Abstract

Estimating the air quality Index (AQI) is one of the crucial topics of air quality research today as it is valuable to evaluate the impacts of air contaminations on human wellbeing in metropolitan regions. The goal of this research is to combine Principal Component Analysis (PCA) and Artificial Neural Network (ANN) approaches in order to obtain accurate findings in determining the Air Quality Index (AQI). This research is focused on the city of Delhi. The principal components score (PCs) of 11 historical air quality and meteorological factors are calculated and used as input variables in ANN models to predict AQI. For techniques to estimate the AQI, ANN and Multiple Linear Regression (MLR) models are used and compared. The other method is to decrease the eleven parameters to eight PCs using PCA. As a result, the eight PCs are fed into ANN (PC-ANN) models as input data. Coefficient of Determination (R2), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE) were used to compare the method’s performance. Given the complexities of the problem of air pollution, When compared to the other models, the PC-ANN model provides the best result. This suggests that the PC-ANN method can be used to make better decisions and solve problems in the field of atmospheric management.

Introduction

Air pollution in India is a growing concern and for the city of Delhi, it is a very serious problem. Energy production from power plants, industry, domestic heating, fuel-burning cars, natural disasters, and other factors are common causes. One of the most serious short-term impacts of air pollution, particularly in metropolitan areas, is human health concerns. The greenhouse effect and global warming are two long-term effects on the global climate. This indicates that perhaps the air we breathe is not pure, but rather polluted, as it contains several hazardous substances and particles that are harmful to human health. The National Ambient Air Monitoring Network collects data on the concentrations of various contaminants in the air, however, this data is difficult to comprehend for the general public. As a result, the Central Pollution Control Board (CPCB) creates a national Air Quality Index (AQI) for Indian cities. The AQI indicates the air quality or the extent to which the air in a certain location is contaminated. This means that AQI provides the actual quality of the air around us in a qualitative form, which is linked to a variety of health effects. Monitoring concentrations of pre-determined air pollutants in residential/commercial/industrial areas are used to compute an air quality index (AQI). Various methods are used to aggregate and convert monitoring data into a single index. As a result, indexing systems and air pollution descriptors frequently differ from one country/region to the next. The air quality indicators allow the public to track the status of their local, regional, and national air quality without having to understand the details of the monitoring data on which they are based.

Air Pollutants Data

Daily averaged Air quality data of Sulphur dioxide (SO2), nitrogen dioxide (NO2), Nitrogen Monoxide (NO), ozone (O3), benzene, toluene and respirable particulate matter (PM2.5) for a period 2013 -2021 at NSIT Dwarka (One of the well know and prominent places in New Delhi) Obtained from central pollution control board (CPCB), Delhi has been used in this study, Apart from these pollutants other parameters such as relative humidity (RH), Wind Speed (WS), Wind Direction (WD) and Solar radiation (SR). PM10 is also one of the major air pollutants criteria which is not considered in this study as sufficient data of PM10 was not available for the period 2013 – 2021.

Meteorological Data

The daily averaged surface meteorological variables such as average temperature (T), Maximum temperature (TM), Minimum temperature (Tm), Atmospheric pressure at sea level (SLP), Average relative humidity (H), Total rainfall and / or snowmelt (PP), Average Visibility (VV), Average windspeed (V), Maximum sustained windspeed (VM) Observed at Safdarjung airport station which is the closest station to the selected air pollutant station in Delhi, The data has been acquired from Indian Meteorological Department (IMD) for the years 2013 – 2021. Safdarjung airport is about 20 kms away from NSIT Dwarka this is the closest station we could find.

Literature Review

Anikender Kumar, PramilaGoyal presented the study that forecasts the daily AQI value for the city of Delhi, India using the previous record of AQI and meteorological parameters with the help of Principal Component Regression (PCR) and Multiple Linear Regression Techniques. They perform the prediction of daily AQI of the year 2006 using previous records of the year 2000-2005 and different equations. After that, this predicted value was compared with an observed AQI of 2006 for the season’s summer, Monsoon, Post Monsoon, and winter using the Multiple Linear Regression Technique [1]. Principal Component Analysis is used to find the collinearity among the independent variables. The Principal components were used in Multiple Linear Regression to eliminate collinearity among the predictor variables and also reduce the number of predictors [1]. The Principal Component Regression gives a better performance for predicting the AQI in the winter season than any other season. In this study, only meteorological parameters were considered or used while forecasting the future AQI but they have not considered the ambient air pollutants that may cause the adverse health effects.

Eman Sarwat and Ghada I. El-Shanshoury’s introduced the study to merge the Principal Component Analysis (PCA) with Artificial Neural Network (ANN) techniques to get the precise results in estimating the Air Quality Index (AQI). This study is applied for Ain Sokhna city for the year 2014. PCA and ANN are merged (PC-ANN) to predict AQI based on 10 historical air quality and meteorological parameters. The feed-forward ANN model, using ten original parameters as inputs, gives a high value of R2 and low values of error rates that contrast to results from the MLR model (Method 1). However, Method 2 (PC-ANN using Varimax method) gives a better prediction of the results as compared to Method l in terms of R2 value and error rate values. Furthermore, the relation between Varimax PC-ANN prediction results and actual values are almost matching. The prediction performance of the Methods 3 and 4 (PC-ANN using Equamax and Quartimax methods, respectively) gives lower effectiveness than Methods 1 and 2, but these methods can predict the AQI within acceptable accuracy when compared with MLR and PCR models. The results proved that the merge of Varimax rotated PC scores with the ANN model is more efficient and precise in the estimation of AQI results. Moreover, using ANN or PC-ANN model gives more accurate estimation results than MLR or PCR. Whereas it also proves that, the ANN and PC-ANN models are very helpful tools in making a decision and solving problems for better atmospheric management of the local environment.

Methodology

Data Pre-Treatment

In this study, five air pollution criteria and six meteorological factors are used. In the preceding part, we discussed air contaminants and meteorological parameters. The original dataset had a total of twenty two parameters including both the pollutants and the meteorological parameters, out of these, Eleven parameters were chosen based on their importance which was found using the correlation matrix. Furthermore to account for the missing values all the rows containing the missing values were dropped.

Chart, treemap chart

Description automatically generated

**Correlation Heatmap**

Chart, histogram

Description automatically generated

**Feature Importance based on correlation coefficients**

AQI Calculation

The generation of sub-indices for each pollutant and the breakpoints (aggregation) of sub indices are the first two processes in calculating the AQI. Each pollutant's breakpoint concentration is based on the Indian NAAQS and epidemiological research that show the danger of adverse health consequences from specific contaminants. Different breakpoint concentrations and air quality standards have been recorded in the literature, which has been recognised (EPA, 1999). In India, a range of index values has been suggested to describe the status of air quality and its consequences on human health which has been mentioned in the table below.

Table

Description automatically generated

**AQI Category for Ozone and Particle pollution**

All the values of NO, NO2, PM2.5 are in (ug/m3). Good: air quality is adequate; but, for some contaminants, a small percentage of people may have a moderate health concern. Moderate: Members of vulnerable groups may experience negative health consequences. Poor: Members of vulnerable groups may suffer more serious health consequences. Very poor: causes a health alert; everyone may suffer from more significant health consequences. Severe: health warnings of emergency situations are triggered. The formula of AQI is given:

Text

Description automatically generated with medium confidence

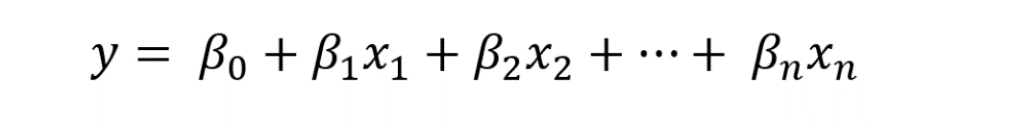
where Ip is the AQI for pollutant 'p,' Cp is the actual ambient concentration of pollutant 'p', BPHi is the breakpoint that is more than or equal to Cp, BPLo is the breakpoint that is less than or equal to Cp, IHi is the subindex value for BPHi, and ILo is the subindex value for BPLo. Each pollutant's AQI (NO, NO2, PM2.5, and SO2) has been determined separately, with the highest among them announced as the AQI of the day, which is employed as one of the input parameters in the statistical models.

Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) approach is used in factor analysis. PCA's main goal is to create a new set of orthogonal and uncorrelated composite variates to account for the overall variance among the 'n' number of subjects (variables) in p-dimensional space. The new set of variates is made up of linear combinations of the original measurements. The linear combinations are created in such a way that each composite variate accounts for a lesser share of the overall variation than the one before it. The variance of the first composite (principal component) will be the highest; the variance of the second will be lower than the first but higher than the third, and so on. If the first few principal components (or, eigenvector-eigenvalue pairs) account for more than 60% of the total variance, there is no need to construct the composite (overall) Air Quality Index with more principle components (PCs). Higher order PCs only explain a small portion of overall variance and are hence considered noise. In our case to account for the maximum variation in the data we had to select the first 8 Principal Components (PCs)

Multiple Linear Regression

In atmospheric modelling, multiple linear regression (MLR) is widely used. By fitting a linear equation to observed data and calculating the percentage contribution of each parameter to atmospheric pollution, this technique has been used to study the link between multiple independent and dependent variables. It is utilised to support the association between air quality and meteorological factors (11 parameters in our case) as independent variables and total AQI data as a dependent variable in this study. The model is obtained using equation:



where,

y is the response variable, x1, x2, xn are the explanatory variable, β0, β1, β2, …, βn are regression coefficients.

Principle Component Regression

PCA (Principle components analysis) and OLS (ordinary least square method) are combined in Principal Component Regression (PCR). Principle components analysis (PCA) is used to breakdown the independent (x) variables into an orthogonal basis (the principal components) and pick a subset of those components as the variables to predict y in principal components regression (PCR). The primary idea underlying PCR is to compute the principal components and then utilise some of them as predictors in a linear regression model fitted using the traditional least squares approach. By employing PCR, it is possible to perform dimensionality reduction on a large dataset and then fit a linear regression model to a reduced collection of variables while retaining the majority of the original predictor’s variability.

Artificial Neural Network

Artificial neural networks (ANNs) have been effectively applied in a variety of problem fields for classification, prediction, and association. ANNs can approximate any nonlinear mathematical function, which is particularly beneficial when the relationship between the variables is unknown or complicated. The multilayer perceptron (MLP), a feed forward network that may use several methods to minimise the objective function, such as backpropagation, conjugate gradient, and others, was examined in this study. The architecture of an MLP ANN is depicted in the diagram below. An ANN's input layer consists of n input units with values xi, I =1, 2, n, and randomly generated initial weights wi, which are commonly in the range [-1,1]. The weighted sum of all xi values is fed into each unit in the hidden (middle) layer. The hidden layer's output, represented as yc, is calculated by adding the inputs multiplied by their weights, as shown in Equation:

A picture containing text

Description automatically generated

where f is the user-selected activation function (sigmoid, tangent hyperbolic, exponential, linear, step or other).

Diagram

Description automatically generated

**ANN Architecture**

In our case we have used a 10 layered Deep feed forward neural network with rectified linear unit as the activation function. The nature of the problem under investigation determines the number of input and output neurons. In this study, the networks are trained, tested and validated with Eights hidden layers and 1024 neurons in the input layer and the neurons in subsequent layers are half the neurons from the layer preceding it. The output neuron (layer) gives the predicted AQI value which gives the best results.

Applying PCA in conjunction with Artificial Neural Networks (PC-ANN)

The goal of using a neural network with PCA is to increase the neural network model's performance. PCA is a multivariate statistical tool extensively used for analysing air pollution data. PCA's purpose is to reduce the number of predictive factors and convert them into new variables known as principle components. These new variables are created by combining original data into separate linear combinations that maintain the highest potential variation. PCA also reduces dataset collinearity, which leads to the worst air pollution concentration forecasts (SOUSA et al., 2007). The PCA-neural network contains fewer input variables than a neural network, which minimises the architecture and computation complexity. The PCs can be calculated using the normalised input data's correlation matrix. The eigenvalues of the correlation matrix 'C' are calculated using the following characteristic equation:



The eigenvalue is **λ**, and the identity matrix is I. A non zero eigenvector e exists for each eigenvalue **λ**, which can be defined as:



The correlation matrix yields the eigenvectors, which have mutually orthogonal linear combinations. The entire amount of variances explained by each of the eigenvectors is represented by their associated eigenvalues. A significant portion of the overall variance can be explained by keeping the top few pairs of eigenvalue–eigenvector or principal components. Noise can be defined as higher order primary components that account for a small portion of the total variance. The ith PC's variance is stated as:

Text

Description automatically generated

The linear combination of the variables, which accounts for the most overall variability in the data, is represented by the first PC1 Which is associated with the greatest eigenvalue. The second PC2 accounts for the most variability, which the PC1 does not account for. The preceding PCs also follow a similar pattern. The initial data set is turned into the orthogonal set by multiplying the eigenvectors after obtaining all of the PCs.

Diagram

Description automatically generated

PC-ANN Architecture

Model Performance Determination

To create exact predictions, four separate criteria were utilised to evaluate the usefulness and performance of MLR, PC-ANN, and PCR models. The following are the criteria that were considered:

Coefficient of Determination (R2), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). In this case, a lower RMSE, MAPE, MAE, and a higher R2 score result in a more accurate prediction.

Graphical user interface

Description automatically generated with medium confidence

Results And Conclusion

Principle Component Analysis

To identify the number of PCs to be used to get accurate predictions and results usually the variance explained by each of the PCs is analyzed, higher the cumulative variance explained by the number of PCs selected better will be the results, In our case we had to choose the first 8 PCs which explained about 99% of the total variation in the data. The remaining higher order PCs are considered as noise and are discarded. A Plot explaining the cumulative variance by all the PCs has been presented in the figure below.

Chart, line chart

Description automatically generated

**Percentage Variance vs Number of Features**

Using MLR and PCR Models to Predict the AQI

MLR is utilised to predict the AQI in this study by justifying the association between air quality and meteorological and pollution factors (11 parameters) as independent variables and total AQI data as a dependent variable. A two-step approach is used to compute PCR. The actual variables are first subjected to a principal component analysis. Following that, the eight producing PCs are used as predictors in an MLR. PCR can show a dataset's inherent linear structure, reducing the number of variables to predict.

ANN and PC-ANN Models for AQI Prediction

For the ANN model 11 original meteorological and pollution factors are used as input, the designed ANN has a total of 8 hidden layers, For the PC-ANN model the original parameters are reduced to 8 principle components using principle component analysis, These 8 PCs are then used as the input to the same ANN model with 8 hidden layers with a single neuron in the output layer for prediction of AQI.

Comparative performance of ANN, MLR, PCR and PC-ANN models

Eleven air pollutant and metrological parameters (original data) are used as input data in ANN and MLR. For the remaining methods the eleven original parameters are reduced into eight PCs using Principle Component Analysis (PCA). Consequently, the generated eight PCs are used as input variables for the ANN and MLR models. RMSE, MAPE, MAE, and R2 were used to compare the model’s performance. The table below shows the comparison of the ANN, MLR, PCR, PC-ANN models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | R2 | RMSE | MAE | MAPE |
| MLR | 0.88641 | 42.8000 | 30.5265 | 0.20237 |
| ANN | 0.99444 | 8.9479 | 4.9945 | 0.02320 |
| PCR | 0.88644 | 41.5212 | 30.6909 | 0.21504 |
| PC-ANN | 0.99691 | 6.9489 | 4.0988 | 0.02531 |

**Model Performance Table**

It is clear from the data above that the PC-ANN model outperforms all the other models in terms of better estimation of the AQI, It can also be seen that in general, models which have been incorporated along principle component analysis are performing better than their base models ( PCR outperforms MLR also PC-ANN outperforms ANN ).

Chart, line chart

Description automatically generated

**Estimated AQI using PC-ANN method Vs Actual value of AQI**

Chart, scatter chart

Description automatically generated

**Estimated AQI using PCR method Vs Actual value of AQI**

Conclusions

PCA and ANN are combined (PC-ANN) in this study to predict AQI using 11 historical air quality and meteorological indicators. There are four different models used, Namely MLR, ANN, PCR, PC-ANN. ANN and MLR models are based on actual raw data where as PCR and PC-ANN models are based on PC scored which are obtained by reducing the original parameters using principle components analysis.

The results show that using Eleven original parameters as inputs, the ANN model produces a high R2 and low error rates, in contrast to the MLR model's outputs. However, in terms of R2 value and error rate values, the PC-ANN model outperforms the ANN, MLR, and PCR models in predicting the results.

Furthermore, when compared to MLR or PCR, utilising an ANN or PC-ANN model yields better accurate estimation results. It also shows that the ANN and PC-ANN models are very useful tools for making decisions and addressing problems in order to improve local atmospheric pollution control.

References

1. Kamal Md. M., Jailani R., & Shauri R. L. A.; Prediction of Ambient Air Quality Based on Neural Network Technique; In SCOReD 2006 -Proceedings of 2006 4th Student Conference on Research and Development "Towards Enhancing Research Excellence in the Region”, pp. 115-119, 2006.
2. SOUSA S. I. V., MARTINS, F. G. ALVIM-FERRAZ M. C. M., PEREIRA M.C. (2007), Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, Environmental Modelling & Software 22,97–103
3. BYUN D. W., CHING J. K. S. (Eds.) (1999), Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) modeling system, EPA/600/R-99/030, Office of Research and Development, U.S. Environmental Protection Agency, Washington, D. C.
4. NAGENDRA S. M. S., VENUGOPAL K., JONES S. L. (2007), Assessment of air quality near traffic intersections in Bangalore city using air quality index, Transportation Research Part D 12, 167–176.
5. Ul-Saufie A. Z., Yahya A. Sh., Ramli N. A. and Abdul Hamid H.; Comparison Between Multiple Linear Regression And Feed forward Back propagation Neural Network Models For Predicting PM10 Concentration Level Based On Gaseous And Meteorological Parameters; International Journal of Applied Science and Technology, Vol. 1, No.4, 2011.
6. Dominick D., Juahir H., Latif M. T., Zain Sh. M. and Aris A. Z.; Spatial assessment of air quality patterns in Malaysia using multivariate analysis; Atmospheric Environment, Vol. 60, pp. 172 181, 2012.
7. KUMAR A., GOYAL P. (2012), Forecasting of Air Quality Index in Delhi Using Neural Network Based on Principal Component Analysis Pure Appl. Geophys. 170 (2013), 711–722.
8. Azid A., Juahir H., Latif M. T., Zain Sh. M. and Osman M. R.; Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia; Journal of Environmental Protection, No. 4, pp.1-10 , 2013.
9. Khan R. A., Zain Sh. M., Juahir H., Yusoff M. K. and Tg Hanidza T. I.; Using Principal Component Scores and Artificial Neural Networks in Predicting Water Quality Index; Chemometrics in Practical Applications Edited by Dr. Kurt Varmuza, Chapter 12, 2012.
10. VAN DEN ELSHOUT S., LEGER K., FABIO N. (2008), Comparing urban air quality in Europe in real time: A review of existing air quality indices and the proposal of a common alternative, Environment International 34, 720–726.
11. KUMAR A., GOYAL P. (2011), Forecasting of air quality index in Delhi using principal component regression technique, Atmospheric Pollution Research 2, 436–444.
12. Isiyaka H. A. and Azid A.; Air Quality Pattern Assessment in Malaysia Using Multivariate Techniques; Malaysian Journal of Analytical Sciences, Vol. 19, No. 5, pp. 966 -978, 2015.
13. SHI J. P., HARRISON R. M. (1997), Regression Modelling of Hourly NOx and NO2 concentrations in urban air in London, Atmospheric Environment 31, 4081–4097.
14. CHELANI A.B., RAO C.V.C., PHADKE K.M., HASAN, M.Z. (2002), Prediction of sulphur dioxide concentration using artificial neural networks, Environment Modelling and Software 17, 161–168.
15. CHANG J.C., HANNA S.R. (2004), Air quality model performance evaluation, Meteorology and Atmospheric Physics 87, 167–196.