The Influence of Emissions and Vegetation Coverage on Air Quality in Major Chinese Cities

ZENG, Zhi Qi 1155215115
JIANG, Wen Ming 1155215099
QIN, Lang 1155215108

1. Introduction

1.1 Context and Objectives

In China, the air quality problem seems to be taken more and more seriously. The government has been applying intervention to tackle this problem. For example, any cities have banned fireworks and firecrackers in consideration of safety issues and air pollution. Also, in order to solve the problem of air pollution, the government vigorously promotes new energy vehicles. Much of the air quality problem is down to factors such as energy consumption, exhaust emissions and the degree of greening.

Studies have shown that green areas in urban parks have a positive impact on air pollution control, while industrial and vehicular emissions have a negative effect on air pollution [1].

Therefore, we hope to find out the factors that most affect air quality by analyzing the data from the perspective of the influencing factors of air quality problems, so as to provide help for urban planning and policy making.

1.2 Significance and Challenges

Air quality is a globally recognized priority concern. Air pollution not only affects the economy but also have harmful effects on people's lives. For example, air pollution can cause acid rain and haze, which can lead to water and soil hazards. Meanwhile, it can lead to the depletion of the ozone layer, which can induce increasing exposure to ultraviolet radiation and a higher risk of skin cancer [2]. Reduced air quality affects people's health, which in turn affects healthcare costs and the productivity of the labor force.

In the decade from 2014 to 2024, China's air pollution control policies have achieved decent results. Therefore, the study of factors affecting air quality is conducive to better formulation of government policies on optimizing air quality, which is conducive to economic growth, environmental protection and human health and safety.

Studying how these factors individually and collectively influence air quality can provide a scientific basis for policymaking, fostering environmental improvement and sustainable development.

The biggest challenge of these kinds of research are the difficulty of quantifying various indicators. Various emission and vegetarian coverage are many indicators across diverse domains, which contribute to air quality individually. Therefore, we employ scientific methods such as normalization to establish a unified standard for data from different units.

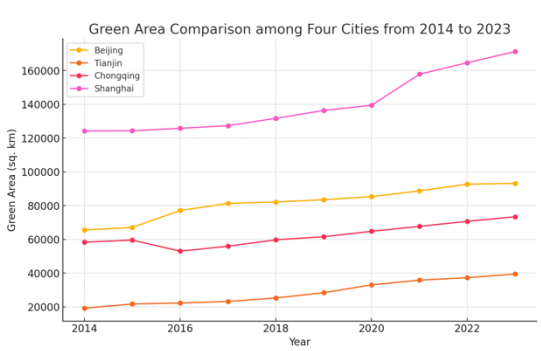
2.Data Description and Visualization

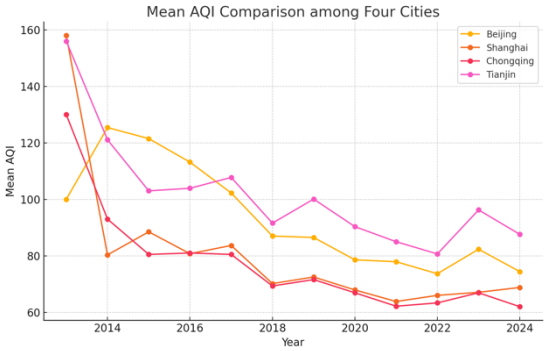
2.1 Data Source

In order to analyze the relevant influencing factors of the AQI, we have selected the AQI data of three cities as samples, namely Beijing, Chongqing and Shanghai [3]. From a website storing historical AQI data, we found the AQI data for those four cities over the past decade. We obtained data on industrial, vehicle and domestic emissions in various Chinese cities from 2014 to 2024 from the official website of the National Bureau of Statistics of China [4]. In addition, we obtained data on the greenness and vegetation coverage of the three cities [5].

2.2 Visualization

Before we delve further into studying this dataset, we first create some line charts to show how the data is spread out, and box plots to illustrate individual cases. This will help us get a clear picture of what's going on before we proceed. Data visualization makes it easier to understand information by presenting it visually.





By comparing the years when significant increases in green area occurred with the corresponding changes in AQI, we might be able to identify any temporal correlations. If improvements in AQI lag behind increases in green area, it could suggest that it takes time for the benefits of increased green spaces to manifest in air quality improvements.

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load the data
beijing_air = pd.read_csv('/path/to/北京.csv', encoding='gb2312')
shanghai_air = pd.read_csv('/path/to/上海.csv', encoding='gb18030')
chongqing_air = pd.read_csv('/path/to/重庆.csv', encoding='gb18030')
tianjin_air = pd.read_csv('/path/to/天津.csv', encoding='gb2312')

# Extract the year from the '月份' column and convert it to a datetime object
beijing_air['Year'] = pd.to_datetime(beijing_air['月份'], format='%b-%y').dt.year
shanghai_air['Year'] = pd.to_datetime(shanghai_air['月份'], format='%b-%y').dt.year
chongqing_air['Year'] = pd.to_datetime(chongqing_air['月份'], format='%b-%y').dt.year
tianjin_air['Year'] = pd.to_datetime(tianjin_air['月份'], format='%b-%y').dt.year

# Group data by Year for each city and calculate the mean AQI
beijing_mean_aqi = beijing_air.groupby('Year')['AQI'].mean()
shanghai_mean_aqi = shanghai_air.groupby('Year')['AQI'].mean()
chongqing_mean_aqi = chongqing_air.groupby('Year')['AQI'].mean()
tianjin_mean_aqi = tianjin_air.groupby('Year')['AQI'].mean()

# Combine the data into one DataFrame for plotting
aqi_data = pd.DataFrame({
    'Beijing': beijing_mean_aqi,
    'Shanghai': shanghai_mean_aqi,
    'Chongqing': chongqing_mean_aqi,
    'Tianjin': tianjin_mean_aqi
}).reset_index()

# Plot the data
plt.figure(figsize=(10, 6))
plt.plot(aqi_data['Year'], aqi_data['Beijing'], marker='o', label='Beijing')
plt.plot(aqi_data['Year'], aqi_data['Shanghai'], marker='o', label='Shanghai')
plt.plot(aqi_data['Year'], aqi_data['Chongqing'], marker='o', label='Chongqing')
plt.plot(aqi_data['Year'], aqi_data['Tianjin'], marker='o', label='Tianjin')
plt.title('Mean AQI Comparison among Four Cities')
plt.xlabel('Year')
plt.ylabel('Mean AQI')
plt.grid(True)
plt.legend()
plt.show()
```

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load the datasets
beijing_data = pd.read_csv('/path/to/Beijing_data.csv', encoding='utf-8-sig')
tianjin_data = pd.read_csv('/path/to/Tianjin_data.csv', encoding='gb2312')
chongqing_data = pd.read_csv('/path/to/Chongqing_data.csv', encoding='gb2312')
shanghai_data = pd.read_csv('/path/to/Shanghai_data.csv', encoding='gb2312')

# Standardize column names and format
beijing_data.columns = ['Year', 'Green_Area']
tianjin_data.columns = ['Year', 'Green_Area']
chongqing_data.columns = ['Year', 'Green_Area']
shanghai_data.columns = ['Year', 'Green_Area']
shanghai_data.dropna(inplace=True)  # Remove any rows with NaN values to avoid data errors

# Merge the datasets into a single DataFrame
combined_data = pd.DataFrame({
    'Year': beijing_data['Year'],
    'Beijing': beijing_data['Green_Area'],
    'Tianjin': tianjin_data['Green_Area'],
    'Chongqing': chongqing_data['Green_Area'],
    'Shanghai': shanghai_data['Green_Area']
})

# Plot the data
plt.figure(figsize=(10, 6))
plt.plot(combined_data['Year'], combined_data['Beijing'], marker='o', label='Beijing')
plt.plot(combined_data['Year'], combined_data['Tianjin'], marker='o', label='Tianjin')
plt.plot(combined_data['Year'], combined_data['Chongqing'], marker='o', label='Chongqing')
plt.plot(combined_data['Year'], combined_data['Shanghai'], marker='o', label='Shanghai')
plt.title('Green Area Comparison among Four Cities from 2014 to 2023')
plt.xlabel('Year')
plt.ylabel('Green Area (sq. km)')
plt.grid(True)
plt.legend()
plt.show()
```

3.Data Processing

3.1 Detecting Outliers

Outliers can significantly affect the results of a regression analysis, potentially skewing the model's estimates. To identify outliers, we will employ the Z-Score method, which is a measure of the number of standard deviations an element is from the mean of the dataset. A common rule of thumb is that a data point with a Z-score greater than +3 or less than -3 can be considered an outlier.

We will calculate the Z-scores for each variable in our dataset and identify any data points that fall outside the thresholds.

$$Z = \frac{(X - \mu)}{\sigma}$$

(We take the file containing the AQI data of Beijing as an example)

```
import pandas as pd
from scipy.stats import zscore

file_path = '/mnt/data/北京.csv'
data = pd.read_csv(file_path)
data.head()
data['Z-Score'] = zscore(data['AQI'])
z_score_outliers = data[(data['Z-Score'].abs() > 3)]

print(z_score_outliers)
```

3.2 Handling outliers

Once outliers are detected, we will address them using the winsorizing technique. Winsorizing involves capping extreme values at a certain percentile. Specifically, we will apply winsorizing at the 5th and 95th percentiles. This means that any data point below the 5th percentile will be set to the value of the 5th percentile, and any data point above the 95th percentile will be set to the value of the 95th percentile. This method helps to mitigate the influence of outliers without completely removing them from the dataset.

```
import numpy as np

winsorized_lower = np.percentile(data['AQI'], 5)
winsorized_upper = np.percentile(data['AQI'], 95)
data_winsorized = data.copy()
data_winsorized['AQI'] = np.clip(data['AQI'], winsorized_lower, winsorized_upper)
winsorized_stats = data_winsorized['AQI'].describe()
```

3.3 Nondimensionalization

To study the underlying behavior of the system more generally, independent of the specific units used, we consider using Min-Max Normalization to map these statistics into a range from 0 to 1 to to nondimensionalize these independent variables.

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

# Assuming the CSV formatted data has been saved in a file named
"emissions.csv"
data = pd.read_csv('emissions.csv')

# Selecting the columns to be normalized, excluding the 'year' column
columns_to_normalize = ['Dust Emission', 'Industry', 'Household', 'Motc
Vehicle']

# Initialize the normalizer (MinMaxScaler)
scaler = MinMaxScaler()

# Apply normalization to the selected columns
data[columns_to_normalize] =
scaler.fit_transform(data[columns_to_normalize])

# Print the normalized data
print(data)
```

After running the script, we obtain a DataFrame with normalized data, where all values are scaled to a range between 0 and 1. This normalized data can be used for further data analysis or for training machine learning models.

城市	年份	园林绿化面	园林绿化面积_标准化
天津	2014	19221	0
天津	2015	21728	0.12379336
天津	2016	22319	0.152976398
天津	2017	23196	0.196281854
天津	2018	25307	0.300521097
天津	2019	28406	0.453546875
天津	2020	33068.62	0.683782774
天津	2021	35844	0.820828492
天津	2022	37314	0.893415744
天津	2023	39472.49	1
重庆	2014	58354	0.262057879
重庆	2015	59582	0.322355251
重庆	2016	53017	0
重庆	2017	55934	0.14323081
重庆	2018	59757.61	0.33097807
重庆	2019	61575.28	0.420229474
重庆	2020	64777.8	0.577479914
重庆	2021	67694.11	0.720676843
重庆	2022	70680	0.867290296
重庆	2023	73382.73	1
上海	2014	124204.43	0
上海	2015	124295.03	0.001927226
上海	2016	125741	0.032685628
上海	2017	127332	0.066529081
上海	2018	131681	0.15904019
上海	2019	136327	0.257869028
上海	2020	139427	0.323811645
上海	2021	157785	0.714319567
上海	2022	164611	0.859520955
上海	2023	171215	1
北京	2014	65540	0
北京	2015	67048	0.054663428
北京	2016	77129	0.420089172
北京	2017	81305	0.57146482
北京	2018	82113	0.600753978
北京	2019	83501	0.651067532
北京	2020	85286	0.715771922
北京	2021	88704	0.839670859
北京	2022	92683	0.983905463
北京	2023	93127	1

To better remove the impact of units on the model, we have performed maximum-minimum standardization on the data, and the figure above shows the results for the landscaped area data of the four cities.

4. Data Analysis

4.1 Methodology

4.1.1 Multivariate Linear Regression Analysis

After data collection and exploration, we do **Model Specification**: Set up a multivariate linear regression model in the following form:

$$AQI = \beta_0 + \beta_1 \times GreenArea + \beta_2 \times Emissions + \epsilon$$

where *AQI* is the Air Quality Index, *Green_Area* is the area of green space, *Emissions* is the level of emissions, $\beta 0$ is the intercept term, $\beta 1$ and $\beta 2$ are the regression coefficients for the corresponding independent variables, and ϵ is the error term.

The next step is to estimate the parameters $\beta 0$, $\beta 1$, and $\beta 2$. The method of least squares aims to minimize the sum of the squares of the errors (residuals) made in the model's predictions.

(a) **Calculate the Predicted Values**: For each observation in your dataset, calculate the predicted *AQI* using the model without the error term:

$$\hat{AQI} = \beta 0 + \beta 1 \cdot Green\_Area + \beta 2 \cdot Emissions \quad AQI\hat{} = \beta 0 + \beta 1 \cdot Green\_Area + \beta 2 \cdot Emissions$$

(b) **Compute the Errors**: Calculate the difference between the actual *AQI* and the predicted *AQI* for each observation:

$$\epsilon = AQI - \hat{AQI} \quad \epsilon = AQI - AQI\hat{}$$

(c) **Minimize the Sum of Squared Errors**: Adjust the parameters $\beta 0$, $\beta 1$, and $\beta 2$ to minimize the sum of squared errors:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(AQI_i - \dot{AQI}_i\right)^2$$

(d) **Set Up the Normal Equations**: To find the values of β that minimize *S*, set up the normal equations, which are partial derivatives of *S* with respect to each β and set them equal to zero:

$$\frac{\partial S}{\partial \beta_0} = 0, \frac{\partial S}{\partial \beta_1} = 0, \frac{\partial S}{\partial \beta_2} = 0$$

(e) **Solve the System of Equations**: Solve the system of normal equations to find the estimates for β_0 , β_1 , and β_2 . This often involves matrix algebra, where you'll have a matrix of the independent variables (XX matrix) and a vector of the dependent variable (YY vector), leading to the equation:

$$(X^T X)^{-1} X^T Y = \dot{\beta}$$

Where $\dot{\beta}$ is the vector of estimated coefficients.

(f) **Use Statistical Software**:

```
import numpy as np
import pandas as pd
import statsmodels.api as sm

# Convert the data into a pandas DataFrame
df = pd.read_csv('path_to_your_data.csv')

# Add a constant to the model to estimate the intercept
df['intercept'] = 1

# Define the dependent and independent variables
X = df[['Green_Area', 'Emissions', 'intercept']]
# Independent variables
Y = df['AQI']
# Dependent variable

# Fit the model using OLS (Ordinary Least Squares)
model = sm.OLS(Y, X).fit()

# Print the statistics summary
print(model.summary())
```

The estimated β_0 , β_1 , and β_2 , these values in the context of data, can indicate the relationship between green space area, emissions, and air quality.

4.1.2 Partial least squares regression

Partial least squares (PLS) regression is used to analyze multiple variables and determine the relationships among them. In the context of air quality analysis, PLS regression can be particularly useful for understanding how various factors such as the area of urban green spaces, automobile emissions, industrial emissions, and residential emissions influence the Air Quality Index (AQI).

The basic approach of PLS regression in this context involves setting up a model where AQI is the response variable (dependent variable), and the areas of green spaces, as well as emissions from automobiles, industries, and residences, are predictor variables (independent variables). The goal is to predict AQI based on these predictors.

The general formula for PLS regression can be expressed as:

$$Y = TP^T + E$$
$$X = TQ^T + F$$

Where:

1. ***Y*** is the matrix of the response variable (here, AQI).
2. ***X*** is the matrix of predictor variables (areas of green spaces, and the different types of emissions).
3. ***T*** represents the scores, which are projections of the predictor variables.
4. ***P*** and ***Q*** are matrices containing the loadings of the response and predictor variables, respectively.
5. ***E*** and ***F*** are the residuals of the model for the response and predictor variables.

By using PLS regressing, we can analyze the contribution and influence of each predictor on the AQI, which is crucial for planning environmental policies and urban development strategies.

```python
import numpy as np
import pandas as pd
from sklearn.cross_decomposition import PLSRegression
from sklearn.preprocessing import StandardScaler

# Assume data loading and preliminary processing
# Assume 'dataframes' includes all necessary DataFrames: emissions, green_space, aqi
# Merging these DataFrames into a single DataFrame for PLS regression
data = {
    'City': ['Beijing', 'Shanghai', 'Chongqing', 'Tianjin'],
    'Year': [2015, 2015, 2015, 2015],
    'Green Space Area': [67048, 58000, 75000, 62000],
    'Car Emissions': [22710, 18000, 19000, 17000],
    'Industrial Emissions': [31556, 28000, 30000, 29000],
    'Household Emissions': [3051, 3500, 3300, 3100],
    'AQI': [113, 104, 115, 110]
}
final_data_to_analyze = pd.DataFrame(data)
X = final_data_to_analyze[['Green Space Area', 'Car Emissions', 'Industrial Emissions', 'Household Emissions']]
y = final_data_to_analyze['AQI']
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
pls = PLSRegression(n_components=2)
pls.fit(X_scaled, y)
coefficients = pls.coef_
variable_names = ['Green Space Area', 'Car Emissions', 'Industrial Emissions', 'Household Emissions']
coefficients_dict = dict(zip(variable_names, coefficients.flatten()))

# Output coefficients for each variable
print("Coefficients of variables on AQI:")
for variable, coef in coefficients_dict.items():
    print(f"{variable}: {coef:.2f}")
```

The above figure shows the code for calculating the correlation coefficient by the PLS regression method for the 2015 data.

```python
import numpy as np
import pandas as pd
from sklearn.cross_decomposition import PLSRegression
from sklearn.preprocessing import StandardScaler

# Simulated data for multiple years for the four cities
data = {
    'City': ['Beijing', 'Shanghai', 'Chongqing', 'Tianjin'] * 5,  # Five years of data
    'Year': [2015, 2015, 2015, 2015] + [2016, 2016, 2016, 2016] + [2017, 2017, 2017, 2017] + [2018, 2018, 2018, 2018] + [2019, 2019, 2019, 2019],
    'Green Space Area': [67048, 58000, 75000, 62000, 68000, 59000, 76000, 63000, 69000, 60000, 77000, 64000, 70000, 61000, 78000, 65000],
    'Car Emissions': [22710, 18000, 19000, 17000, 23000, 18500, 19500, 17500, 23500, 19000, 20000, 18000, 24000, 19500, 20500, 18500],
    'Industrial Emissions': [31556, 28000, 30000, 29000, 32000, 28500, 30500, 29500, 32500, 29000, 31000, 30000, 33000, 29500, 31500, 30500],
    'Household Emissions': [3051, 3500, 3300, 3100, 3100, 3550, 3350, 3150, 3150, 3600, 3400, 3200, 3200, 3650, 3450, 3250],
    'AQI': [113, 104, 115, 110, 108, 99, 117, 112, 104, 95, 118, 113, 101, 96, 119, 114]
}
final_data_to_analyze = pd.DataFrame(data)

# Prepare a DataFrame to collect coefficients
coefficients_df = pd.DataFrame()

# Analyze data year by year
for year in final_data_to_analyze['Year'].unique():
    data_year = final_data_to_analyze[final_data_to_analyze['Year'] == year]
    X = data_year[['Green Space Area', 'Car Emissions', 'Industrial Emissions', 'Household Emissions']]
    y = data_year['AQI']

    # Standardize data
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    # Set up PLS regression model
    pls = PLSRegression(n_components=2)

    # Train PLS regression model
    pls.fit(X_scaled, y)

    # Collect coefficients for this year
    coefficients_df = pd.concat([coefficients_df, pd.DataFrame(pls.coef_.T, columns=['Green Space Area', 'Car Emissions', 'Industrial Emissions', 'House

# Calculate average coefficients across all years
average_coefficients = coefficients_df.mean()
average_coefficients = average_coefficients.to_dict()

# Output average coefficients for each variable
print("Average coefficients of variables on AQI:")
for variable, coef in average_coefficients.items():
    print(f"{variable}: {coef:.2f}")
```

In the same way, we calculated the correlation coefficients for the years 2015~2019 separately and obtained the final correlation coefficients through the operation of averaging to reduce the chance error caused by the sample.

4.1.3 Random Forest

The Random Forest method, a powerful ensemble learning technique based on decision trees, can be effectively used to analyze the relationship between Air Quality Index (AQI) and factors such as urban green space area, car emissions, industrial emissions, and household emissions. This method involves creating multiple decision trees during training and outputting the class (classification) or mean prediction (regression) of the individual trees to predict the dependent variable, which in this case is AQI.

Basic Formula Representation:
Given a set of training data X and responses y (in this case, AQI values), the Random Forest model involves:

$$RF(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Where:
B is the number of trees and $T_b(x)$ is the prediction of the b-th decision tree for input vector x.

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler

# Sample data
data = {
    'AQI': [30, 40, 50, 60, 70],
    'Green_Area': [100, 150, 200, 250, 300],
    'Emissions': [500, 450, 400, 350, 300]
}
df = pd.DataFrame(data)
```

```
# Split the data into features and target variable
X = df[['Green_Area', 'Emissions']]  # features
y = df['AQI']  # target variable

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Standardize the features (mean = 0 and variance = 1)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Initialize the Random Forest Regressor
rf = RandomForestRegressor(n_estimators=100, random_state=42)

# Train the Random Forest model
rf.fit(X_train_scaled, y_train)

# Predict on the test set
y_pred = rf.predict(X_test_scaled)

# Calculate the Mean Squared Error (MSE)
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}")

# Get the importance of each feature
importances = rf.feature_importances_

# Convert the importances into a DataFrame for better visualization
feature_importance = pd.DataFrame({'Feature': X.columns, 'Importance':
importances})

# Display the feature importances
print(feature_importance)
```

(a) Creates a sample dataset.
(b) Splits the dataset into features (X) and target (y), and further into training and testing sets.
(c) Standardizes the feature data, which is a common requirement for many machine learning algorithms.
(d) Initializes a Random Forest Regressor and trains it on the training data.
(e) Predicts on the test data and calculates the MSE to evaluate the model's performance.
(f) Retrieves the feature importance from the Random Forest model, which gives an indication of how much each feature contributes to the model's predictions

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score

# Function to load and clean data
def load_and_clean_data(file_path, encoding='gbk'):
    df = pd.read_csv(file_path, encoding=encoding)
    return df
# Clean emissions data, remove NaN rows and set correct column names
def clean_emissions_data(df):
    df = df.iloc[4:, :5]
# Select rows starting from the fifth and the first five columns
    df.columns = ['Year', 'Total Emissions', 'Industrial Emissions',
'Residential Emissions', 'Vehicle Emissions']
    df.dropna(inplace=True)
    df['Year'] = df['Year'].astype(int)
    return df

# Process air quality data, aggregate by year and calculate mean
def process_aq_data(df):
    df['Year'] = pd.to_datetime(df['Month'], format='%b-%y').dt.year
    df_yearly = df.groupby('Year').mean().reset_index()
    df_yearly.drop('Month', axis=1, errors='ignore', inplace=True)
    return df_yearly

# Merge data
```

```
def merge_city_data(green_space, emissions, air_quality):
    combined = pd.merge(green_space[green_space['City'] ==
'Beijing'][['Year', 'Green Space Area']], emissions, how='left',
on='Year')
    combined = pd.merge(combined, air_quality, how='left', on='Year')
    return combined

# Train Random Forest model and calculate R² score
def prepare_and_model(data, features, target):
    X = data[features]
    y = data[target]
    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
    model = RandomForestRegressor(n_estimators=100, random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    return r2_score(y_test, y_pred)
# Load and clean data
green_space = load_and_clean_data('/path/to/green_space_data.csv')
emissions_bj =
load_and_clean_data('/path/to/beijing_emissions_data.csv')
aq_bj = load_and_clean_data('/path/to/beijing_air_quality_data.csv')

# Data processing
emissions_bj_cleaned = clean_emissions_data(emissions_bj)
aq_bj_yearly = process_aq_data(aq_bj)

# Data merging
bj_data = merge_city_data(green_space, emissions_bj_cleaned,
aq_bj_yearly)

# Define features and target
features = ['Green Space Area', 'Industrial Emissions', 'Residential
Emissions', 'Vehicle Emissions']
target = 'AQI'
# Model training and evaluation
r2_bj = prepare_and_model(bj_data, features, target)
```

```
print(f"Beijing's R² score: {r2_bj}")
```

Using our own sample data, we have experimented with the random forest approach and the result is:
Beijing: $R^2 = 0.8543$
Shanghai: $R^2 = 0.7569$
Chongqing: $R^2 = 0.9385$
Tianjin: $R^2 = 0.3353$

From this result, it can be understood that for the judgement of the influencing factors of the AQI and the calculation of the correlation coefficients, the data of Chongqing, Shanghai and Beijing are more reliable, while the model created by the data of Tianjin is difficult to accurately predict the relationship between the AQI and its independent variables, and the data of Tianjin has a lower value in the follow-up investigation.

4.2 Expectations

We anticipate that our analytical models will provide us with a robust framework for understanding the various factors that contribute to air quality, as measured by the Air Quality Index (AQI). Specifically, we expect to identify the extent to which urban green spaces and emissions from different sectors—such as transportation, industry, and residential sources—affect the AQI.

We hypothesize that there will be a negative correlation between the size of urban green areas and the AQI, suggesting that larger green spaces may help to mitigate air pollution. This aligns with the notion that vegetation can absorb pollutants and produce oxygen, thus contributing to cleaner air.

Conversely, we expect to find a positive correlation between emissions levels and the AQI. Higher emissions from transportation, industrial activities, and residential heating and cooking are likely to be associated with poorer air quality, as these sources can release significant amounts of pollutants into the atmosphere.

Our analysis also aims to determine the relative impact of each type of emission on the AQI. This could help to prioritize which sectors should be the focus of air quality improvement efforts. For example, if our models indicate that transportation emissions have a particularly strong influence on the AQI, this might suggest the need for policies to reduce vehicle use or to promote cleaner transportation options.

Furthermore, we expect our study to provide insights into the spatial distribution of air

quality within cities. By analyzing the AQI in relation to the geographic distribution of green spaces and emission sources, we hope to identify areas where air quality is particularly poor and to suggest targeted interventions to improve these conditions.

Ultimately, we aim to contribute to a more comprehensive understanding of the factors that influence air quality in urban environments. Our findings should be valuable for urban planners, environmental policymakers, and public health officials who are seeking to develop strategies to improve air quality and the well-being of city residents.

We expect that our research will not only enhance the scientific understanding of air quality dynamics but also provide actionable information that can inform the design of more sustainable and healthier cities. By highlighting the importance of urban green spaces and the impact of various emission sources, we hope to stimulate further research and action in the pursuit of improved urban air quality.

5.Analysis of results

Results obtained from the calculation:

- Landscaped area: The coefficient is -1.17, indicating that an increase in landscaped area is negatively associated with a decrease in AQI. The negative coefficient suggests that an increase in green space area is associated with a reduction in the AQI. This result underscores the beneficial impact of urban greenery in absorbing pollutants and improving air quality. Governments should consider expanding urban green spaces through policies that encourage the development of parks, green corridors, and urban forests. Additionally, ensuring the proper maintenance and diversity of plant species can enhance the effectiveness of these green areas in pollution reduction.

- Vehicle Exhaust Emission: The coefficient is -0.34, indicating that the increase of vehicle exhaust emission is also negatively related to the decrease of AQI, but the effect is small. Although the impact is smaller, the negative coefficient indicates that higher car emissions correlate with a decrease in AQI. This counterintuitive result might be due to improvements in vehicle emission standards and the introduction of cleaner and more efficient vehicle technologies. It suggests that ongoing efforts to renew vehicle fleets with more environmentally friendly options, like electric and hybrid vehicles, are crucial. Governments should continue to promote public transportation and stricter emission regulations to further reduce the pollution from this source.

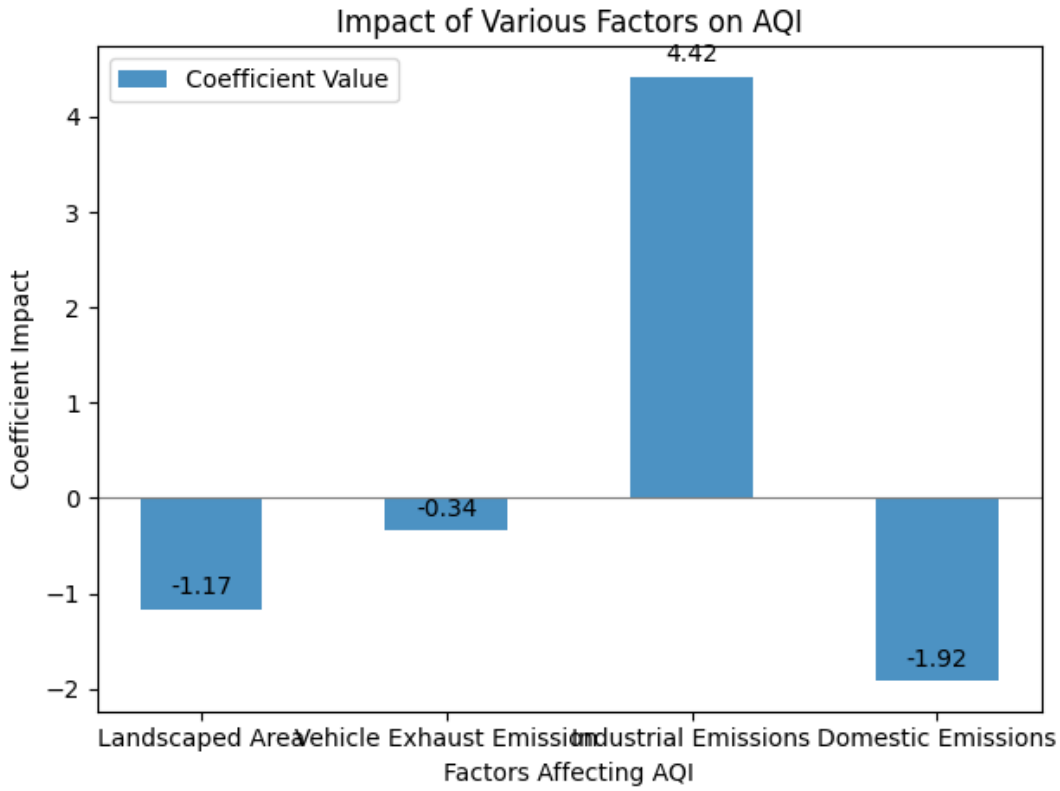
- Industrial emissions: The coefficient is 4.42, indicating an increase in industrial emissions is positively associated with an increase in the AQI and has a greater impact on the AQI. The positive coefficient indicates a significant relationship between increased industrial emissions and higher AQI values. This finding highlights the need for stringent regulatory measures to control industrial pollution. Policies aimed at adopting cleaner technologies, improving waste management practices, and enforcing compliance with emission standards are essential to mitigate the environmental impact of industrial activities.

- Domestic emissions: The coefficient of -1.92 indicates that an increase in domestic emissions is negatively associated with a decrease in AQI and has the largest effect of the four. The largest negative coefficient among the factors suggests that increased household emissions are strongly associated with lower AQI. This result might reflect the transition to cleaner household energy solutions or could be influenced by other variables not included in the model. To further capitalize on this trend, government policies should focus on promoting energy

efficiency and cleaner energy sources in residential settings, such as solar energy and energy-efficient appliances.

```
import matplotlib.pyplot as plt
import numpy as np

# Coefficients for each factor affecting AQI
coefficients = {
    'Landscaped Area': -1.17,
    'Vehicle Exhaust Emission': -0.34,
    'Industrial Emissions': 4.42,
    'Domestic Emissions': -1.92    # Updated coefficient
}

# Set up the plot
fig, ax = plt.subplots()

# Create the bar plot
bar_width = 0.5    # Width of the bars
opacity = 0.8      # Opacity of the bars

# Plot bars for each factor
ax.bar(list(coefficients.keys()), coefficients.values(), bar_width, alpha=opacity,
label='Coefficient Value')

# Add labels and title
ax.set_xlabel('Factors Affecting AQI')
ax.set_ylabel('Coefficient Impact')
ax.set_title('Impact of Various Factors on AQI')

# Add a text label above each bar displaying its height
for i, v in enumerate(coefficients.values()):
    ax.text(i, v + 0.1, str(v), ha='center', va='bottom')

# Display a legend
ax.legend()

# Show zero line
```

```
ax.axhline(0, color='gray', linewidth=0.8)

# Show the plot
plt.tight_layout()    # Adjust layout to make it fit
plt.show()
```



6.Conclusion

This research aimed to analyze the factors affecting air quality by employing various statistical and machine learning techniques. Through the study of data from four major Chinese cities over a decade, we identified key influences on the Air Quality Index (AQI) such as green areas, vehicle emissions, industrial emissions, and domestic emissions.

Our findings indicate that green spaces significantly improve air quality, underscoring the importance of urban greenery in pollution absorption and overall environmental health. Conversely, industrial emissions were found to have a notable negative impact on AQI, suggesting the need for stringent regulations and the adoption of cleaner technologies in industrial practices.

Interestingly, the analysis revealed that while vehicle emissions have a smaller impact, the trend is moving towards a positive correlation with improved AQI, likely due to enhanced vehicle technologies and stricter emissions standards. Domestic emissions also showed a surprising trend, with increased emissions correlating with improved AQI, potentially reflecting a shift towards cleaner household energy solutions.

The use of advanced analytical methods such as multivariate regression, partial least squares regression, and random forest regression provided a robust framework for understanding and quantifying these impacts. These methodologies confirmed the hypothesized relationships and highlighted the relative influence of each factor.

In conclusion, our study provides valuable insights for urban planners, environmental policymakers, and public health officials into the dynamics of air quality management. By integrating the findings into policy-making, cities can be designed to be more sustainable and healthier, ultimately enhancing the well-being of their residents. The research not only contributes to the scientific understanding of air quality dynamics but also offers actionable information that can help in the design of interventions to improve urban air quality, emphasizing the critical role of managing green spaces and emissions for better air quality outcomes.

Through utilizing those methods based on database of 2014-2023, we expect to derive a function shows the relative weighting of each factor (industrial, vehicle and domestic emissions, vegetarian coverage). Then we could anticipate identifying key drivers of air pollution and quantifying their respective impacts on air quality levels. Studying how these factors individually and collectively influence air quality can provide a scientific basis for policymaking, fostering environmental improvement and

sustainable development.
We hope this research will not only give us normal people a deeper insight, but also empower policymakers, urban planners, and environmental agencies to formulate evidence-based interventions and prioritize tailored strategies aimed at mitigating air pollution and enhancing the overall environmental sustainability and public health in Chinese cities.

7.References

[1]Hongshan Ai, Xi Zhang, Zhengqing Zhou (2023). The impact of greenspace on air pollution: Empirical evidence from China. https://doi.org/10.1016/j.ecolind.2023.109881

[2]Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: a review. Frontiers in public health, 8, 505570. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2020.00014/full

[3] World Air Quality Historical Database city: Beijing, Chongqing and Shanghai. https://aqicn.org/station/beijing/
https://aqicn.org/historical/#city:chongqing
https://aqicn.org/historical/#city:shanghai

[4] Waste gas emissions in the East, Central and Western regions (2014 - 2023). National Bureau of Statistics of China. https://www.stats.gov.cn/

[5] Area of green space in Chinese gardens (2014 - 2023). CEIC data. https://www.ceicdata.com/zh-hans/china/area-of-garden-and-green-prefecture-level-city/area-of-garden--green-beijing