiles directly. The Mapbox API allows for programmatic access of satellite data [13]. Mapbox tiles are returned as raster tiles, containing the pixel-based data stored as a grid structure in jpg format with 90% quality maximum. Although not ideal in terms of georeferencing and resolution, in this study, the Mapbox Raster Tiles API was used for further analysis. This approach allowed for availability of satellite imagery of any given area in the world, permitting the analysis of any arbitrary city.

Figure 1: Random tile from Mapbox Satellite [5]. API request: https://api.mapbox.com/v4/mapbox.satellite/16/34869/22726@2x.jpg90?access_token={token}

All Mapbox - tiles were obtained with a zoom level of 16, to have a sufficient resolution for subsequent tree detection. Each tile has a dimension of 512x512 pixels and was to be requested separately. Mapbox tiles were to be indexed as described in the Slippy Map Tilenames specification [14], where the entire topology of the world is transformed into a submanifold, i.e. a 2-D representation of smaller subtiles, depending on the zoom level. The higher the level of the zoom specified, the higher the number of tiles the world is divided into, starting with 1 tile to cover the whole world at zoom level 1. At a zoom level of 16 the world is divided into 4 295 million subtiles. The Slippy Map then lets you choose an arbitrary 512x512 pixel tile, at any given longitude and latitude [Figure 1]. The tiles are not indexed by longitude and latitude as one would expect but are to be described by a special file naming convention due to the ellipsoid shape of the earth. For each separate zoom level, while the longitude (width) in degrees remains constant, the latitude (height) will change as you move upwards throughout the poles [14]. As a result, the longitude gets described by:

$$x = \left\lfloor \frac{lon + 180}{360} \cdot 2^z \right\rfloor$$

and the latitude will be deviated as:

$$y = \left\lfloor \left(1 - \frac{\ln\left(\tan\left(lat \cdot \frac{\pi}{180}\right) + \frac{1}{\cos\left(lat \cdot \frac{\pi}{180}\right)}\right)}{\pi} \right) \cdot 2^{z-1} \right\rfloor$$

In addition to the special file name convention, the Mapbox Satellite API request takes 2 more arguments: the format of the returned tile and the access token. For our analysis, we chose the highest quality jpg image request (@2x.jpg90), returning a 90% quality jpg image of a satellite raster tile with the highest possible pixel density. The access token can be easily received by making an account. Finally, the Mapbox Satellite API request for 1 raster tile would look like this:

https://api.mapbox.com/v4/mapbox.satellite/{tile}@2x.jpg90?access_token={access_token}

2. 2. Detectree

To classify trees in the downloaded (Mapbox-) satellite images, it is not sufficient to read out the respective pixel value, i.e. the color, to solve the binary classification problem. Instead, one must refer to a 27-dimensional feature vector. These pixel features are based on the methods of Yang et al [15]. Here, each pixel is converted into the feature vector, with 6, 18, and 3 features capturing the properties of color, texture, and entropy, respectively. The binary classification at the pixel level between tree-like and non-tree-like pixels is taken by the Detectree module with the pixel_features.PixelFeaturesBuilder() function. The output of the function is a matrix in which each row represents a pixel of the original RGB image, with the respective features of the pixel added as a column.

The structure and in other words the content or the high-level semantics of an orthophoto can be represented by means of Gist descriptors [16] as described by Yang et al. [15]. The exact computation of image properties is based on mathematical convolution with Gabor filters on 3 frequencies and 4, 8, and orientations, respectively, as well as the computation of a common 8x8x8 color histogram in Lab color space. Using the Detectree library, the image property can be computed using the compute_image_descriptor_from_filepath function.

The task of classifying tree/non-tree pixels becomes a supervised learning problem, where a classifier that maps the pixel features to the tree/non-tree responses is trained and later used to classify the values for the remaining pixels [16].

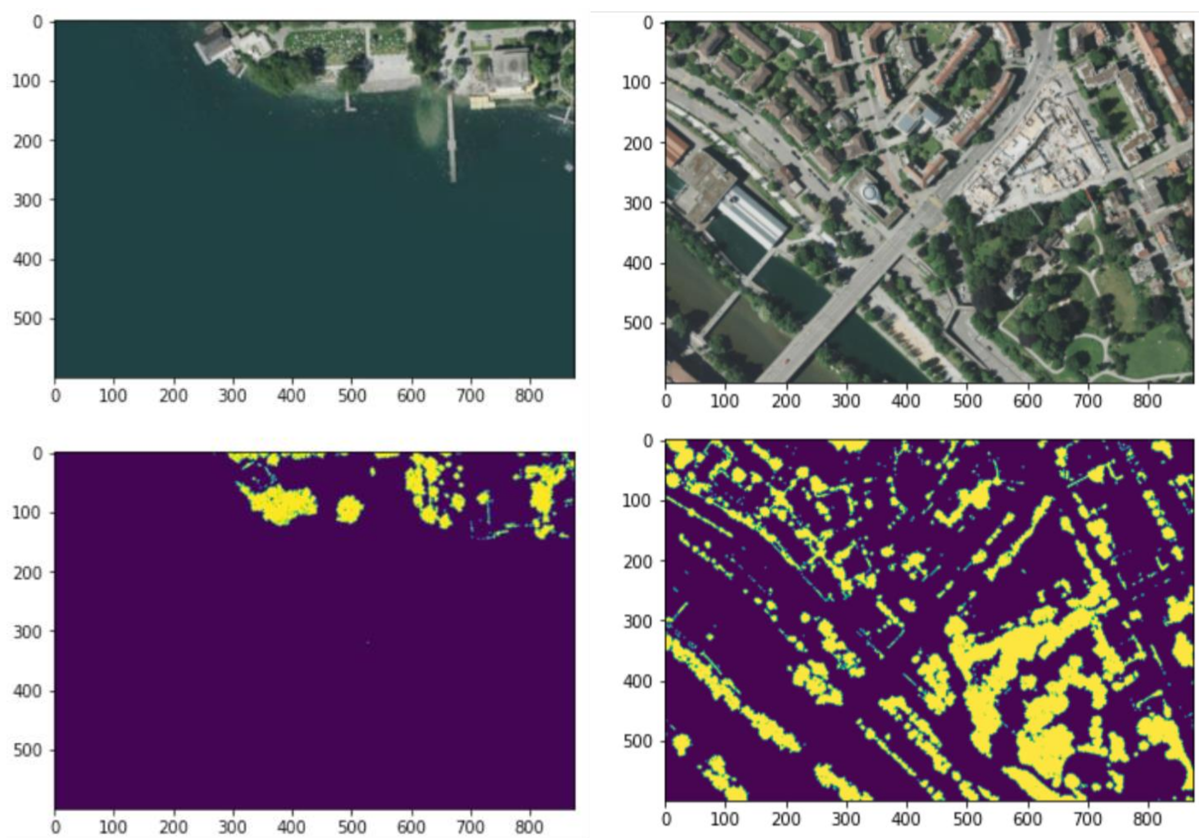


Figure 2: The response tiles provide ground truth masks for the training tiles

Before the actual classification takes place, the system must be trained on suitable data. For this purpose, orthophotos of the city of Zurich were used in this study, since this city also provided binary LIDAR image data that represents the ground truth of the tree/non-tree masks [17]. The training data could be randomly selected as a subset of the original photos, but the individual tiles differ greatly in terms of the imaged surface and may not have any trees at all. This problem was also considered previously in Yang et al [15], where a training pattern selection procedure was proposed that aims to find the set of tiles that is most representative of the dataset. For this purpose, we used the TrainingSelector function of the Detectree library, which, depending on the mode, divides the images into the subsets of test and 1% training data. In this study, the cluster-II method of Yang et al [15] was used for this purpose. The cluster-II method follows the following procedure. First, the image features are computed for each tile of the dataset. Then, one applies k-means to the image properties to divide the dataset into 4 first level clusters. These then differ in terms of their surface properties. In the next step, within each cluster of the first level, another k-means is applied to the image properties. In the final step, those tiles of each first-level cluster that are closest to the centroid of that cluster are then selected to train a separate classifier for the tiles in that first-level cluster. To handle the high computational effort, PCA is used to reduce the dimensionality of the image features to a desired level in our case to 24-dimensions.

The respective tiles selected for training must now be linked to the ground truth of the tree/non-tree masks previously loaded.

Considering the split between training and test data, four AdaBoost classifiers will be trained when using the Cluster II method, one for each first-level tile cluster. Training an AdaBoost ensemble is a computationally intensive task that is performed in the train_classifiers method of the ClassifierTrainer class. The individual trained four AdaBoost classifiers can then be saved as a joblib file (see models directory). From this point on, these classifiers can be used to distinguish tree-like and non-tree-like pixels on previously unseen images. However, one should carefully check whether the new image data also come from the same source and have the same properties regarding resolution, color, and other parameters. This is the only way to guarantee that the classifiers are trained on the model for which they are supposed to make predictions later. In our first step, we did not train the classifier using the same data source. However, similar image quality leads to significantly worse prediction accuracy.
So, the original "raw_tiles" of each city are imported and successively predicted using the classifiers and then stored as "predicted_tiles".

Evaluation and validation:
We can determine the proportion of predicted tree-like and non-tree-like pixels of each image using our program pixel_color.py. A list of the "predicted_tiles" from the corresponding directory is created and in the next step the proportion of tree-like pixels is determined. After running the classification function, the image is further refined to neglect weakly detected tree-like pixels. The tree-like pixels are then stored as white pixels, while the non-tree-like pixels are stored as black ones. Using an analysis of the colors in the image, the proportion of classified pixels can be determined.

To provide an accuracy of the algorithm, we divided the "predicted_tiles" in the program according to high and low tree coverage, respectively, and later randomly selected two tiles from each group. These tiles were then compared to the tree coverage we manually determined to maintain the accuracy of the classification. Since the conditions and the characteristics of the images from discrete cities differ, we performed this validation for multiple cities [Figure 3].

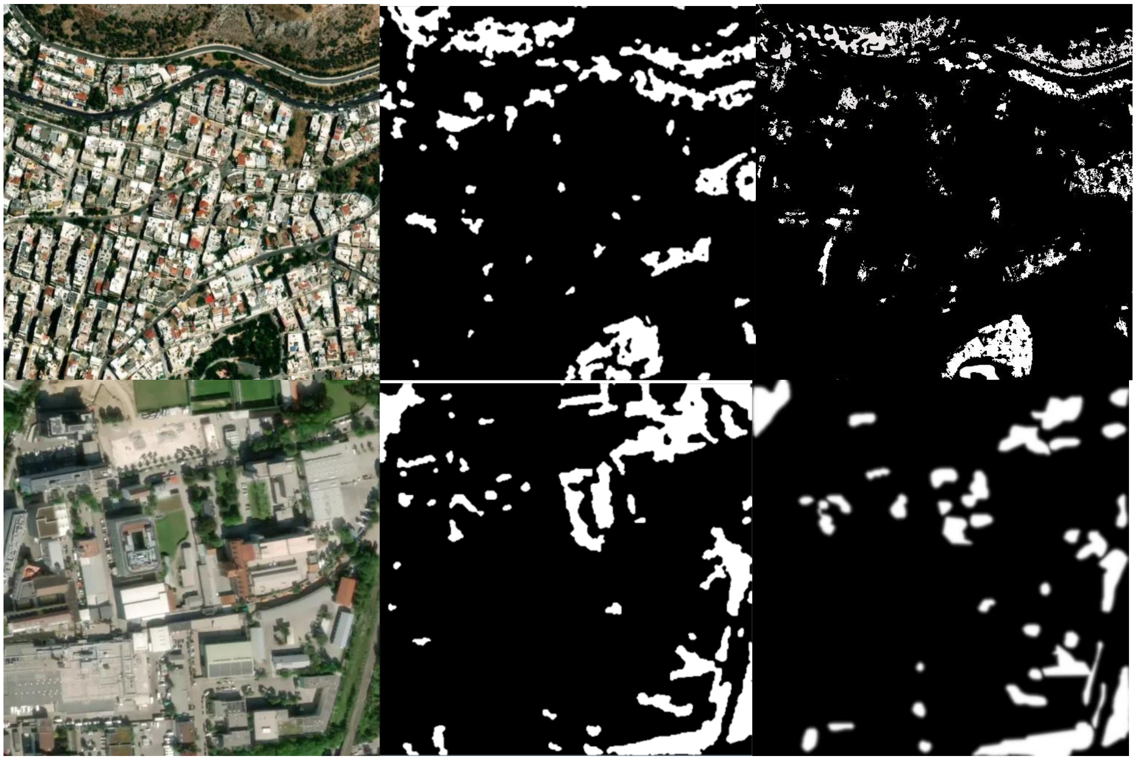


Figure 3: Classification accuracy of satellite image tiles by Detectree. From left to right: Original orthophoto from Mapbox, predicted tree like pixels in white, manually determined tree like pixel in white.

To create a ground truth mask manually, we use the image editing program GIMP. We select all recognizable trees and color them white, while the rest is colored black. With this method we create a mask from which we can determine the percentage of trees which is the white area [Figure 3, 3rd image from left]. To determine the average accuracy, we calculate the ratio of the tree share between the classification algorithm as well as the manually determined tree share. The mean accuracy of the classification algorithm is 75.28% in the analyzed urban areas. Finally, with this relatively high accuracy, we can determine the tree coverage of any city. However, it should be noted that the accuracy can be further improved if, for training the classifier, the LIDAR generated ground truth masks were taken at the same time as the orthophotos and if the images were provided at a higher resolution. Especially the shadows cast by houses and trees increase the tree coverage determined by the algorithm compared to the actual value. However, one must keep in mind that the false detection of tree-like pixels is, to a good approximation, a systematic error that occurs in all cities in a similar order of magnitude depending on individual parameters.

2. 3. External data sets

To compare data on the proportion of trees in cities with other variables, data from external studies were used. In order to better understand the impact of urban forestation on social and community factors, data from surveys and studies on these topics were applied. City-data on pollution, quality of life, cost of living, purchasing power, citizen satisfaction, health care, and property price-to-income ratios were collected. Much of the data for the first analysis comes from the Numbeo.com website and, according to the company, is not subject to influence by government organizations [17]. Numbeo's statistics and surveys have already been cited by many well-known international magazines and appear to be reputable enough to perform truthful analyses. The values used for the analysis were available for the year 2022 and thus matched the time period of our analysed satellite imagery. The data were reported as indexes and calculated in different ways. It is often a composite of citizen surveys and data from professional organizations. In other cases, the data is available from governments or city administrations. The quality-of-life index is an aggregation of all variables of a city and thus gives an overall view of the living conditions. Included in its calculation were "purchasing power index (the higher the better), pollution index (the lower the better), housing cost to income ratio (the lower the better), cost of living index (the lower the better), safety index (the higher the better), health care index (the higher the better), commute time index (the higher the better), and climate index (the higher the better)" [18]. A report on the quality of life in European cities from the European Commission was included as an additional data source for the variable "citizen satisfaction". The percentage of citizens who are satisfied living in the respective city was determined [19].

For the second analysis, a dataset on the proportion of trees in cities was used. Since our first analysis could not include so many samples, we chose to study the dataset of 78 analysable cities from the European Environment Agency [20]. In this dataset, the proportion of trees in cities was determined based on an analysis of land atlases. The area used as the basis for a city's analysis is very important to the percentage of trees and can lead to large differences in small changes. It should be noted that for this external dataset, the cities were partially delimited over a larger area than in our own detectree procedure. Since the cities often have a higher tree density at the edges, the tree fractions of the external data set are higher. Data from this dataset were available for 2021, so we used the Numbeo dataset from 2021 for the second analysis. The "citizen satisfaction" variable was dropped in the second analysis and replaced by the "health care index."

For both the first and second analyses, data from the Numbeo dataset were downloaded as a CSV file. The tree share data were entered manually into the CSV file as this was not possible otherwise due to the formats. The data was now in the format to start the analysis in python.

3. Data Analysis

3. 1. Data analysis with detectree data set

The tree share percentages obtained by detectree are combined with the quality of life data in one data set. As an additional binary variable "high_l" is integrated, which is output as "low" for a tree share below 25% and as "high" for a tree share above 25%. All values are in a width between 0 and 200 but are not normalized initially to maintain clarity.

Variable	Level	Scope
tree	numerical	1-100 (percent)
pollution	numerical	0-200 (Index)
life_quality	numerical	0-200 (Index)
living_costs	numerical	0-200 (Index)
PPI	numerical	0-200 (Index)
satisfaction	numerical	1-100 (percent)
property_income	numerical	0-200 (Index)
high_l	binary	“high” or “low” tree share
city	categorical	City name

The binary variable "high_l" that divides the data into weakly or heavily forested cities was **box plotted** at the beginning with each variable. This should give a first impression how far apart the median and the interquantile range (IQR) are for the categories "low" and "high" for the respective variables. The diagrams generated with the package matplotlib.pyplot should illustrate first rough differences between the categories.

In order to see possible strong differences of the two expressions "high" and "low" tree share in the average values for the different variables a table was generated for the respective **means**. These initial analyses should help identify for which variables a difference in tree cover takes on differentiated characteristics.

On all numerical variables from the data set, **min-max normalization** in the range 0 to 1 were applied before performing linear regression. The normalization was chosen to prevent possible bias in the results for the coefficients and to better identify a positive or negative influence of a variable. The normalization was performed with the support of the python numpy package. For **linear regression**, we use the Ordinary least squares function from the statsmodels library to obtain the best regression equation for our data. We specify our data set and choose "tree" as the dependent variable. The explanatory variables are "pollution", "life_quality", "living_costs", "PPI", and "satisfaction", as these gave the most promising results from the initial exploratory analyses. Even though it is probably not possible to literally say that these variables have an influence on the number of trees in cities, the regression equation is chosen around them on purpose in order to obtain results that deal with the variable of most interest, tree share. In the first line of code, we create the model and in the second line we search for the best fitting model. Afterwards we output the summary of our regression model.

```
model = smf.ols("tree ~ pollution + life_quality + living_costs + PPI + satisfaction", data=df)
m = model.fit()
print(m.summary())
```

It must be added that we used only n=12 samples and therefore the results should be considered with caution and may not be generally valid. The analysis is intended to give a first indication of whether the variables have any “influence” on the tree share in cities (or the other way around).

In the next step, a second additional regression model was created in the same way. This time with "pollution" as the dependent variable and "tree" share as the explanatory variable. This was done to identify the influence that trees in cities have on air pollution, as this is of particular importance politically and socially. In addition, the regression model was then plotted using the matplotlib library to graphically display a **regression line**.

In addition, the **Pearson correlation coefficient** between "pollution" and "tree" was calculated to look more closely at the correlation between the two variables. With the help of the seaborn library a **correlation matrix** for all numerical variables was created to get additional information about the correlations.

3.2. Data analysis with European environment agency data set

A second analysis was performed with the European environment agency dataset to perform a linear regression and correlation matrix with more samples. The changed variables "health_care" instead of "satisfaction" and without the binary variable "high_l" can be read in the table below. The city names were omitted in this data set.

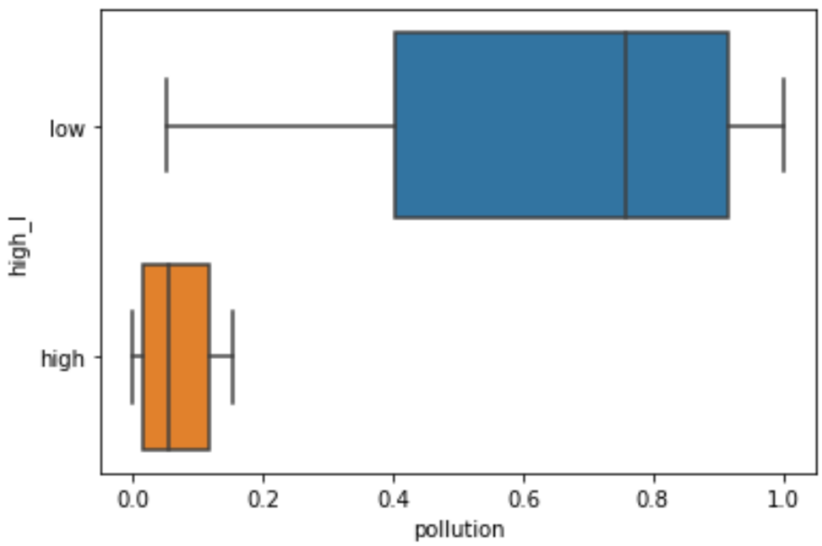
Variable	Level	Scope
tree	numerical	1-100 (percent)
pollution	numerical	0-200 (Index)
life_quality	numerical	0-200 (Index)
cost_of_living	numerical	0-200 (Index)
PPI	numerical	0-200 (Index)
Health_care	numerical	0-200 (Index)
property_income	numerical	0-200 (Index)

The numerical variables were min-max normalized between 0 and 1. Again, the ordinary least squares function from the statsmodels library was used to set up a regression model. The explanatory variables are "pollution", "life_quality", "cost_of_living", "PPI", and "health_care" and the dependent variable is "tree". In the following, as in the first analysis, a summary of the regression model was created and a regression line with "pollution" as the intercept and "tree" as the explanatory variable was plotted. A correlation matrix was then calculated to analyze the correlations between the variables based on a more valid data set.

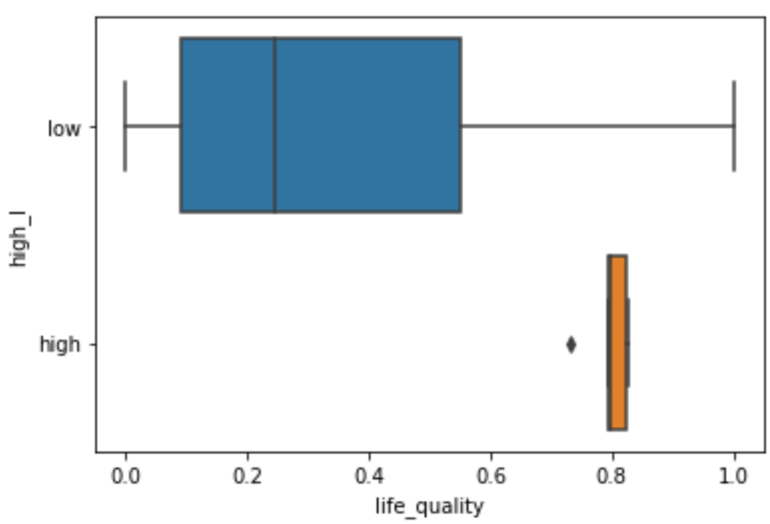
4. Results

4.1 Results from Analysis 1

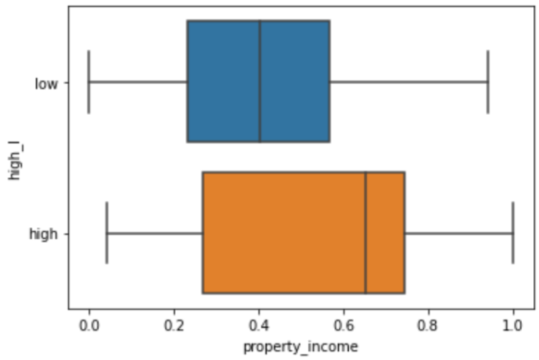
We see that for the pair of variables we are particularly interested in, pollution and tree fraction, the median and all quartiles of the boxplot for the two categories are very far apart. For the cities classified in the category "high" tree share, all quartiles for pollution are much lower than for the cities classified in the category "low" tree share. Leaving aside the small number of samples in the categories, we can assume through this initial visuaization that a tree percentage below 25% is an indicator of greater pollution in the city.

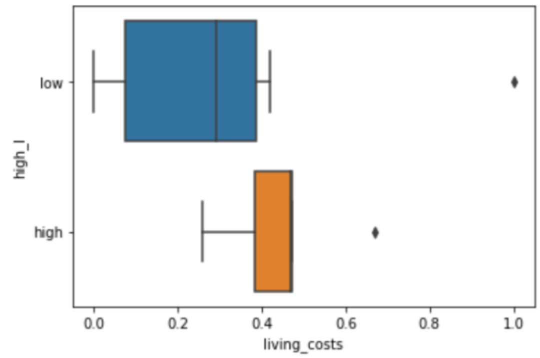


Quality of life appears to be very close for the cities classified as "high". Again, both the median and the inner quartiles are in a higher range as far as the quality-of-life index is concerned. This could indicate that the quality of life is influenced by the proportion of trees in cities lying above or below 25%.

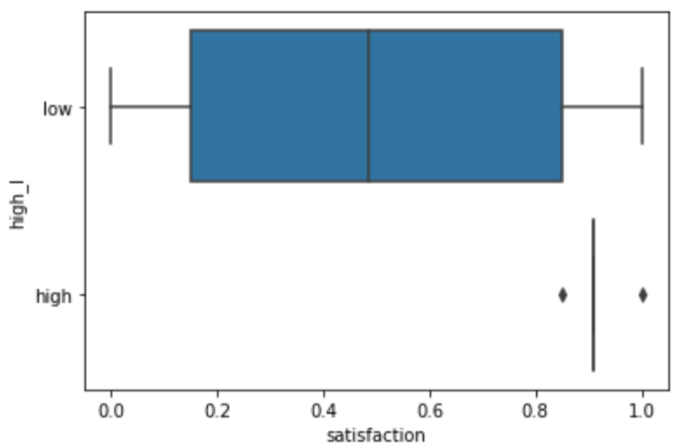


However, in the area of cost of living as well as real estate income index, the high forest cities seem to perform "worse". A tree coverage ratio above 25% leads to a higher median housing cost to income ratio and also to a higher median cost of living. This could be due to the fact that more ecologically friendly cities are more expensive in their prices for existentially important goods.





Satisfaction is higher in the median and inner quartiles for cities with "high" tree ratios and appears to be distributed in a similar range for these cities as well. In cities with "low" tree ratios, satisfaction among citizens is highly dispersed. It is difficult to interpret from this plot whether the degree of tree coverage really directly affects resident satisfaction. Firstly, the satisfaction of the citizens is possibly strongly co-determined by all the variables already considered, secondly, we use the same data as for these first variables.



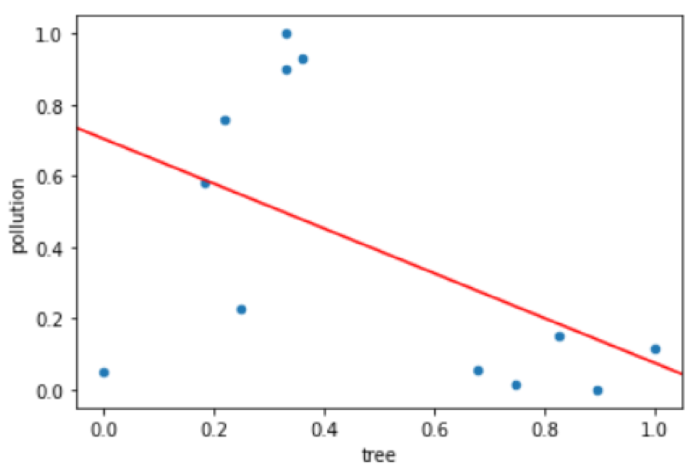
From the average table, similar slopes can be seen for all variables as from the boxplots, although the real estate income ratio is almost the same and the range in pollution is again clearly far apart. Cities with less than 25% tree coverage have, on average, a pollution index more than three times (3.20) as high as cities with higher tree coverage.

high_l	tree	pollution	life_quality	living_costs	PPI	satisfaction	property_income
high	43.98	18.8	178.24	75.28	79.68	94.20	12.14
low	15.82	61.0	130.37	61.46	62.72	80.43	11.05

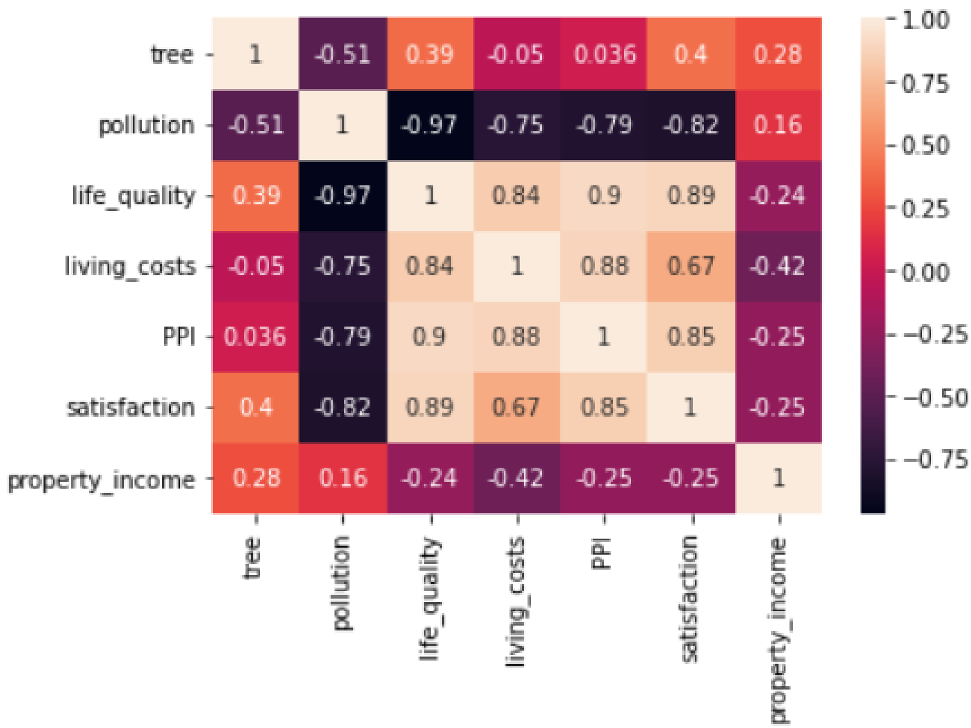
The **regression model** explains 82.0% of the change in our tree variable (**r-squared**) and the **adjusted r-squared** is still 66.9%, which could iniciate an “influence” of the explanatory variables on the dependent variable tree proportion. However, the coefficients are all non-significant below a 10% significance level, which makes the analysis **invalid**. The output variable Prob (F-statistic) also tells us that the effect of the chosen variblen is zero with a probability of 3.10%. Nevertheless, it should be noted that this regression model indicates that a one unit change in "pollution" results in a unit reduction of -0.2026 in tree cover.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   tree   R-squared:                       0.820
Model:                            OLS   Adj. R-squared:                  0.669
Method:                 Least Squares   F-statistic:                     5.455
Date:                Mon, 20 Jun 2022   Prob (F-statistic):             0.0310
Time:                        22:08:44   Log-Likelihood:                 7.2023
No. Observations:                  12   AIC:                            -2.405
Df Residuals:                       6   BIC:                            0.5048
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        0.3924      0.932      0.421      0.688      -1.889       2.673
pollution       -0.2026      0.939     -0.216      0.836      -2.501       2.095
life_quality     1.2819      1.642      0.781      0.465      -2.735       5.299
living_costs    -0.6175      0.532     -1.161      0.290      -1.919       0.684
PPI             -1.3895      0.727     -1.911      0.105      -3.169       0.390
satisfaction     0.3768      0.425      0.887      0.409      -0.663       1.417
==============================================================================
Omnibus:                        0.739   Durbin-Watson:                   2.120
Prob(Omnibus):                  0.691   Jarque-Bera (JB):                0.684
Skew:                           0.435   Prob(JB):                        0.710
Kurtosis:                       2.217   Cond. No.                         60.1
==============================================================================
```

The second regression model also might not be valid with a Prob (F-Statistic) of 8.92% and no p-value for the coefficients under 10%. Nevertheless, the regression line with pollution as the dependent variable and tree as the explanatory variable is interesting to look at because again a negative slope (tree coefficient: -0.6285) can be observed. This would mean that higher tree cover would bring about less pollution in cities.



Looking at the correlation matrix, we notice that tree cover is strongly negatively correlated with pollution (**pearson**: *-0.51*; **p-value**: 0.0891), and positively correlated with quality of life (**pearson**: *0.39*; **p-value**: 0.2134) and satisfaction (*0.40*), but not so clearly correlated with the other variables. However, the pollution index correlates very strongly with the satisfaction (-*0.82*), PPI (-*0.79*), cost of living (-*0.75*), and quality of life (**pearson**: *-0.97*; **p-value**: 0.0000) variables. The correlation matrix suggests that urban pollution is strongly correlated with many quality-of-life characteristics. Comparing these results with our initial boxplots, it can be seen that low tree cover in cities as well as high pollution levels may be related to lower cost of living but also lower purchasing power, quality of life and citizen satisfaction.



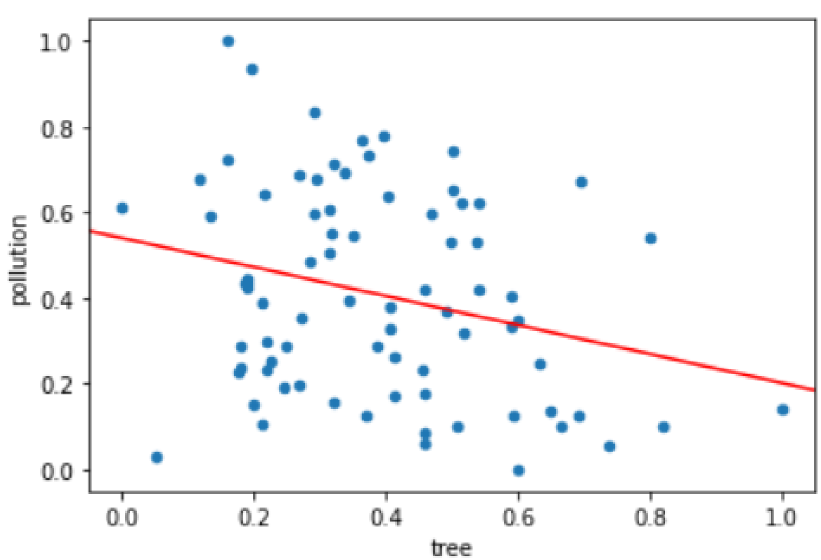
4.2 Results from Analysis 2

In the second analysis, **linear regression** was started immediately to obtain more valid coefficients. The first regression model of the second analysis explains only 8.5% of the change in the dependent variable tree proportion and the **adjusted r-squared** is only 2.2%. It should be noted that with other models and fewer variables, no more valid regression models could be found to explain a relationship with the tree proportion. The output variable Prob (F-statistic) tells us that the effect of the selected variable is zero with a probability of 25.5%. The model is too invalid to draw conclusions from the coefficients.

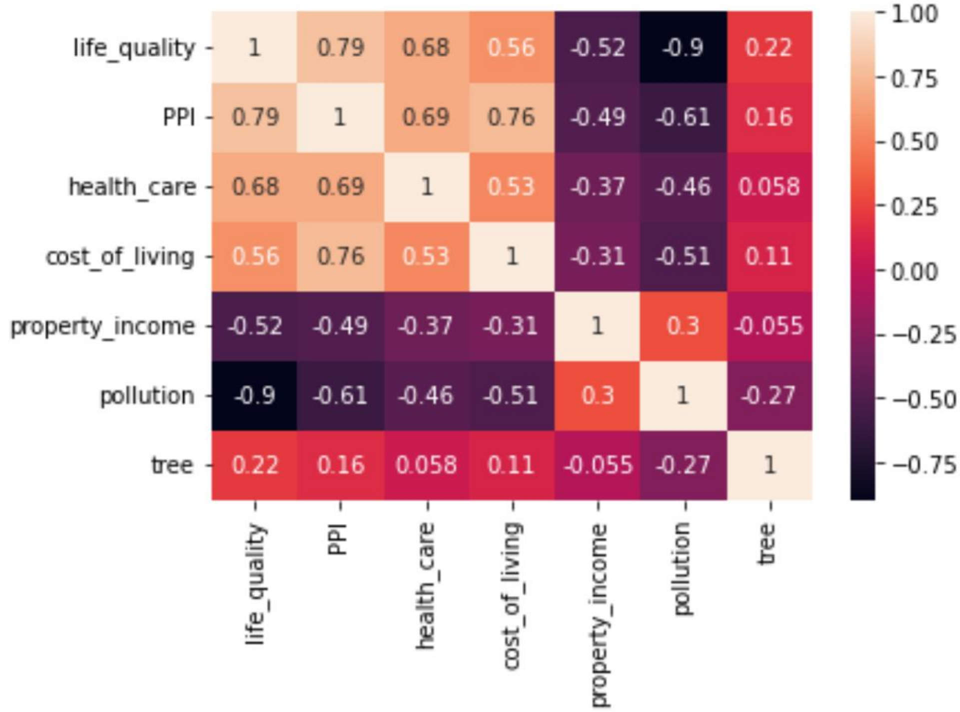
For the second regression model with "pollution" as the dependent variable and "tree" as the only explanatory variable, more valid values can be read. The second regression model again explains only 7.3% of the change in the dependent variable tree proportion and the adjusted r-squared is only 6.1%. However, this time the output variable **Prob (F-statistic)** tells us that the effect of the selected variable is zero with a probability of 1.64%. The p-value of the coefficient "tree" is 0.016, which is below a significance level of 5%. One unit of the variable tree share decreases the dependent variable pollution by 0.338 units, there is a **negative relationship** between the two variables.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              pollution   R-squared:                       0.073
Model:                            OLS   Adj. R-squared:                  0.061
Method:                 Least Squares   F-statistic:                     6.020
Date:                Mon, 20 Jun 2022   Prob (F-statistic):             0.0164
Time:                        18:09:31   Log-Likelihood:                 4.0489
No. Observations:                  78   AIC:                            -4.098
Df Residuals:                      76   BIC:                            0.6156
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5396      0.059      9.070      0.000       0.421       0.658
tree          -0.3380      0.138     -2.453      0.016      -0.612      -0.064
==============================================================================
Omnibus:                        7.628   Durbin-Watson:                   0.426
Prob(Omnibus):                  0.022   Jarque-Bera (JB):                3.258
Skew:                           0.200   Prob(JB):                        0.196
Kurtosis:                       2.082   Cond. No.                         6.04
==============================================================================
```

The regression line clearly shows the negative correlation between "tree" and "pollution". This regression model, this time built with a larger external data set, confirms our assumption that there may be a negative relationship between tree cover and pollution in a city.



The variables "pollution" and "tree" are negatively correlated (pearson: *-0.27*; p-value: 0.0164). Looking at the correlation matrix, we notice that pollution and quality of life are again strongly negatively correlated (*-0.90*). The correlations are not as strong as in our first data set with fewer variables, but the values are more valid and still have lower p-values. Similar to the first analysis, the variable "pollution" has a strong negative correlation with several variables related to the quality of life in cities.



4.3 Results Conclusion

The first analysis with the dataset containing only 12 samples resulted in few valid analyses, but a direction of the relationships could be identified, especially between the percentage of trees, pollution and the quality of life in cities. While in the first analysis partly very extreme correlations between these three variables were calculated, the results in the second analysis with more samples were less extreme but yielded the same correlations. It can be concluded from these analyses that the proportion of trees in cities is related to lower pollution and better quality of life. The exact way in which these factors influence each other should be worked out in further research based on this work’s foundation.

5. Discussion about implications and implementation

5.1 Determination of the value for climate offset measures

Offsetting one's own carbon footprint is not only done by voluntary climate-protecting private individuals, the protection of the climate and the "reparation" of what one has done to the environment is also coming more and more into the focus of companies. A large number of companies that continue to emit many emissions can still call themselves carbon neutral. "Only a few of them have achieved their climate neutrality by completely reducing their emissions. Most are made climate neutral by offsetting [...] the 'unavoidable emissions.'" [21]. This works by organizations running climate protection projects and selling the carbon saved from the air through these projects in the form of certificates. For example, if a company emits 30 tons of CO_2 per year, it can buy CO_2 certificates worth 30 tons of carbon and call itself climate neutral. The price of these certificates is then about the same as the costs that were necessary to reduce 30 tons of CO_2, which might have been for planting trees or developing carbon storage.

So far, this neutralization has only been carried out on a voluntary basis in most industries. However, the climate neutrality of a company is of great importance to some shareholders, and this is becoming more and more important in Europe. Investors see the long-term imperative for a company to remain carbon neutral, as stricter regulations and growing pressure from society will make this unavoidable. Also, an increasing number of customers pay attention to a company's carbon footprint before buying their products or services. One reason for this attention is that we are all threatened by climate change in our way of life, but some contribute more and some less to this outcome. The pressure that those who do the damage to society must also fix it or pay for it keeps growing. There must be a social balance and a company can accomplish this through compensatory measures. In addition, many climate projects in the EU are implemented by governments using carbon tax funds to re-balance the pollution of the contributors. To achieve this social balance efficiently and as accurately as possible, however, the effects of both emission discharges and countermeasures must first be known. In order for the damage to be paid fairly, the state, society and companies must know what damage they are causing and what repairs and societal gains they are achieving through certain measures. This work was conducted to provide guidance for the latter.

5.2 The need for measurability of social benefits from urban forestation.

By determining the influence of trees in cities on social factors, a metric can be developed that evaluates measures for tree planting in urban areas. The damage caused by emissions to society is roughly known since the effects are familiar (at least at present). To measure the real costs of a contribution to climate change for the whole society, of course, future consequences have to be taken into account. Such calculations are already being made by governments and NGOs but will not be discussed further in this paper. The question of how much it can be worth to society to expand green spaces in cities should be answered in part.

Among the various actions taken to reduce atmospheric pollutants, the planting of trees is a major element. “EU forests, for example, absorb the equivalent of nearly 10% of total EU greenhouse gas emissions each year.” [22]. Especially in cities, where more pollutants accumulate in the air but there are few areas and difficult circumstances for planting trees, the creation of new tree areas can bring enormous benefits. At the 26th UN Climate Change Conference "COP26" in Glasgow, global necessities for combating climate change were discussed. One of the topics discussed was the increased use of green spaces and trees in cities. The world economic forum, for instance, attributes many functions to trees in cities: "They sequester carbon, but they also provide support to health and mental wellbeing services, around schools they dampen noise, and improve educational attainment, provide a sustainable building material, and for finance and insurance enthusiasts, trees are a natural hedge, providing stability and reducing flood risk induced by extreme weather.” [23].

If it could now be determined more precisely to what extent these benefits are caused by planting trees in cities, then a value could be attached to this measure. If such a benchmark exists, planting projects in cities can be carried out by climate protection organizations or companies themselves, and certificates can be issued. The value of a climate protection project has so far mainly been calculated in such a way that the resulting certificates receive the differential value of the project. The differential value results from the pollution level that would have arisen or not been cleaned up without the project and the lower pollution level that results from the climate project. What has been missing so far in the calculation of the value of a sustainability project are its additional benefits for society. These social gains are often important benefits that accrue in addition to carbon reductions. Governments and climate protection programs should include these "co-benefits" in the calculation of the value of a certificate [24]. “In many development contexts, and in many specific communities, these co-benefits are concrete and near-term, and are often seen as more directly valuable than carbon benefits.” [24]. One obstacle due to which the positive social side effects of sustainability projects are not attributed to their value is the difficulty of measuring such effects. In this work, smaller samples were used to draw attention to the fact that trees in cities do exert a positive social influence on various factors. By determining the percentage of trees in different cities, comparisons can be made and the influence of a lower or higher percentage of trees can be estimated. The work can be built upon by including larger samples and more detailed data on social aspects to determine an accurate value for the climate protection project "Reforestation of cities". This will help both governments and companies to more precisely regulate the carbon market and initiate a stronger positive impact on society. Data-driven analysis of environmental and social effects will be necessary to guide the future market. Satellite imagery processing and analysis will be an important tool for large-scale studies.

6. Appendix

Git repository: detectree.git

Detectree - Data aggregation:

1. convert_coordinates.py (get the relevant tiles)
2. mapbox_city.py (download the relevant tiles from docs.mapbox.com)

Detectree - Pixel Classification:

1. train_classifier.py (train the classifier for tree like pixel recognition)
2. Ibk_detectree.ipynb (Notebook to classify tree like pixels for Innsbruck)
3. Detectree_skopje.ipynb (Notebook to classify tree like pixels for Skopje for further information)
4. models (Folder with the already trained AdaBoost classifier)

Detectree - Pixel Analysis:

1. pixel-color.py (Analyze the tree share in the classified tiles)

Data analysis:

1. Analysis_data_lab_1.ipynb (entire analysis 1)
2. first_data.csv (Dataset for analysis 1)
3. Analysis_data_lab_2.ipynb (entire analysis 2)
4. second_data.csv (Dataset for analysis 2)

Results:

1. Analysis_data_lab_1.ipynb (Results Analysis 1)
2. Analysis_data_lab_2.ipynb (Results Analysis 2)

References:

[1] *Untc*. (2016, November 4). Retrieved June 14, 2022, from https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-7-d&chapter=27&clang=_en

[2] Raihan, A., Begum, R. A., Mohd Said, M. N., & Abdullah, S. M. (2019). A review of emission reduction potential and cost savings through forest carbon sequestration. *Asian Journal of Water, Environment and Pollution*, *16*(3), 1-7. https://doi.org/10.3233/ajw190027

[3] Nowak, D. J., Crane, D. E., & Stevens, J. C. (2006). Air pollution removal by urban trees and shrubs in the United States. *Urban Forestry & Urban Greening*, *4*(3-4), 115-123. https://doi.org/10.1016/j.ufug.2006.01.007

[4] Jiang, B., Chang, C., & Sullivan, W. C. (2014). A dose of nature: Tree cover, stress reduction, and gender differences. *Landscape and Urban Planning*, *132*, 26-36. https://doi.org/10.1016/j.landurbplan.2014.08.005

[5] Nowak, D. J., Civerolo, K. L., Trivikrama Rao, S., Gopal Sistla, Luley, C. J., & E. Crane, D. (2000). A modeling study of the impact of urban trees on ozone. *Atmospheric Environment*, *34*(10), 1601-1613. https://doi.org/10.1016/s1352-2310(99)00394-5

[6] Bolund, P., & Hunhammar, S. (1999). Ecosystem services in urban areas. *Ecological Economics*, *29*(2), 293-301. https://doi.org/10.1016/s0921-8009(99)00013-0

[7] Schaefer, F., & Blanke, M. (2014). Opportunities and challenges of carbon footprint, climate or CO2 labelling for horticultural products. *Erwerbs-Obstbau*, *56*(2), 73-80. https://doi.org/10.1007/s10341-014-0206-6

[8] Ruberg, S. (2022, April 23). Major cities work to remove barriers for tree canopy coverage in underserved communities. NBC News. Retrieved June 14, 2022, from https://www.nbcnews.com/science/science-news/cities-look-trees-combat-heat-islands-growth-slow-rcna25327

[9] Nowak, D. J., & Greenfield, E. J. (2018). Declining urban and community tree cover in the United States. *Urban Forestry & Urban Greening*, *32*, 32-55. https://doi.org/10.1016/j.ufug.2018.03.006

[10] Ehrenberg, R. (2015). Global forest survey finds trillions of trees. *Nature*. https://doi.org/10.1038/nature.2015.18287

[11] Bosch, M. (2020). DetecTree: Tree detection from aerial imagery in Python. *Journal of Open Source Software*, *5*(50), 2172. https://doi.org/10.21105/joss.02172

[12] Orthofoto Download Tirol (2022), https://tiris.maps.arcgis.com/apps/webappviewer/index.html?id=5849fe1df5994dc8a3c1e4675682d2fd

[12] Mapbox Satellite (2022), https://www.mapbox.com/maps/satellite

[14] Slippy map tilenames (2022), Retrieved June 15, 2022, from https://wiki.openstreetmap.org/wiki/Slippy_map_tilenames

[15] Yang, L., Wu, X., Praun, E., & Ma, X. (2009). Tree detection from aerial imagery. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*. https://doi.org/10.1145/1653771.1653792

[16] Olivia, A., & Torralba, A. (2001). *Geolion*. Retrieved June 14, 2022, from https://www.geolion.zh.ch/geodatensatz/show?gdsid=343

[17] *Europa: Lebensqualität-index 2021*. (2021). Lebenshaltungskosten. Retrieved June 19, 2022, from https://de.numbeo.com/lebensqualit%C3%A4t/rankings-nach-region?title=2021®ion=150

[18] *Europa: Lebensqualität-index 2022*. (2022). Lebenshaltungskosten. Retrieved June 19, 2022, from https://de.numbeo.com/lebensqualit%C3%A4t/rankings-nach-region?title=2022®ion=150

[19] Bolsi, P., Castelli, C., Hombres, B., Dominicis, L., Montalto, V., Pontarollo, N., Poelman, H., & Ackermans, L. (2020). *REPORT ON THE QUALITY OF LIFE IN EUROPEAN CITIES, 2020*. European Commission. https://ec.europa.eu/regional_policy/sources/docgener/work/qol2020/quality_life_european_cities_en.pdf

[20] *Urban tree cover*. (2021, December 1). European Environment Agency. Retrieved June 19, 2022, from https://www.eea.europa.eu/data-and-maps/dashboards/urban-tree-cover

[21] *Freiwilliger Markt fur CO2-kompensation im Umbruch*. (2021, December 7). EY US - Building a better working world. Retrieved June 19, 2022, from https://www.ey.com/de_de/decarbonization/freiwilliger-markt-fuer-co2-kompensation-im-umbruch (Translations were conducted by the authors)

[22] European Commission. (2022). *Forests and agriculture*. Climate Action. Retrieved June 19, 2022, from https://ec.europa.eu/clima/eu-action/forests-and-agriculture_de

[23] World Economic Forum. (2022, January 19). *Turning COP26's energy and enthusiasm into citywide change*. World Economic Forum. Retrieved June 19, 2022, from https://www.weforum.org/agenda/2022/01/turning-cop26-s-energy-and-enthusiasm-into-citywide-change/

[24] Lou, J., Hultman, N., Patwardhan, A., & Qiu, Y. L. (2022). Integrating sustainability into climate finance by quantifying the Co-benefits and market impact of carbon projects. *Communications Earth & Environment*, *3*(1). https://doi.org/10.1038/s43247-022-00468-9

[25] Wirth, S., & Liedke, J. (2022, June 1). *CO2-preis: Was die CO2-steuer fur Verbraucher bedeutet*. Handelsblatt - Nachrichten aus Finanzen, Wirtschaft und Politik. Retrieved June 19, 2022, from https://www.handelsblatt.com/finanzen/steuern-recht/steuern/co2-preis-was-die-co2-steuer-fuer-verbraucher-bedeutet/26228322.html

[26] *CO2-zertifikate: Bund nimmt Milliarden beim Emissionshandel ein*. (2019, December 26). Handelsblatt - Nachrichten aus Finanzen, Wirtschaft und Politik. Retrieved June 19, 2022, from

https://www.handelsblatt.com/politik/deutschland/co2-zertifikate-bund-nimmt-milliarden-beim-emissionshandel-ein/25366938.html