CHAPTER ONE

Introduction

1.1 INTRODUCTION

Air pollution can be defined as the presence in the external atmosphere of one or more contaminants (pollutants), or combinations thereof, in such quantities and of such duration as may be or may cause injury to human health, plant or animal life, or property (materials), or which unreasonably interfere with the comfortable enjoyment of life, or property, or the conduct of business (Canter, 1996).

Breathing is not optional, it is essential even for a short time, and air has to be used as it is found. The Times newspaper (1881) quoted:

'The air we receive at our birth and resign only when we die is the first necessity of our existence'.

Despite its essential ingredient to life, air quality has been historically variable and frequently to the detriment of human health. Nevertheless, our quality of life dramatically improved during the twentieth century. Now, however, a growing body of research has found that certain pollutants may affect human health at lower concentrations than had previously been thought. This concern has heightened public anxiety to the importance of improving and managing air quality. It is paramount that a resource as important as air quality is protected and managed for future generations (Department of the Environment (DoE), 1993).

1.2 AIR POLLUTION - A CONCERN

Concerns about air quality have probably been around as long as mankind. From the moment fire was invented air pollution became a problem (Brimblecombe, 1987) and it has been a problem ever since. Concerns have occurred periodically throughout history, and are well documented (e.g. Ashby and Anderson, 1981; Brimblecombe, 1987; National Society for Clean Air and Environmental Protection (NSCA), 2000a).

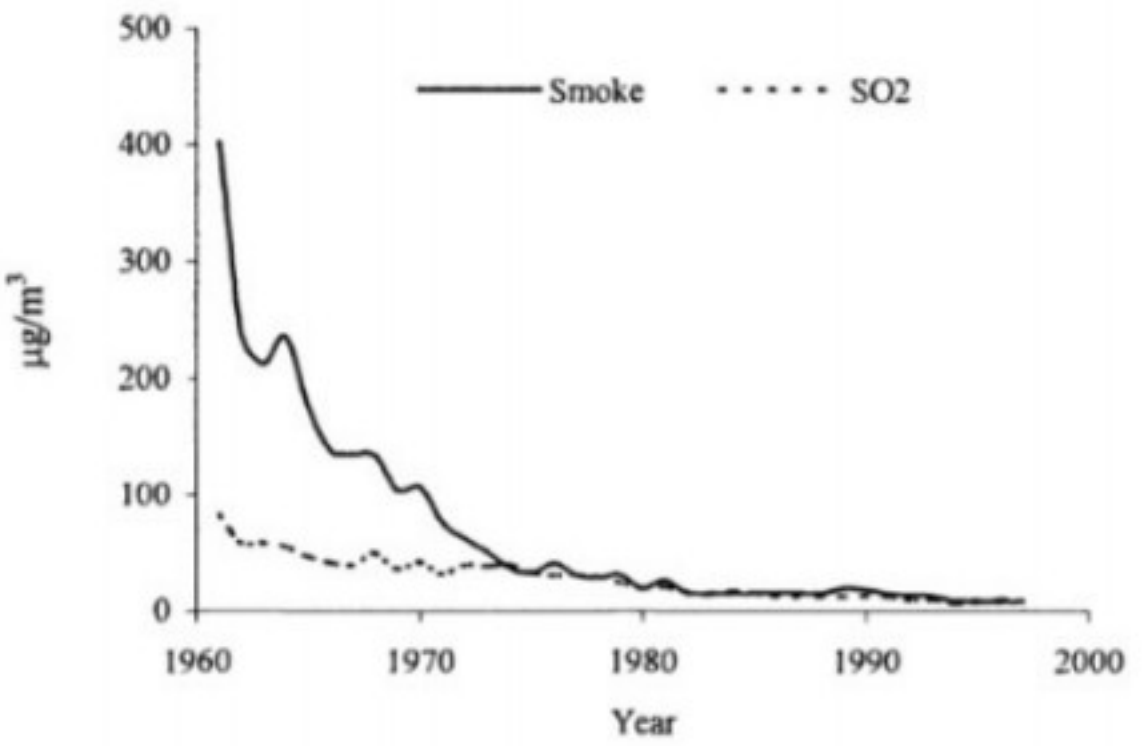
Perhaps the historical air quality problems of the UK are the best documented. In 1257 in England, Queen Eleanor, wife of Henry III, was obliged to leave Nottingham on account of smoke nuisance. In 1273 a Royal proclamation was issued by Edward I to prohibit the use of sea coal in open furnaces because of the prejudicial effects to health from smoke emissions. Three hundred years later in 1578 it was written that Queen Elizabeth I:

'findeth hersealfe greatly greved and annoyed with the taste and smoke of sea-cooles'.

'examine the nature, causes and effects of air pollution and the efficacy of present preventive measures; to consider what further measures are practicable; and to make recommendations'.

This culminated in the introduction of the Clean Air Act (CAA) in 1956, which was later amended and extended by the 1968 CAA. The Acts constituted the operative legislation against pollution by smoke, grit and dust from domestic fires and other commercial and industrial processes not covered by the Alkali Acts and other subsequent pollution legislation. Again the improvement in air quality was discernible. Annual average SO_2 and smoke levels diminished considerably; the most dramatic reductions were achieved in urban areas (Fig. 1.2).

Figure 1.2 Historical smoke and SO_2 concentrations ($\mu g/m^3$) for Glasgow, Scotland (Data from the Department of the Environmental, Transport and the Regions (DETR), 2000b)



The reduction in the concentration of SO_2 and smoke was brought about by the burning of cleaner fuels, especially the use of gas, the use of tall stacks on power stations, their relocation outside cities, and the decline of heavy industry (DoE, 1993). The 1956 and 1968 Acts have subsequently been consolidated and their provisions re-enacted in the 1993 CAA. Occurrences of poor air quality have also been found in other countries (Chapter 2, Section 2.3 and Chapter 3, Section 3.2.1).

For most developed and developing countries motor traffic emissions now pose a principal threat to air quality, particularly in urban areas. Petrol and diesel engines emit a wide variety of pollutants, principally carbon monoxide (CO), oxides of nitrogen (NO_x), volatile organic compounds (VOCs) and particulates, which have an increasing impact on air quality. Whilst improvements in motor exhaust emission controls and fuel technology (Chapters 7 and 8) have resulted in an improvement in air quality (e.g. the elimination of lead (Pb) in fuel), concerns still persist about the elevated levels of pollution including the occurrence of photochemical smogs (or hazes) (Chapters 2 and 3). Photochemical reactions resulting from the action of sunlight on nitrogen dioxide (NO_2) and VOCs from

Table 2.1 Concentration (ppm) of gases comprising the atmosphere (Stern, 1976; O'Neill, 1985)

Molecular species	*Background concentration*	*Polluted air*	*Percentage weight*
N_2	780840		75.5
O_2	209480		23.2
Ar	9340		1.3
CO_2	314-318[a]		
Ne	18.20		$1.3x10^{-3}$
He	5.24		
CH_4	1.0-2.0	1-10	$72x10^{-6}$
Kr	1.1		$0.45x10^{-3}$
H_2	0.5		$23x10^{-6}$
N_2O	0.25-0.5		
CO	0.1	5-10	
O_3	0.01-0.07	0.1-0.5	
NO_2	0.001-0.02	0.2-1.0	
NO	0.002-0.002		
SO_2	0.0002-1.0	0.02-2.0	
H_2S	0.0002		
C			$9.3x10^{-3}$
Xe			$40x10^{-6}$
S			$70x10^{-9}$

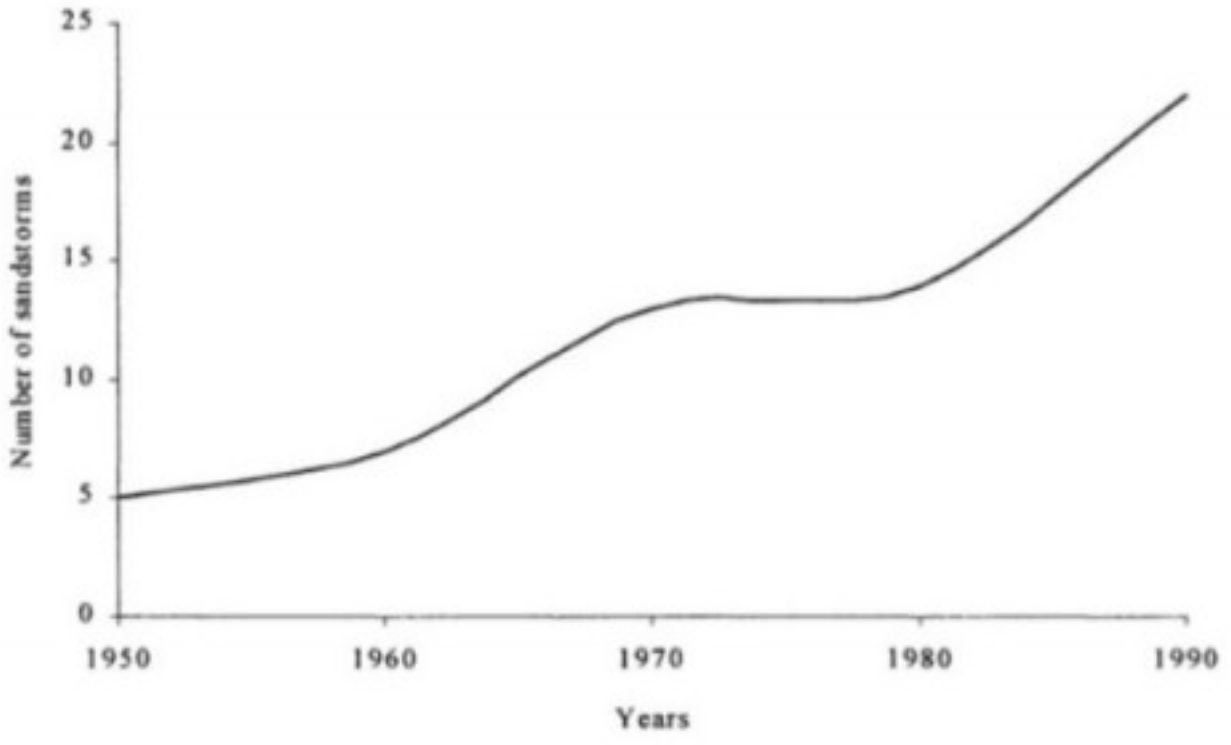
a) Section 2.4.3

2.3 AIR POLLUTION SOURCES

Air pollution sources may be either anthropogenic or natural. However, as human activity disturbs natural systems, the distinction may become blurred. Natural sources include dust storms, volcanic action, forest fires, etc. For some pollutants, e.g. SO_2, natural sources exceed anthropogenic sources on a global scale. Incursions from the stratosphere increase ground level (tropospheric) concentrations of O_3. However, when considering the effects of air pollutants on health, especially in urban areas where population densities are high, anthropogenic sources are very important and are those to which attention is usually directed with a view to control.

Many pollutants are emitted into the atmosphere from naturally occurring sources. An example of a natural pollution problem is dust storms. In north China (China Daily, 2000) in the Capital Circle, Beijing and Tianjin municipalities and Zhangjiakou and Chengde in Hebei Province, large areas are affected by strong winds and elevated dust levels from northern Inner Mongolia. The areas are also exposed to soil erosion, dry climates and water shortages. The number of dust storms has risen in recent years (Fig. 2.1). Efforts are currently being made to control desertification in the Capital Circle area by 2010 through tree and fauna planting schemes.

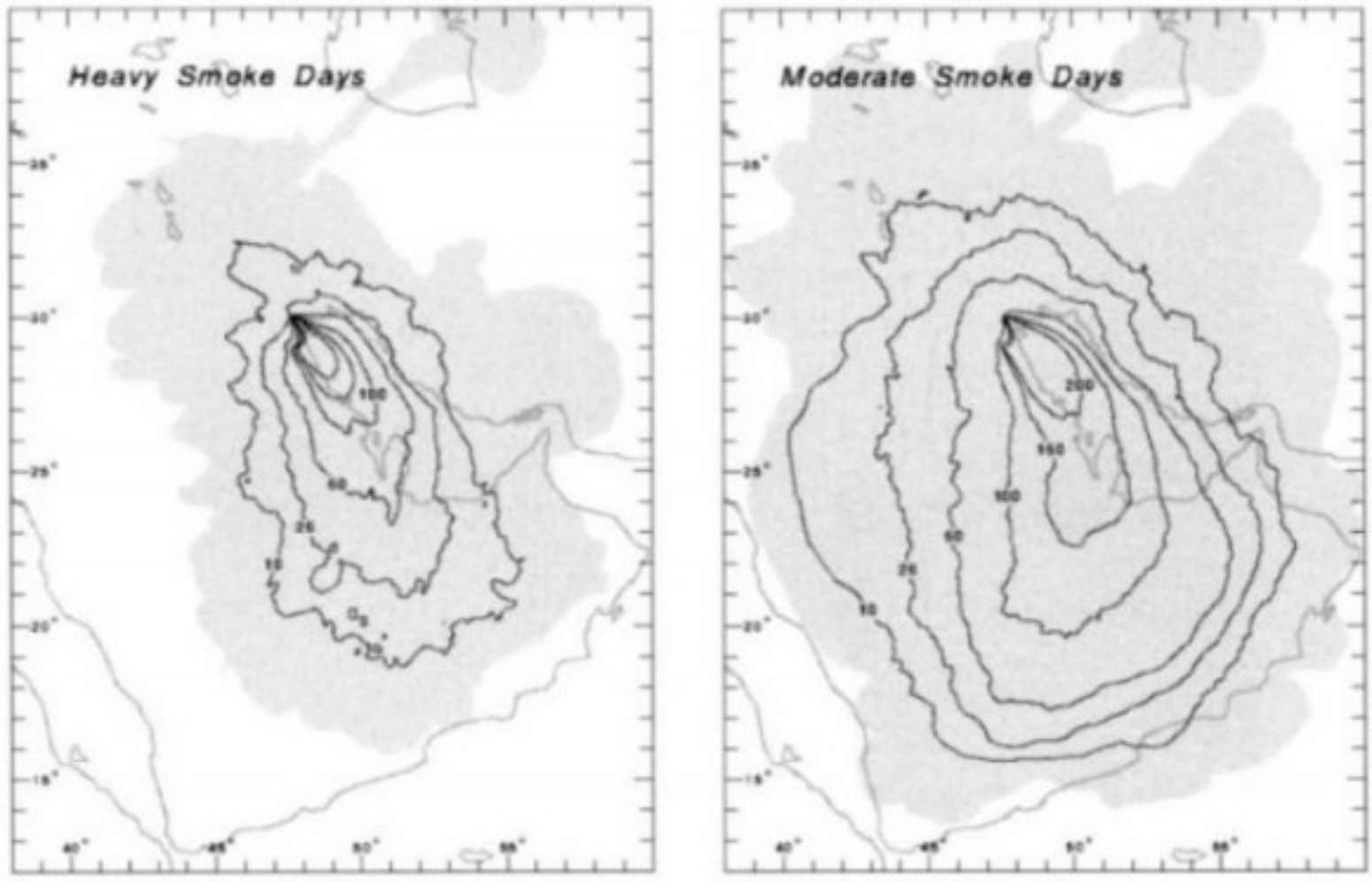
Figure 2.1 Frequency of dust storms in North China in the Capital Circle area (Data from the China Daily, 2000)



Pollutants can be emitted from point or stationary, area or mobile (linear) anthropogenic sources (Case Study 3, Chapter 10.4). Each source has its own distinct emission characteristics. Point sources include stacks, flues, etc. Area sources include groupings of usually small sources spread over a delineated area (e.g. industrial complexes) and mobile sources include motor vehicle, aircraft, etc. The form of the emission from these sources may be either controlled, uncontrolled, accidental, intentional or fugitive (e.g. uncontrolled minute leaks) (Harrop, 1999) (Chapter 4, Section 4.4). Case Study 12 (Chapter 10, Section 10.13) details the quantification of fugitive emissions from an oil terminal.

In the past, accidental releases of emissions to air have caused severe impacts on human health and air quality (Fig. 2.2). Some accidents have had extremely adverse effects on human health. Such well documented episodes include Bhopal (India) and Seveso (Italy). In December 1984, in Bhopal, approximately 2,500 people died when approximately 40t of methyl isocyanate was accidentally released from the Union Carbide chemical plant. Nearly 200,000 people were also injured, mainly from respiratory and eye injuries. In Seveso, in July 1976, an explosion occurred at a chemical plant making trichlorophenol as a herbicide, which released various chemicals, including dioxin (2,3,7,8-tetrachlorodibenzoparadioxine (TCDD). A cloud of dioxin, trichlorophenol, ethylene glycol and caustic soda dispersed into the surrounding environs. Within several weeks, fauna and flora died and residents surrounding the plant were admitted to hospital. No deaths were linked to the accident, but because of the possible link between dioxin and genetic mutations, 90 women decided to have abortions (Elsom, 1987). More than 700 people living close to the plant were evacuated and another 5000 people in a less contaminated area were permitted to stay at home but had restrictions placed on them with regard to raising animals, gardening and children playing (Fuller, 1977). As a result of the Seveso accident the European Union (EU) introduced the 'Seveso Directive' in 1984 (Chapter 7, Section 7.9.4).

Figure 2.3 Dispersion of particulate emissions from the intentional burning of oil wells in the Gulf War, 1991 (KuDA, 2000)



Contours indicate the number of days with smoke overhead (based on afternoon satellite image from the NOAA-11 polar orbiting meteorological satellite). The grey background areas indicate the areas for which smoke was detected overhead on at least one day. The analysis includes the period during the Gulf War, as well as for the rest of the year until the oil fires were extinguished.
Courtesy of Kuwaiti Data Archive (KuDA) Project 2000 at the National Centre for Atmospheric Research (NCAR) in Boulder, Colorado, USA. NCAR is sponsored by the National Science Foundation.

haze blanketed vast areas for weeks with reported extreme levels of pollution and visibility <200m in the interior of East Kalimantan. In Sarawak, where an emergency was declared, the air pollution index (API) exceeded 800 (Chapter 7, Section 7.3).

The deliberate burning of vegetation to rejuvenate vegetation growth is another contributor to air quality. The burning of grass on the Savannah (Fig. 2.4) is a widespread practice, however it causes extensive air quality and smoke/particulate nuisance problems to communities.

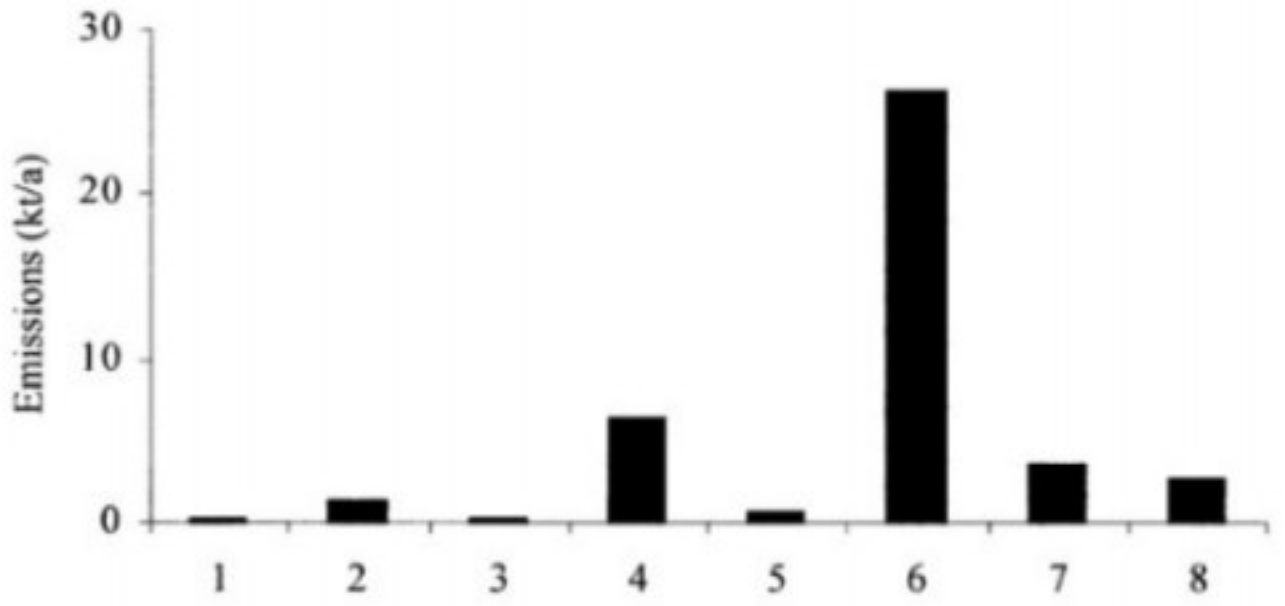
2.4 TYPES OF POLLUTANTS AND THEIR SOURCES

Generally air pollution may be either gaseous or particulate in nature. However, other forms of pollution of potentially equal concern may be physical (e.g. radiation) or heat related.

Fine and coarse particles generally have distinct sources and formation mechanisms, although there may be some overlap (Table 2.2). Primary fine particles are formed from condensation of high temperature vapours during

total emissions (Fig. 2.5), although in the UK it is estimated that benzene levels will decline by almost 40% by 2010 on a 1995 base (DoE, 1997a). In Europe benzene emissions from existing petrol storage and handling facilities will come under the control of the EC Directive (94/63/EC) on controlling VOC emissions resulting from storage and distribution of petrol to service stations. Other EC Directives will further reduce motor vehicle emissions for cars, light vehicles and heavy goods vehicles (HGVs) sold from 2001 and 2006 as part of the EC Auto-Oil programme. The programme will reduce the amount of benzene and aromatics in petrol from the year 2000 as well as reduce the sulphur content of fuels from 2000 and again from 2005. The reduction of fuel sulphur content will help to reduce the deterioration in catalyst performance and therefore help to abate benzene emissions (DETR, 1998).

Figure 2.5 UK emission sources for benzene (Data from Goodwin *et al.*, 1999; SEPA, 2000)



1. *Public power, co-generation and district heating*
2. *Commercial, institutional and residential combustion*
3. *Industrial combustion*
4. *Production process*
5. *Extraction and distribution of fossil fuels*
6. *Road transport*
7. *Other mobile sources and machinery*
8. *Waste treatment and disposal*

2.4.1.2 1,3-Butadiene

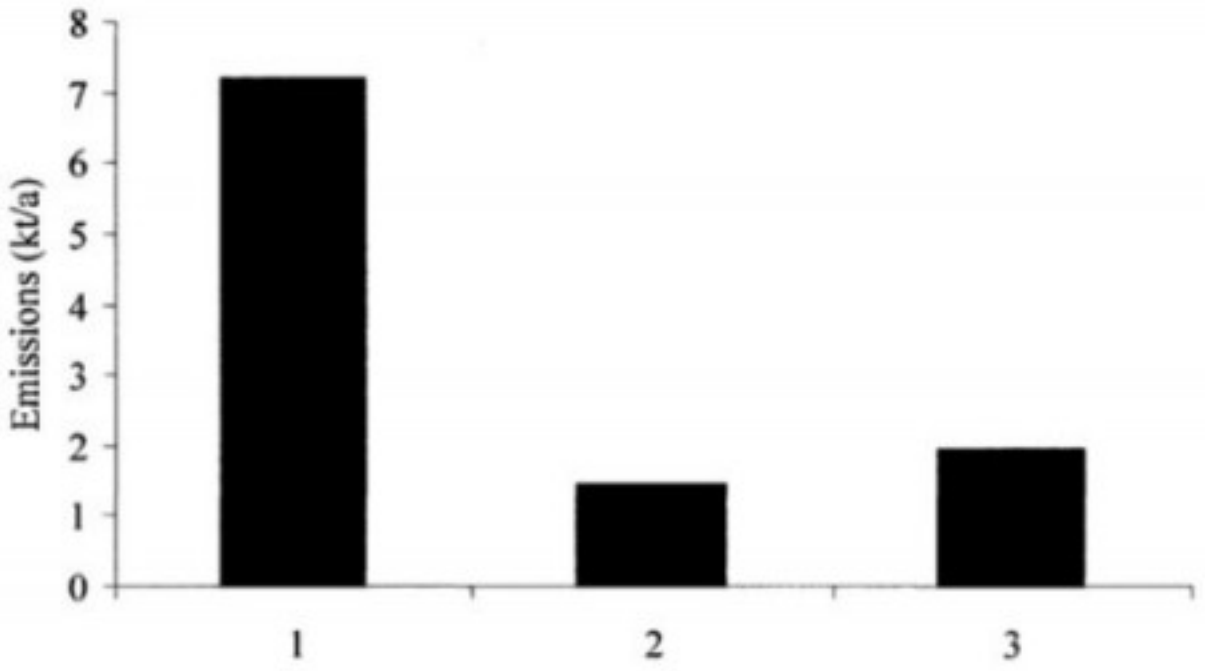
1,3-butadiene is a VOC arising from the combustion process of petroleum products. It disperses relatively rapidly within the atmosphere. Motor vehicles are the principal source of emissions. For example in the UK they account for 68% of total emissions whilst industrial chemical processes account for 18.3% (Fig. 2.6). In the UK emissions from petrol engine motor vehicles are anticipated to have declined by about 55% by 2000 on 1992 values and by 73% by 2010 (DoE, 1997a). Table 2.4 details typical levels of 1,3-butadiene in air for selected sites and countries. In the EU emissions of 1,3-butadiene will be reduced by the introduction of EC Directive (94/63/EC) and the Auto-Oil programme (DETR, 1998).

Table 2.4 Summary of 1,3-butadiene measurements (μg/m^3) (DETR, 2000b; USEPA, 2000)

Country	*Concentration*
UK	0.1-1.7[a, b]
Scotland (Edinburgh)	0.2
US	0.22[c]

a) *Range of annual averages for UK sites for 1999*
b) *Includes Scotland*
c) *From USEPA Aerometric Information System (AIRS), USEPA (2000)*

Figure 2.6 UK emission sources for 1,3 butadiene (Data from Goodwin *et al.*, 1999; SEPA, 2000)



1. *Road transport*
2. *Other mobile sources and machinery*
3. *Chemical industry*

2.4.2 Carbon monoxide

Carbon monoxide (CO) is a colourless, odourless gas formed during the incomplete combustion of carbon containing materials. CO is a relatively stable compound which converts to CO_2 within the atmosphere as a result of a reaction with hydroxyl radicals (Scottish Office, 1998). Typically, concentrations fall relatively rapidly with distance from source. For example, CO levels resulting from motor vehicle emissions typically return to background levels within 200 m of a road (Hickman and Colwill, 1982). Consequently air quality impacts are generally localised.

Natural ambient concentrations of CO range between 0.01-0.23 mg/m^3 (WHO, 1987). In urban environments, mean concentrations over 8 hours are usually less than 20 mg/m^3 and 8 hour maximum levels are usually less than 60 mg/m^3. Concentrations of CO can be high in vehicles, underground car-parks, road

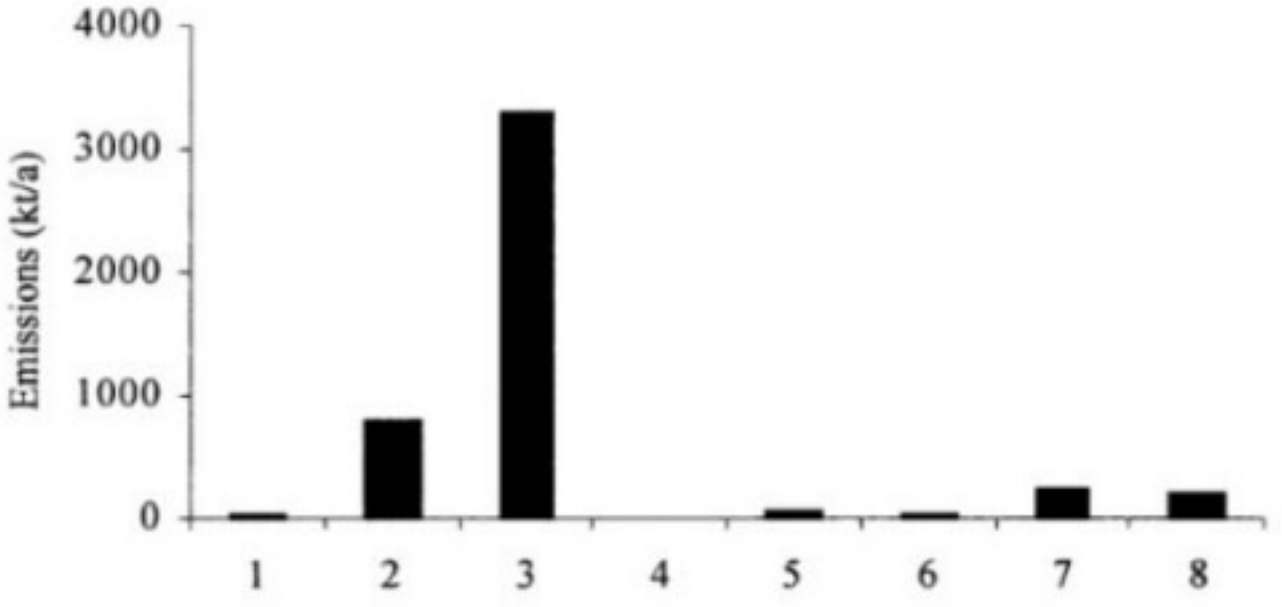
tunnels and in other indoor environments where combustion engines operate with inadequate ventilation. Case Study 2 details the concentrations found in a poorly ventilated indoor area and the mitigation measures used to reduce concentrations to an acceptable level (Chapter 10, Section 10.3). In these circumstances, mean concentrations of CO can reach up to 115 mg/m^3 for several hours (WHO, 2000b). Table 2.5 shows typical levels of CO in air for selected countries.

Table 2.5 Summary of CO measurements

Country	*Concentration*	*Source*
Canada[a,b]	0.1-1.2	Environment Canada, 2000
Costa Rica[c]	9.3-10.3	Ministerio de Planificacion, 2000
Japan[a,d]	0.6	Japanese Government, 2000b
New Zealand[e]	5-10 urban <0.1 rural	New Zealand Government, 2000
Singapore[a] (1998)	0.6 (2.0)	Singapore Government, 2000
UK[a,f]	0.2-2.1	DETR, 2000b
US[g]	0.08-0.2	USEPA, 2000

a) Average, mg/m^3
b) Range of annual average levels for all NAP stations and cities for 1998
c) 1995 – 1997 San Jose (sampling duration not known) µg/m^3
d) Annual average levels from 145 general environmental monitoring stations for 1998
e) mg/m^3, 8 hour
f) Range of annual average values for UK automated sites for 1999
g) New York for 12 sites units are mg/m^3 (maximum 1 hr)

Figure 2.7 UK emission sources for CO (Data from Goodwin *et al.*, 1999; SEPA, 2000)



1. *Waste treatment and disposal*
2. *Other mobile sources and machinery*
3. *Road transport*
4. *Extraction and distribution of fossil fuels*
5. *Production process*
6. *Industrial combustion*
7. *Commercial/institutional/residential combustion*
8. *Public power/co-generation/district heating*

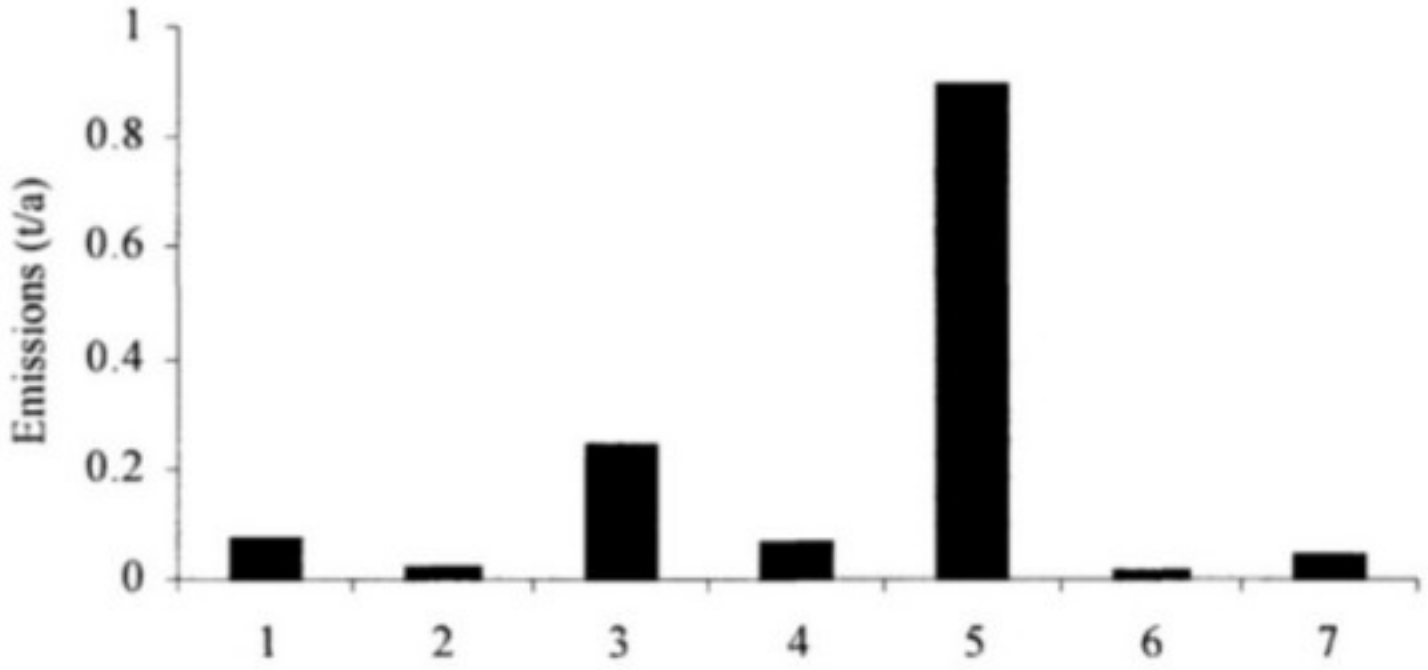
(WHO, 2000b). Table 2.6 shows typical levels of lead in air levels for selected and countries.

Table 2.6 Summary of lead measurements ($\mu g/m^3$)

Country	*Annual concentration*	*Source*
Canada[a]	0-0.63	Environment Canada, 2000
Costa Rica[b]	0.2-0.5	Ministerio de Planificacion, 200
Singapore (1998)	0.1-0.2	Singapore Government, 2000
UK[c]	0.004-1.43	DETR, 2000b
US[d]	0.02-0.2	USEPA, 2000
Venezuela[e]	0.4-0.9	MARNR[f], 1997

a) Range of annual average levels for all NAP stations and cities for 1998
b) 1995 – 1997 San Jose (sample duration not known)
c) Annual average levels for 1999 for UK monitoring sites
d) New York, average for 6 sites
e) San Cristobal, Valencia and Puerto Ordaz (1997)
f) Ministerio del Ambiente y de Los Recursos Naturales Renovables

Figure 2.8 UK emission sources for lead in air (Data from Goodwin *et al.*, 1999; SEPA, 200



1. *Public power, co-generation and district heating*
2. *Commercial, institutional and residential combustion*
3. *Industrial combustion*
4. *Production process*
5. *Road transport*
6. *Other mobile sources and machinery*
7. *Waste treatment and disposal*

2.4.5 Nitrogen dioxide

Nitrogen is a constituent of both the natural atmosphere and of the biosp
When industrial processes release nitrogen to atmosphere it is consider
'pollutant' because of its chemical form (NO, NO_2, and N_2O). These NO_x ca
toxic to humans, to biota, and they also affect the chemistry of the g

(WHO, 2000b). Table 2.6 shows typical levels of lead in air levels for selected sites and countries.

Table 2.6 Summary of lead measurements ($\mu g/m^3$)

Country	*Annual concentration*	*Source*
Canada[a]	0-0.63	Environment Canada, 2000
Costa Rica[b]	0.2-0.5	Ministerio de Planificacion, 2000
Singapore (1998)	0.1-0.2	Singapore Government, 2000
UK[c]	0.004-1.43	DETR, 2000b
US[d]	0.02-0.2	USEPA, 2000
Venezuela[e]	0.4-0.9	MARNR[f], 1997

a) Range of annual average levels for all NAP stations and cities for 1998
b) 1995 – 1997 San Jose (sample duration not known)
c) Annual average levels for 1999 for UK monitoring sites
d) New York, average for 6 sites
e) San Cristobal, Valencia and Puerto Ordaz (1997)
f) Ministerio del Ambiente y de Los Recursos Naturales Renovables

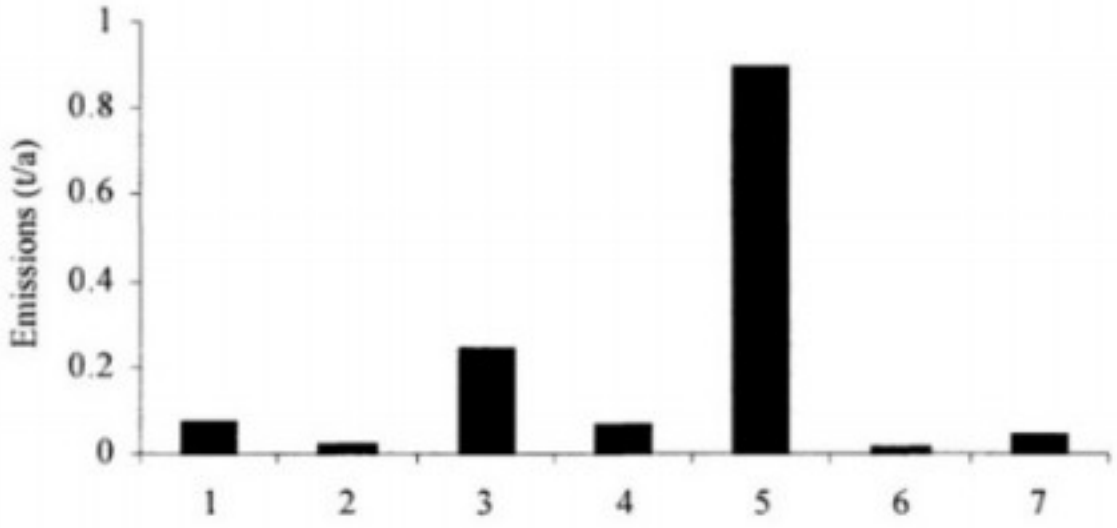


Figure 2.8 UK emission sources for lead in air (Data from Goodwin *et al.*, 1999; SEPA, 2000)

1. *Public power, co-generation and district heating*
2. *Commercial, institutional and residential combustion*
3. *Industrial combustion*
4. *Production process*
5. *Road transport*
6. *Other mobile sources and machinery*
7. *Waste treatment and disposal*

2.4.5 Nitrogen dioxide

Nitrogen is a constituent of both the natural atmosphere and of the biosphere. When industrial processes release nitrogen to atmosphere it is considered a 'pollutant' because of its chemical form (NO, NO_2, and N_2O). These NO_x can be toxic to humans, to biota, and they also affect the chemistry of the global

measures will be a Directive setting National Emission Ceilings for 2010; a Directive on the Sulphur Content of Liquid Fuels setting a maximum permissible sulphur content of heavy fuel oil of 1% from 2003 and gas oil (0.1% from 2008) and a LCP Directive. In addition, further reductions have been agreed in emission levels for motor vehicles as part of the Auto-Oil programme. This programme has also led to an agreement to reduce the fuel sulphur content from 2000 and 2005.

Figure 2.14 Trend of total sulphur emissions (thousand short tons) for the US (Data from USEPA, 2000)

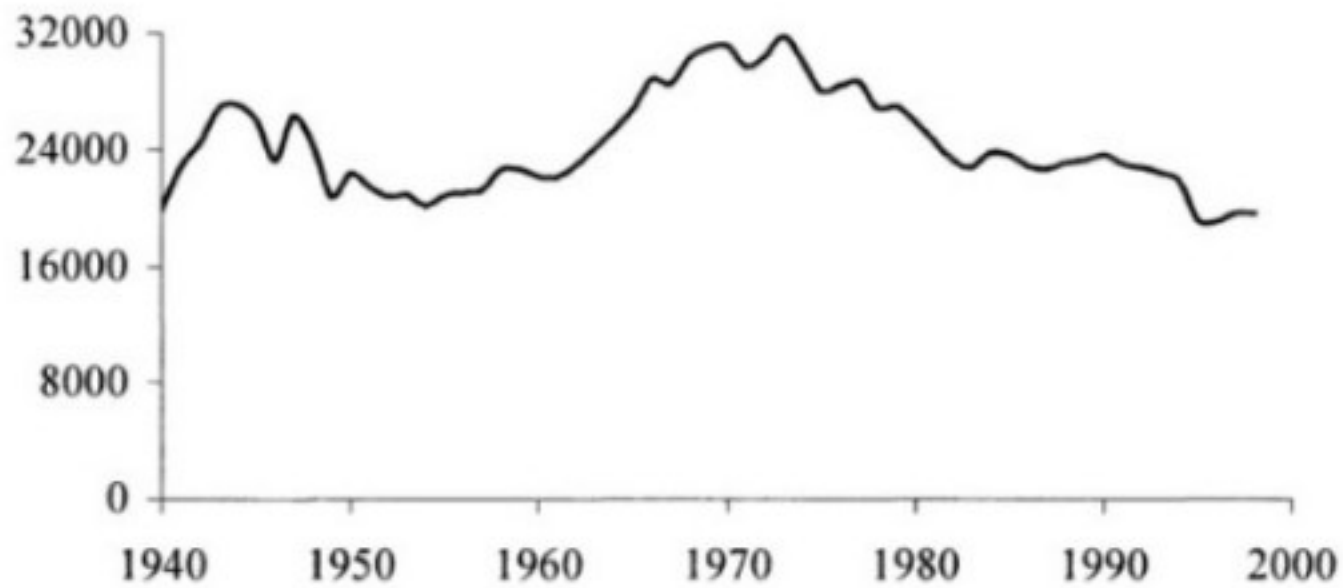
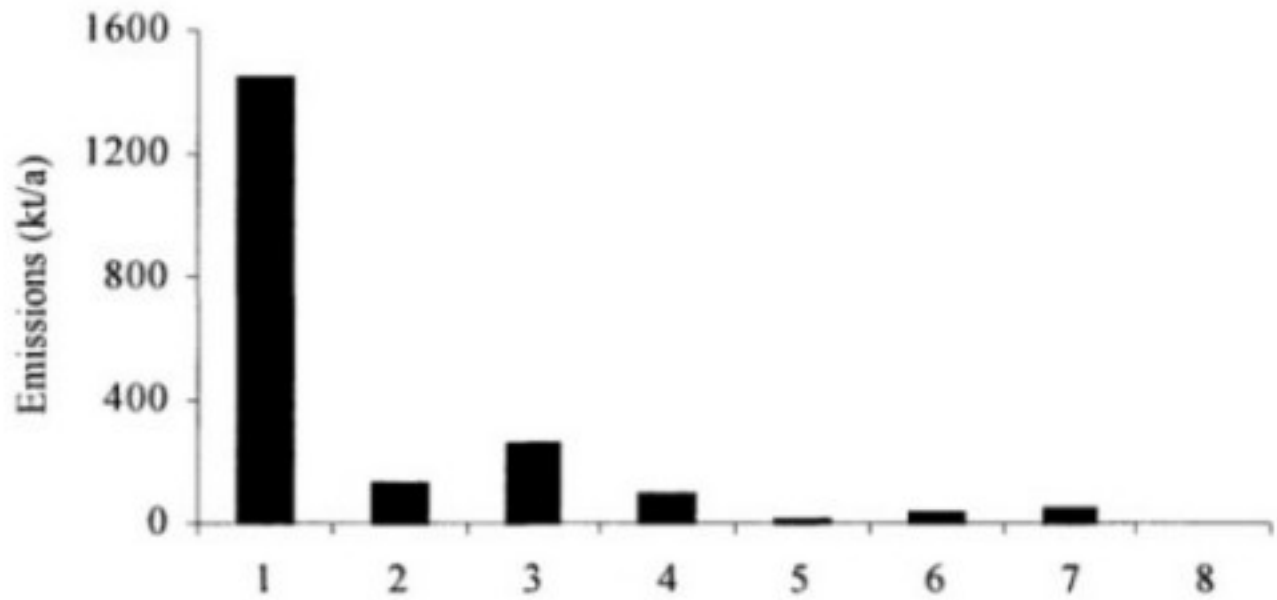


Figure 2.15 UK emission sources for SO_2 (Data from Goodwin *et al.*, 1999; SEPA, 2000)



1. *Public power, co-generation and district heating*
2. *Commercial, institutional and residential combustion*
3. *Industrial combustion*
4. *Production process*
5. *Extraction and distribution of fossil fuels*
6. *Road transport*
7. *Other mobile sources and machinery*
8. *Waste treatment and disposal*

Figure 2.10 Trends and comparison of annual averages of NO_2 for Scottish sites (Data from DETR, 2000b)

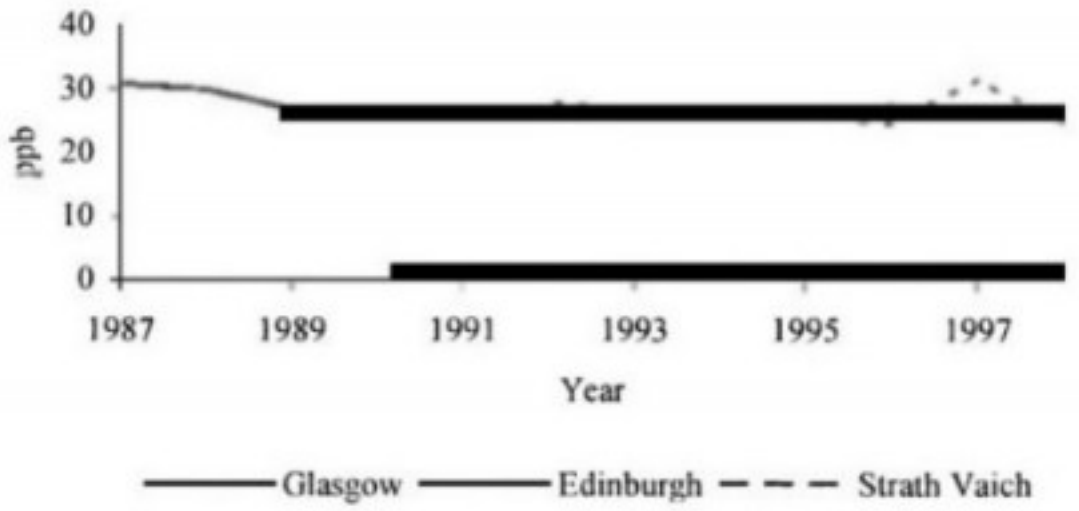
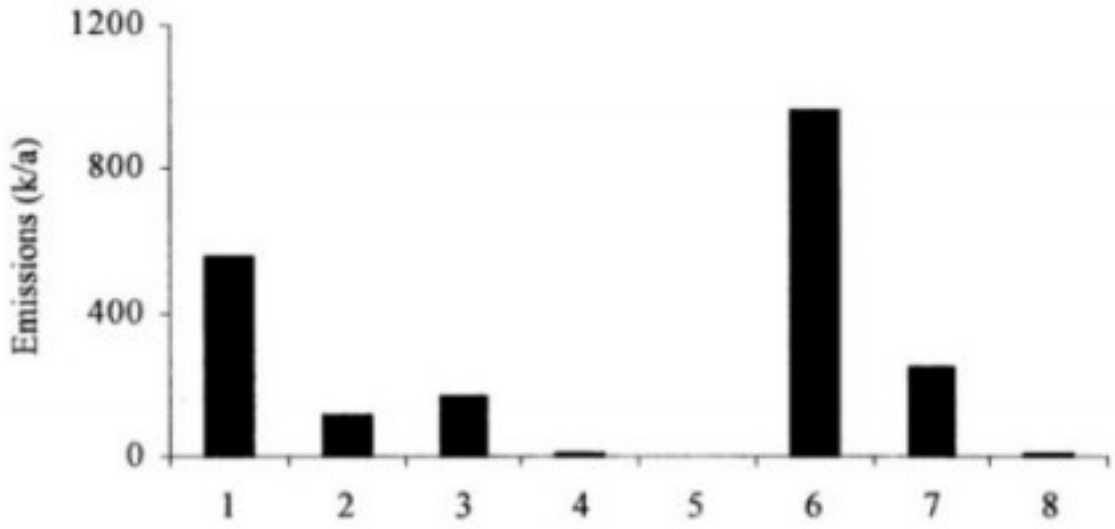


Figure 2.11 Smog in Lima, Peru (courtesy of Cordah Limited)

significant anthropogenic sources as it is a secondary pollutant formed from the reaction between NO_x (the sum of NO_2 and NO) and VOCs in sunlight to form photochemical smog (Photochemical Oxidants Research Group (PORG), 1997). The only significant reaction producing O_3 in the atmosphere is:

$$O + O_2 + M = O_3 + M \qquad (2.3)$$

Figure 2.9 UK emission sources for NO_2 (Data from Goodwin *et al.*, 1999; SEPA, 2000b)



1. *Public power, co-generation and district heating*
2. *Commercial, institutional and residential combustion*
3. *Industrial combustion*
4. *Production process*
5. *Extraction and distribution of fossil fuels*
6. *Road transport*
7. *Other mobile sources and machinery*
8. *Waste treatment and disposal*

Table 2.7 Summary of NO_2 measurements ($\mu g/m^3$)

Country	*Annual concentration*	*Source*
Canada[a]	9.6-59.2	Environment Canada, 2000
Costa Rica[b]	40.3-46.4	Ministerio de Planificacion, 2000
Japan[c]	32.5	Japanese Government, 2000a
New Zealand	5-30 (24 hr) urban 0-1 (24 hr) rural	New Zealand Government, 2000
Singapore (1998)	34	Singapore Government, 2000
UK[d]	5.1-92.0	DETR, 2000b

a) Range of annual average levels for all NAP stations and cities for 1998
b) 1995 – 1997 San Jose (NO_x) (sampling duration not known)
c) Average levels from over 1400 general environmental monitoring stations for 1998
d) Range of annual average levels for UK automated sites for 1999

2.4.6 Ozone

There are two zones of O_3. O_3 in the stratosphere (15-50 km above the earth's surface) forms what is known as the 'ozone layer' and is essential in limiting the level of UV irradiation reaching the earth's surface (Chapter 3, Section 3.8.2). O_3 in the troposphere, the level that contains human life, is the other zone of interest.

Photochemical smog is frequently the most visible form of air pollution (Fig. 2.11) and is of particular concern in urban areas during spring and summer. O_3 is an acidic colourless gas, which acts as a very strong oxidising agent. There are no

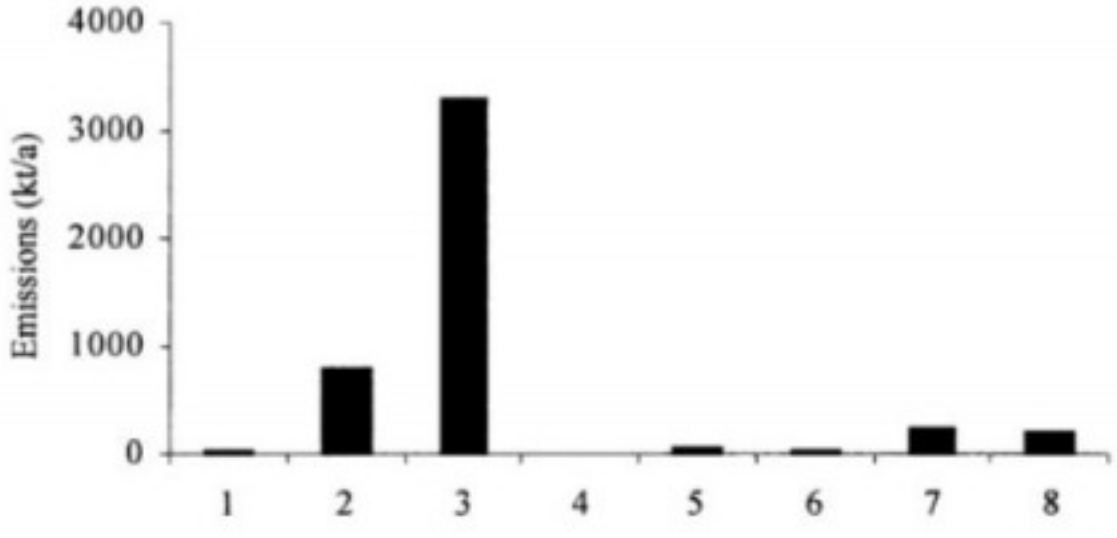
tunnels and in other indoor environments where combustion engines operate with inadequate ventilation. Case Study 2 details the concentrations found in a poorly ventilated indoor area and the mitigation measures used to reduce concentrations to an acceptable level (Chapter 10, Section 10.3). In these circumstances, mean concentrations of CO can reach up to 115 mg/m^3 for several hours (WHO, 2000b). Table 2.5 shows typical levels of CO in air for selected countries.

Table 2.5 Summary of CO measurements

Country	*Concentration*	*Source*
Canada[a,b]	0.1 - 1.2	Environment Canada, 2000
Costa Rica[c]	9.3 - 10.3	Ministerio de Planificacion, 2000
Japan[a,d]	0.6	Japanese Government, 2000b
New Zealand[e]	5-10 urban <0.1 rural	New Zealand Government, 2000
Singapore[a] (1998)	0.6 (2.0)	Singapore Government, 2000
UK[a,f]	0.2-2.1	DETR, 2000b
US[g]	0.08-0.2	USEPA, 2000

a) Average, mg/m^3
b) Range of annual average levels for all NAP stations and cities for 1998
c) 1995 – 1997 San Jose (sampling duration not known) µg/m^3
d) Annual average levels from 145 general environmental monitoring stations for 1998
e) mg/m^3, 8 hour
f) Range of annual average values for UK automated sites for 1999
g) New York for 12 sites units are mg/m^3 (maximum 1 hr)

Figure 2.7 UK emission sources for CO (Data from Goodwin *et al.*, 1999; SEPA, 2000)



1. *Waste treatment and disposal*
2. *Other mobile sources and machinery*
3. *Road transport*
4. *Extraction and distribution of fossil fuels*
5. *Production process*
6. *Industrial combustion*
7. *Commercial/institutional/residential combustion*
8. *Public power/co-generation/district heating*

IMBALANCED DATASET
NEW DELHI:

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	Severe	239	239
2	Moderate	485	485
3	Very Poor	514	514
4	Poor	534	534
5	Satisfactory	108	108

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	Severe	478	478
2	Moderate	485	485
3	Very Poor	514	514
4	Poor	534	534
5	Satisfactory	432	432

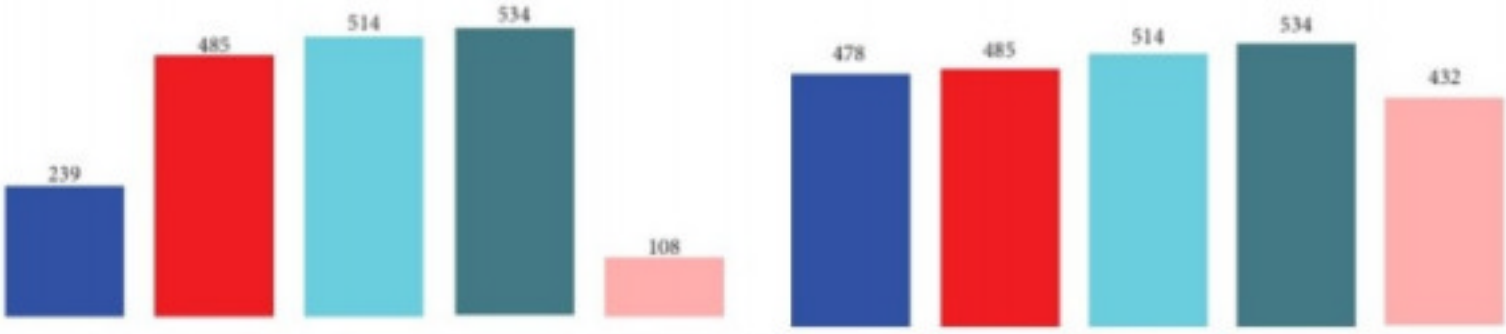


Figure 2: Balanced and imbalanced data values for New Delhi city.

IMBALANCED DATASET
BANGALORE:

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 1 (0%)

No.	Label	Count	Weight
1	Moderate	479	479
2	Satisfactory	810	810
3	Poor	12	12
4	Good	59	59
5	Very Poor	1	1

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 1 (0%)

No.	Label	Count	Weight
1	Moderate	958	958
2	Satisfactory	810	810
3	Poor	768	768
4	Good	944	944
5	Very Poor	1	1

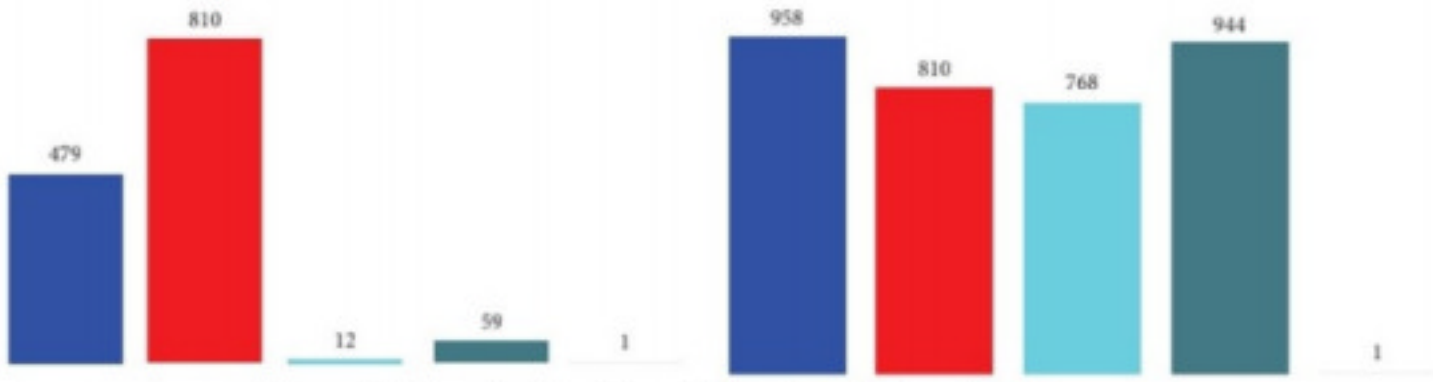


Figure 3: Balanced and imbalanced data values for Bangalore city.

neurons of the brain, gradient boost-techniques utilizing an ensemble of weak prediction models, decision tree-which works by making predictive models using data, and k-nearest neighbor-a lazy learning nonparametric supervised method.

The proposed algorithms used and compared are given below.

4.1. Synthetic Minority Oversampling Technique (SMOTE) Algorithm. Synthetic samples are created for the minority class using this oversampling technique. It aids in making an imbalanced dataset balanced. This approach helps with beating the issue of overfitting brought about by arbitrary oversampling.

4.2. Support Vector Regression. It is a discrete value prediction technique that uses supervised learning. For comparable purposes, SVMs and support vector regression are likewise used. Finding the most appropriate line is the main tenet of SVR. In SVR, the hyperplane with the most points is the line that fits the data the best.

IMBALANCED DATASET
KOLKATA:

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 1 (0%)

No.	Label	Count	Weight
1	Moderate	151	151
2	Satisfactory	278	278
3	Good	119	119
4	Poor	119	119
5	Very Poor	66	66
6	Severe	13	13

BALANCED DATASET

Selected attribute
Name: AQI_Bucket Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

No.	Label	Count	Weight
1	Moderate	302	302
2	Satisfactory	278	278
3	Good	238	238
4	Poor	238	238
5	Very Poor	264	264
6	Severe	208	208

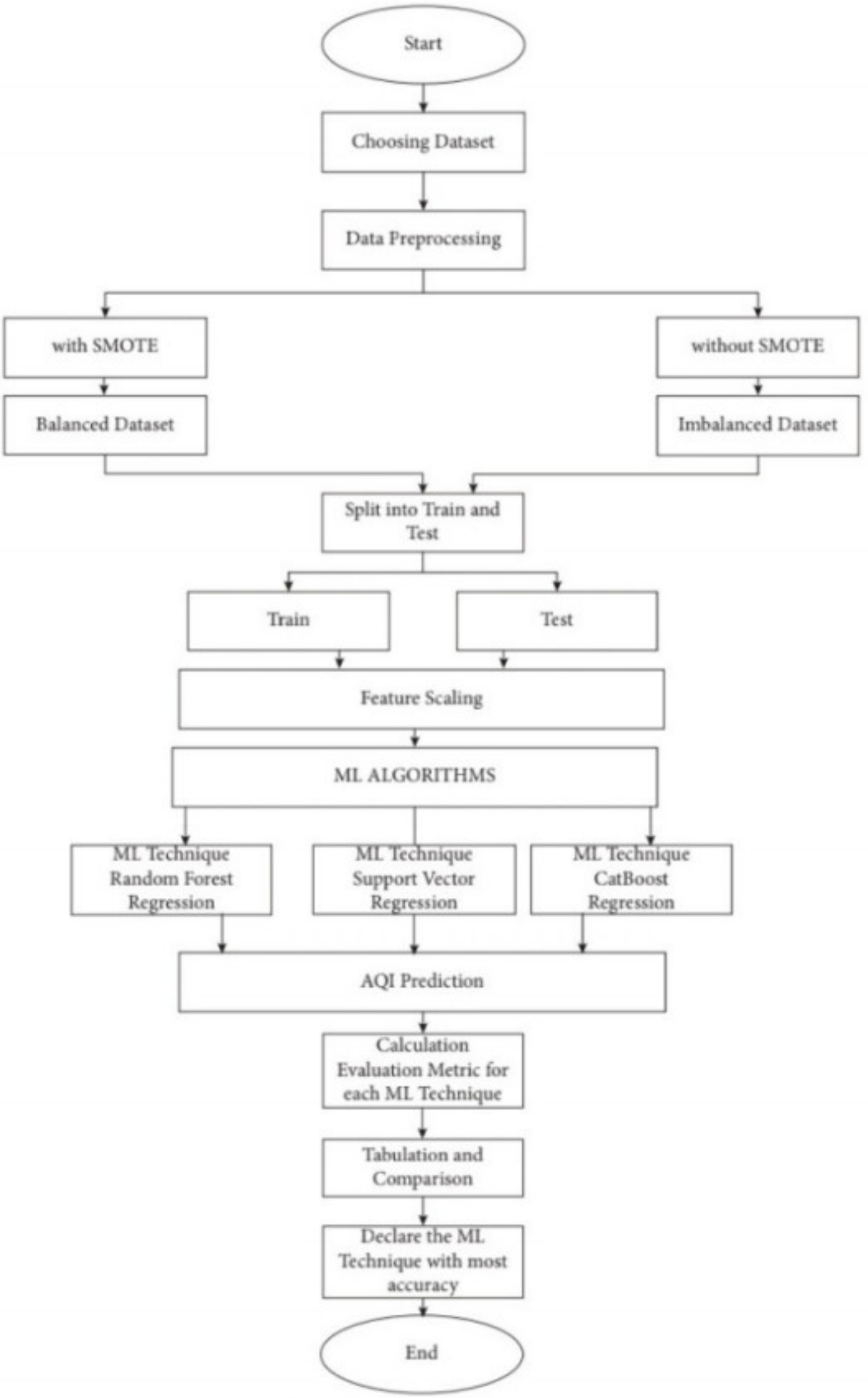


Figure 6: Flowchart for the proposed methodology.

to the Bangalore, Kolkata, Hyderabad, and New Delhi datasets and compare their accuracies to figure out what best fits our use case.

The picked algorithms have the highest accuracy based on our extensive literature survey as logged in Table 1, used for the AQI prediction. The algorithms being used for prediction are support vector regression (SVR), random forest regression (RFR), and CatBoost regression (CR). These algorithms will be provided with a suitably large dataset of cities, such as New Delhi, Bangalore, Kolkata, and Hyderabad, and will provide a practical environment.

The dataset used will be cleaned, reduced, and prepared according to our requirements and the data will be split into training and testing data. The plan is to use the simplest, most

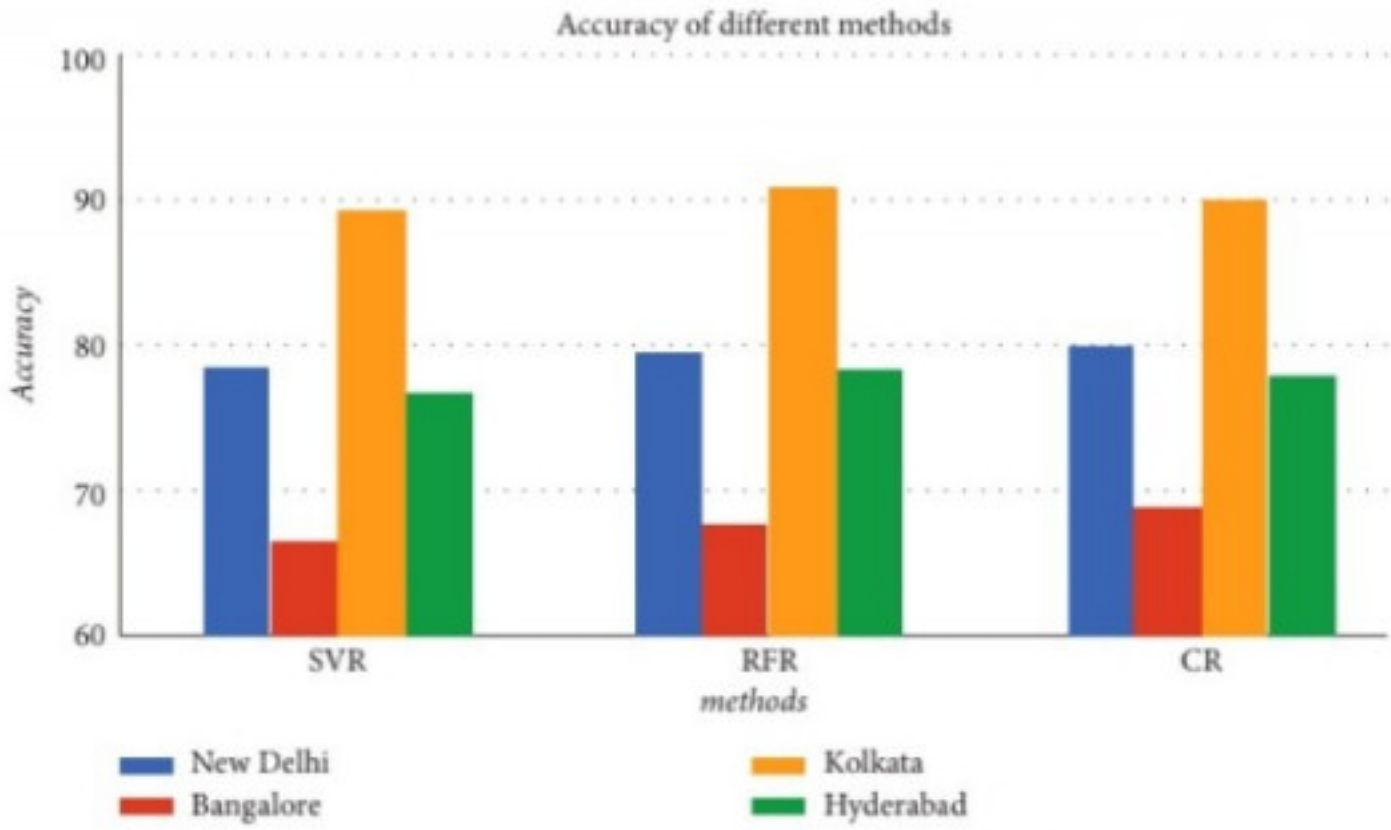


Figure 7: Accuracy comparison of algorithms for four cities.

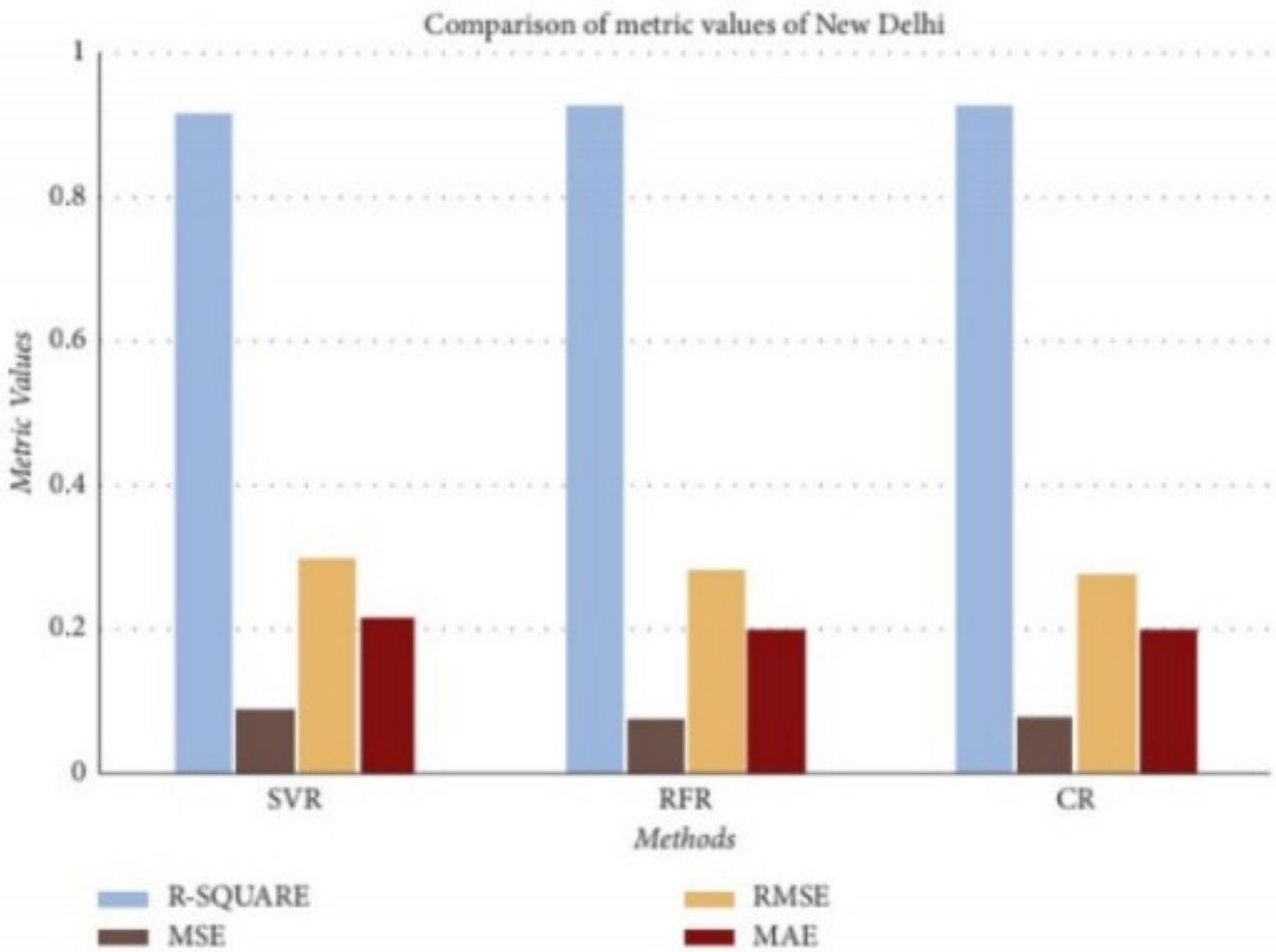


Figure 8: The comparison between R-square, MSE, RMSE, and MAE of support vector regression, random forest regression, and CatBoost regression of the New Delhi city imbalanced dataset.

straightforward implementation in order for the algorithms to be applied easily in a real-life use case. Then, different parameters will be taken to finalize and draw up a comparison between these 3 algorithms and then come to the conclusion to show which is the most accurate. The comparison can bring out important information about AQI prediction methods and even help us choose the most suitable one. A comparison of the accuracy levels obtained with an imbalanced dataset and a balanced dataset with the help of the SMOTE algorithm will also be done.

Hence, the methodology is a step-by-step process in which the first step is to find a suitable dataset and clean it. After this, further data preprocessing is applied which makes use of SMOTE in order to balance the dataset. Both balanced and imbalanced datasets will be preserved and used in order to bring to light any differences in performance that may arise due to balancing. Following this, in a standard machine learning procedure, the dataset is split into train and test to train the models and test their accuracies against real data. Feature scaling and normalization are carried out.

Now, each regression model which has been picked, namely, random forest, support vector regression, and CatBoost, are used for prediction and its accuracy is gauged, for each balanced and imbalanced dataset as mentioned previously. They are compared using metrics such as RMSE and R-SQUARE. Finally, all the data and results have been

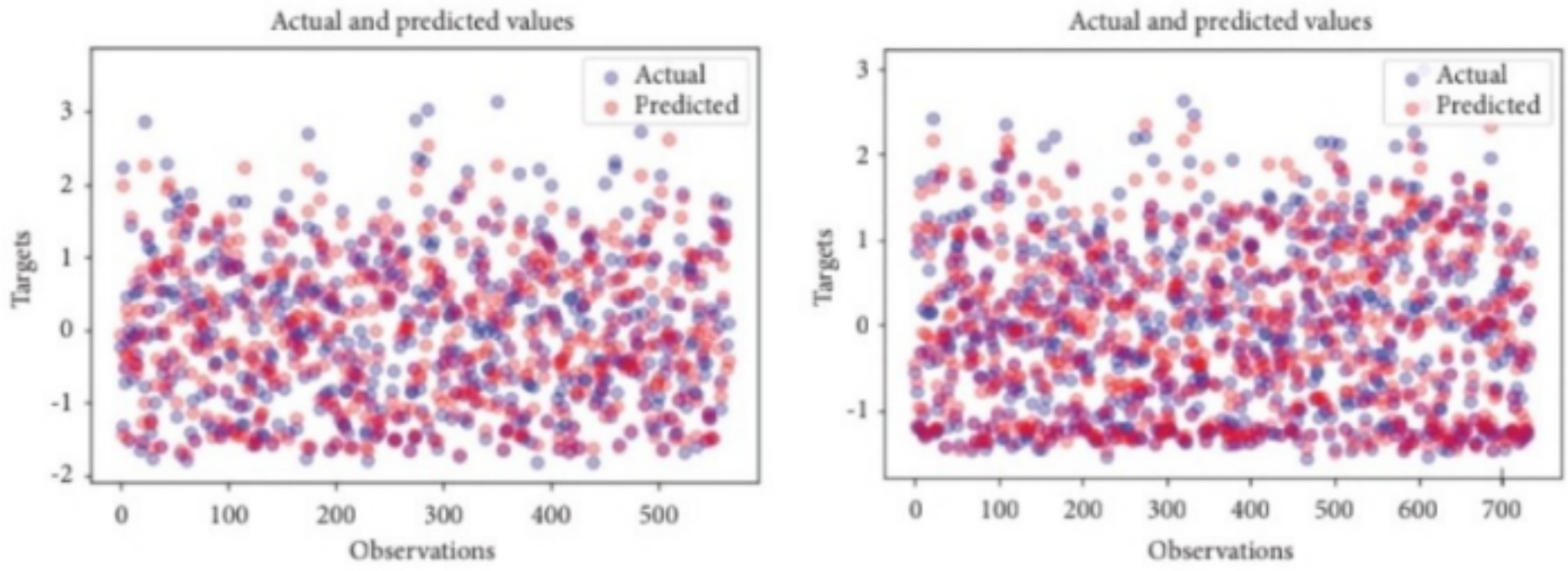


Figure 15: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi-SVR.

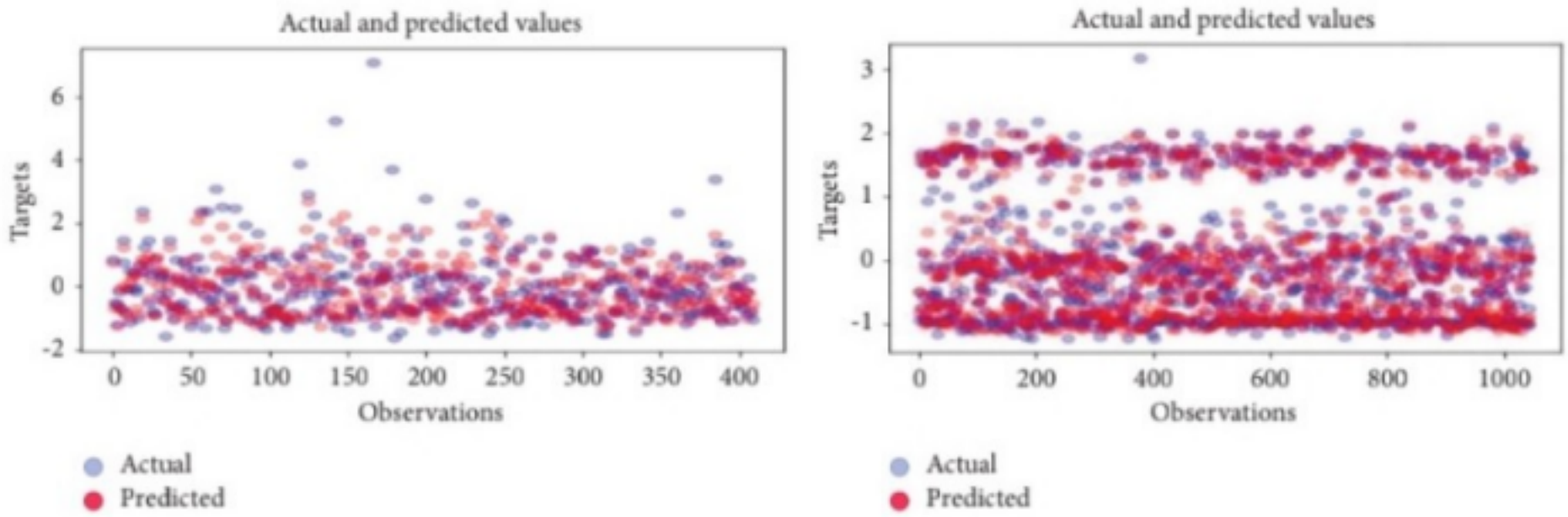


Figure 16: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore-SVR.

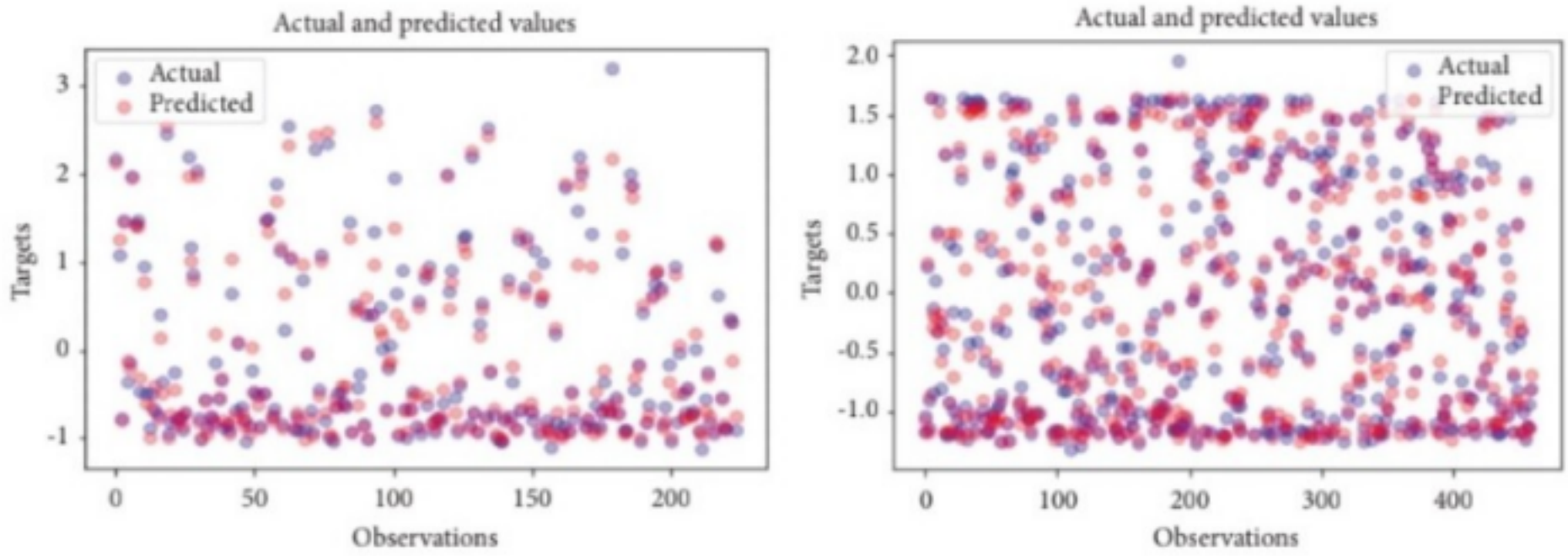


Figure 17: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata-SVR.

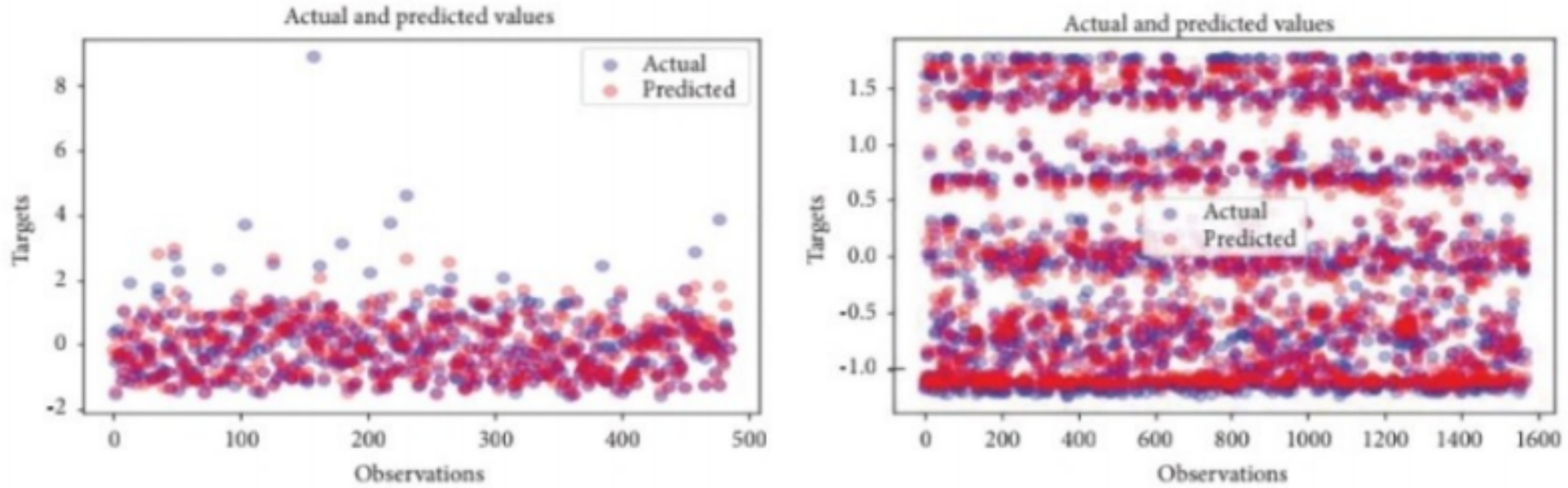


Figure 18: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad-SVR.

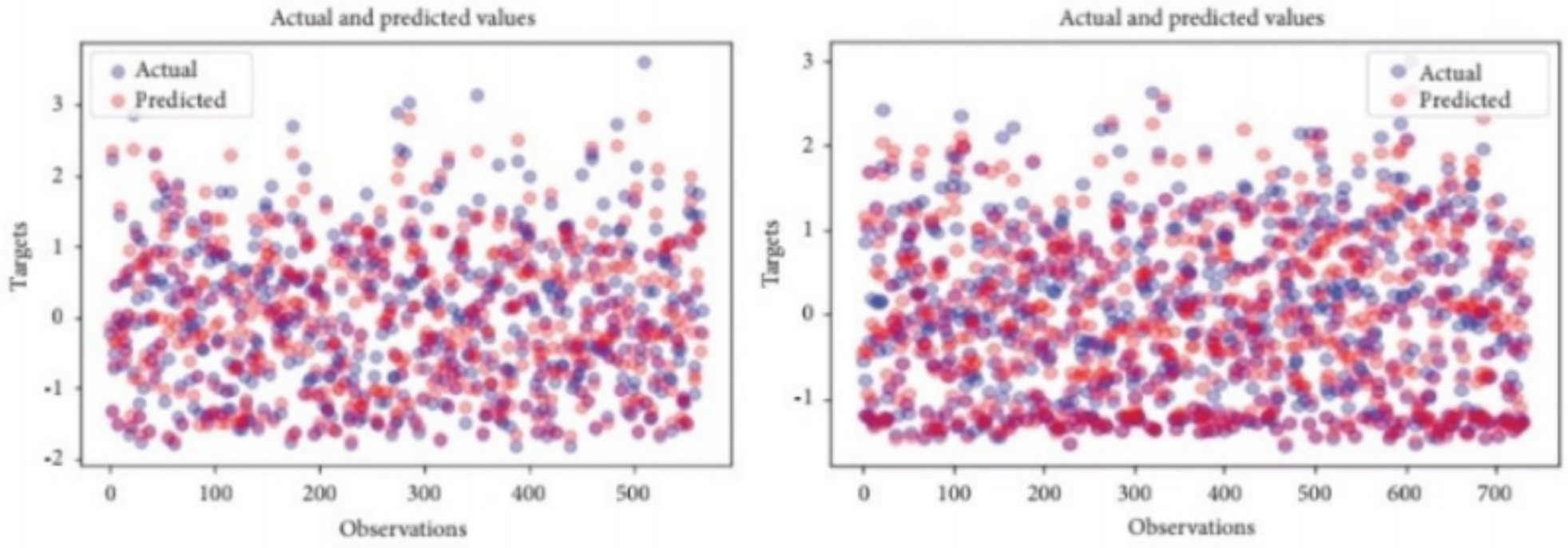


Figure 19: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi-RFR.

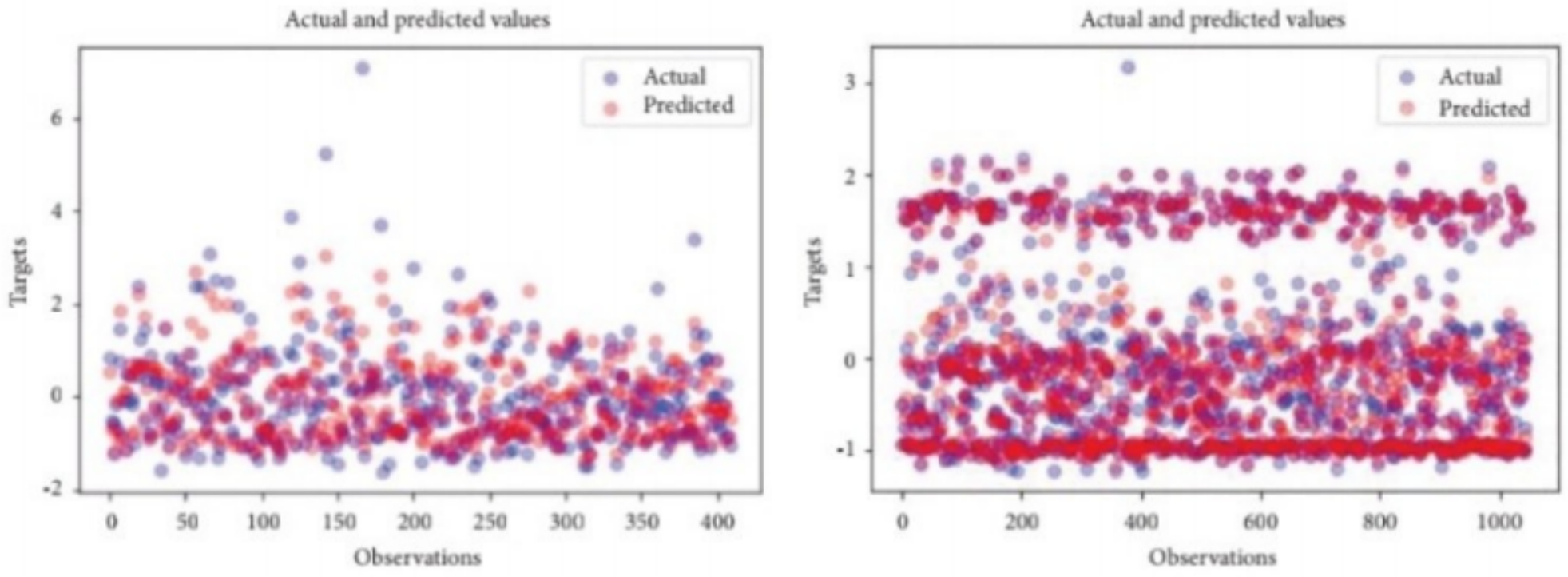


Figure 20: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore-RFR.

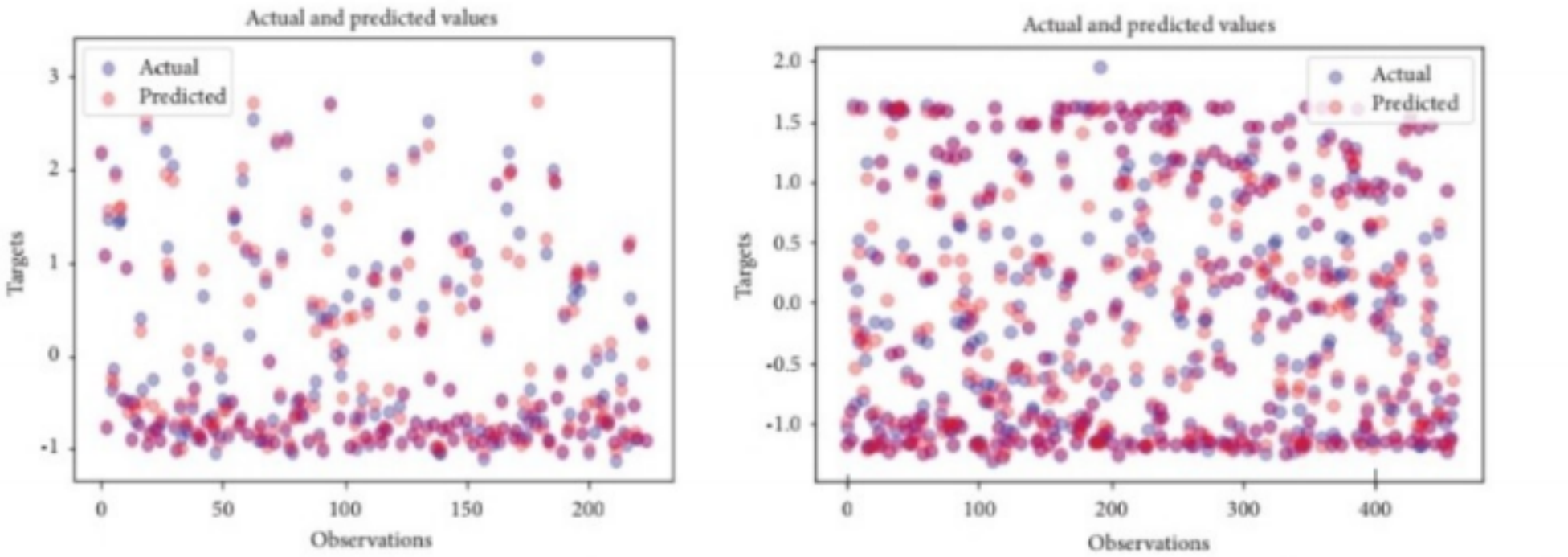


Figure 21: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata–RFR.

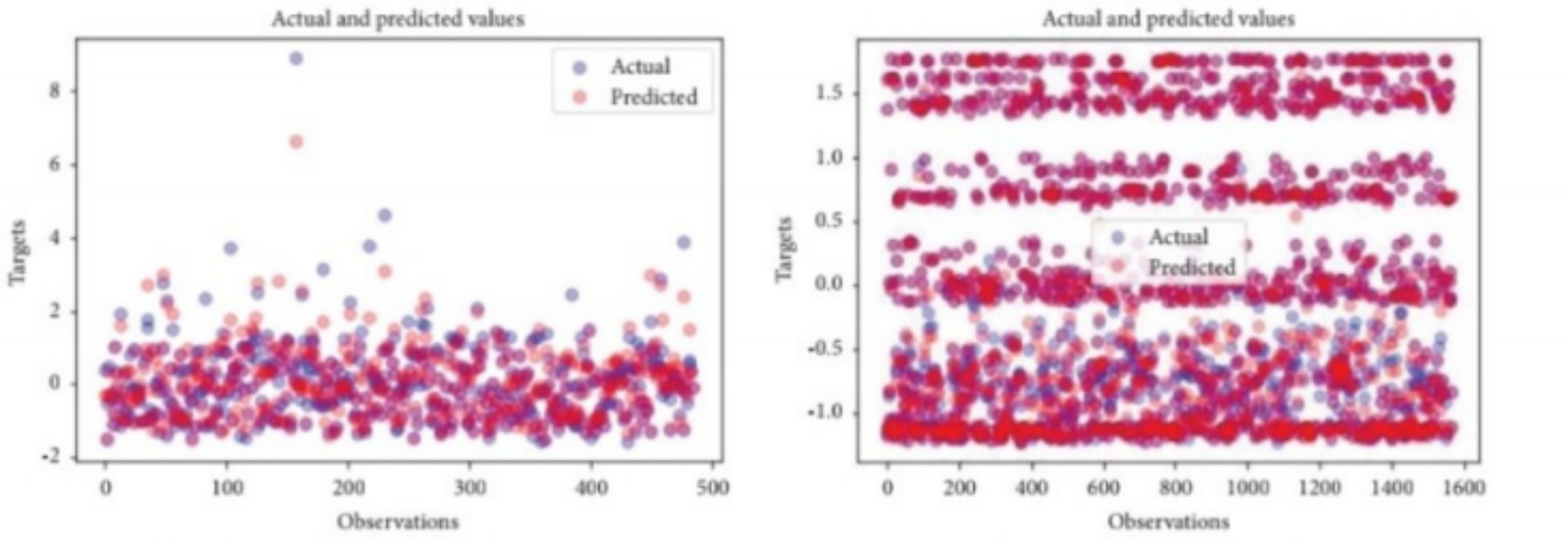


Figure 22: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad–RFR.

range of 0.2–0.5 demonstrate that the model can reasonably predict the data. It is shown in the equation

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(X_i - X_i\hat{})^2}{m}}, \tag{3}$$

where

(a) x_i = The i^{th} observed value
(b) $x_i\hat{}$ = The corresponding predicted value
(c) n = The number of observations

(iv) MAE evaluates the absolute distance of the observations to the predictions on the regression line. It is shown in the equation

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{n} |X_i - X|, \tag{4}$$

where

(a) n is the number of errors
(b) Σ is the summation symbol (which means "add them all up")
(c) $|x_i - x|$ is the absolute errors

(v) Accuracy is used as a measurement to calculate how well a model is finding patterns and identifying relations in the dataset and it is shown in the equation

$$\text{Accuracy} = (1 - MAE) * 100. \tag{5}$$

This gives the accuracy in percentage.

6. Results and Discussion

In the proposed work, the dataset mentioned above has been cleaned such that it only has the values for the cities of New Delhi, Bangalore, Kolkata, and Hyderabad. The dataset was used in two ways, once in an imbalanced version and then in a balanced version using SMOTE. Graphs were plotted and it was seen that there was an increase in the accuracies of the models which had the balanced dataset. For prediction purposes, three algorithms were run on it, namely, support

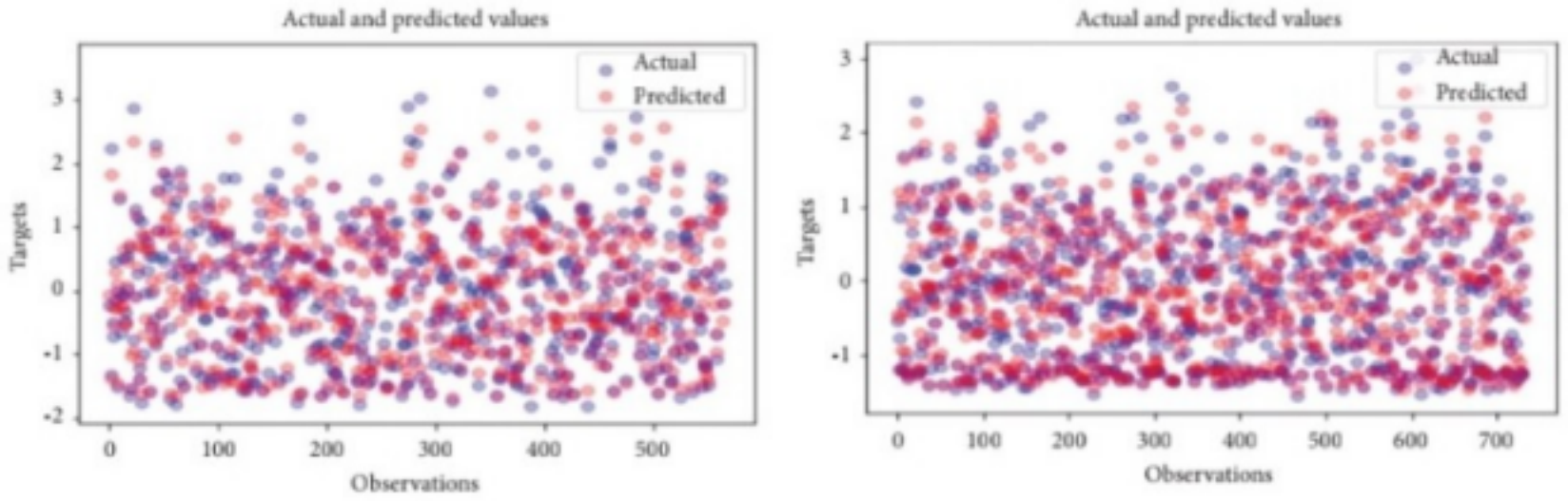


Figure 23: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for New Delhi–CR.

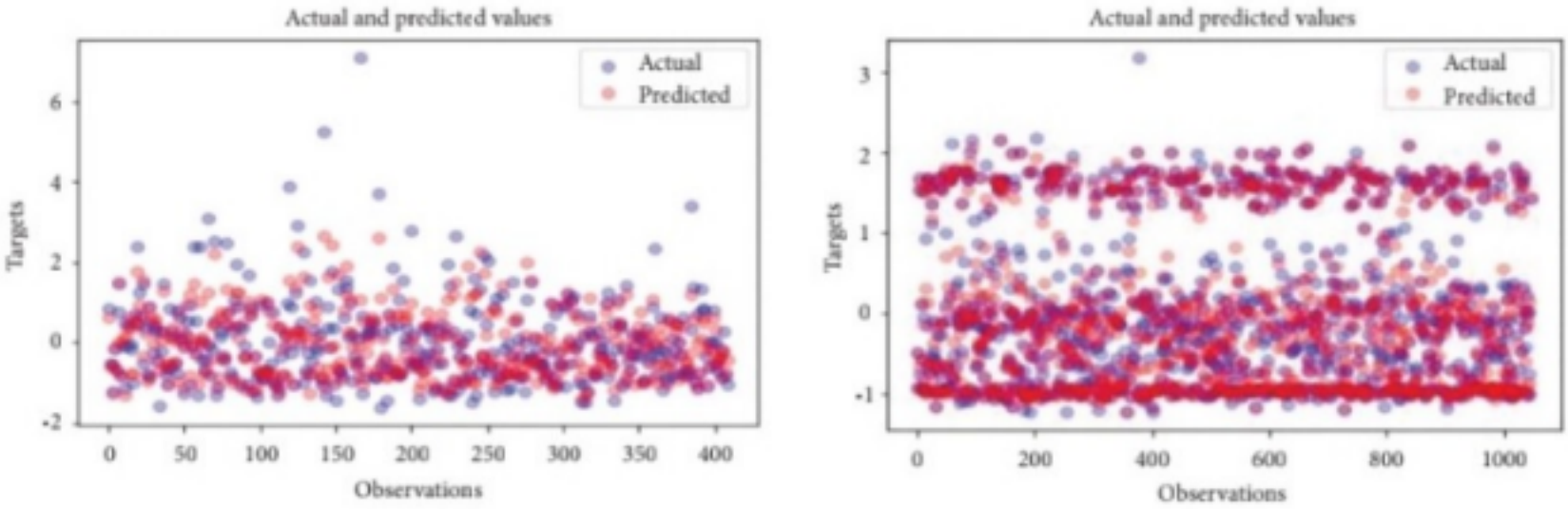


Figure 24: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Bangalore–CR.

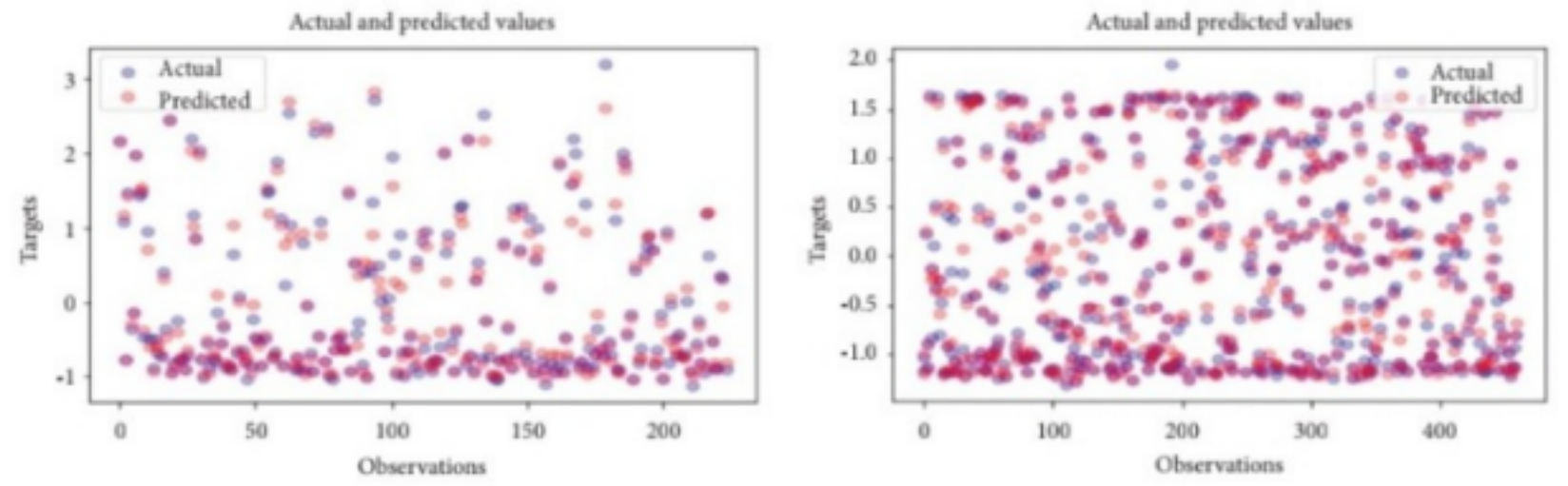


Figure 25: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Kolkata-CR.

vector regression, random forest regression, and CatBoost regression. Plotted graphs between the test data and the predicted data were shown as well. The metrics calculated in each algorithm are R-SQUARE, mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). Comparative tables, graphs and scatter plots were drawn for balanced and imbalanced dataset results to show how using a balanced dataset when used provides higher accuracies in each algorithm.

According to the research in this paper, the choice to use statistical metrics, such as RMSE, R-SQUARE and so on, has been understood and referred to in papers [30–33], as well as how to effectively implement them. Metrics are used to track and gauge a model's performance (during training and testing). These metrics provide information on the precision of the forecasts, and the amount of departure from the actual values since all of the algorithms utilized are based on regression models.

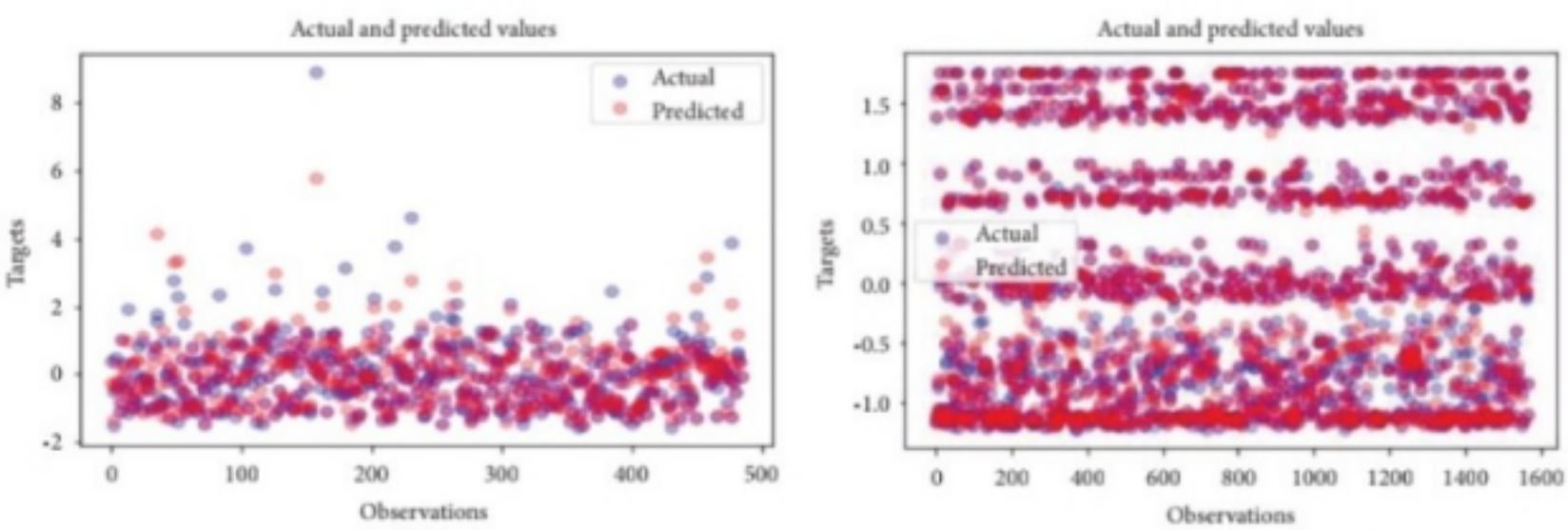


Figure 26: Scatter plots showing actual and predicted values for the imbalanced dataset (without using SMOTE) and balanced dataset (with using SMOTE) for Hyderabad–CR.