Life Sciences 269 (2021) 119093

Contents lists available at ScienceDirect

Life Sciences

journal homepage: www.elsevier.com/locate/lifescie

Statistical analysis of COVID-19 infection caused by environmental factors: Evidence from Pakistan

Muhammad Fawad [a,b], Sumaira Mubarik [c], Saima Shakil Malik [d], Jingli Ren [a,*]

[a] Zhengzhou Key Laboratory of Big Data Analysis and Application, Henan Academy of Big Data, Zhengzhou University, Zhengzhou 450052, China
[b] School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China
[c] Department of Epidemiology and Biostatistics, School of Health Sciences, Wuhan University, Wuhan, Hubei 430071, China
[d] Department of zoology, University of Gujrat, Gujrat 50700, Pakistan

ARTICLE INFO

Keywords:
COVID-19
Environmental factors
Public health
Principal components regression

ABSTRACT

Background: Coronavirus disease 2019 (COVID-19) has become a severe public health problem around the globe. Various epidemiological, statistical, and laboratory-based studies have shown that the role of temperature and other environmental factors has important influence in the transmission of coronaviruses. Scientific research is needed to answer the questions about the spread and transmission of the infection, whether people could be avoided from being infected with COVID-19 in next summer.
Aim: We aim to investigate the association of daily average temperature, daily average dew point, daily average humidity, daily average wind speed, and daily average pressure with the infection caused by this novel coronavirus in Pakistan.
Key findings: First, we check the correlation between environmental factors and daily infected cases of COVID-19; among them, temperature and dew point have positive linear relationship with daily infected cases of COVID-19. The thought-provoking findings of the present study suggested that higher temperature and dew point can contribute to a rise in COVID-19 disease in four provinces of Pakistan, possible to genome modifications and viral resistance to harsh environment. Moreover, it is also observed that humidity in Punjab and Sindh, and wind speed in Balochistan and Khyber Pakhtunkhwa have influenced the spreading of daily infected COVID-19 cases.
Significance: Current study will serve as a guideline to develop understanding of environmental factors that influence COVID-19 spread, helping policymakers to prepare and handle a catastrophe resulting from this pandemic.

1. Introduction

Coronavirus disease 2019 (COVID-19), caused by the novel coronavirus is an infectious disease and initially documented in Wuhan city, China, at the end of December 2019 [1]. Afterward, it spread rapidly to other provinces and countries, although the Chinese government has taken strict and timely measures to stop the trafficking at the National and International levels [2]. The World Health Organization (WHO) [3] has declared it as public health emergency and a pandemic once it starts spreading to various countries outside of China. As per 27th December 2020, COVID-19 is affecting 213 countries and territories and there are about 80,794,083 confirmed cases with 1,766,513 total deaths, worldwide.

The situation may be different in Pakistan as the first case has been confirmed by the Ministry of Health, government of Pakistan on February 26, 2020 in Karachi, Sindh province [4]. Additionally, another case was confirmed by the Pakistan Federal Ministry of Health in Islamabad on the same day [5]. It has been observed that just in a couple of weeks we have confirmed 20 COVID-19 positive cases from 471 suspected ones with highest numbers in the Karachi, Sindh followed by the areas of Gilgit Baltistan and it keeps on rising drastically with every single day. All the initially confirmed cases had a travel history from Syria, Iran and United Kingdom (London) [6]. On 27th December 2020, there are about 471,335 confirmed cases in Pakistan and situation is still getting worse day by day with the average diagnosis of 2000 plus COVID-19 positive cases per day. The present scenario of Pakistan is not satisfactory, being a developing, heavily populated country with far less facilities than the required [3]. Therefore, we are facing problems in

* Corresponding author.
E-mail address: renjl@zzu.edu.cn (J. Ren).

https://doi.org/10.1016/j.lfs.2021.119093
Received 14 December 2020; Received in revised form 2 January 2021; Accepted 10 January 2021
Available online 18 January 2021

terms of economy, health care infrastructure and diagnostic capacity [7].

COVID-19 infection could cause severe respiratory illness, analogous to severe acute respiratory syndrome coronavirus (SARS-CoV). Largely, dry cough, fever, fatigue, shortness of breath, aches, pains and loss of taste/smell are common symptoms recorded at the onset of disease [8]. Due to contagious nature, COVID-19 has affected the lives of humans on a large scale and obviously restricted their movement. To reduce or prevent the virus spread, in Pakistan, it is suggested that people should not move around unless it is obligatory. Besides, people are prohibited to enter some of the metropolitan cities or their parts except in necessary conditions [9]. Such preventive and strict measures not only illustrate the severity of the situation but further confirms the mobility of virus through contact and travelling. Various studies have confirmed its human-to-human transmission through direct contact and respiratory droplets [10–12].

In addition to human-to-human transmission, meteorological parameters are thought to be effective factors in terms of the transmission, viability and broad range viral spread [13,14]. Various studies have reported the correlation between temperature and COVID-19 infection in China [15,16], Turkey [17] and Singapore [18] and demonstrated that the mean temperature and the number of COVID-19 cases have a positive linear association. Another study has also demonstrated the impact of humidity and temperature cause variations on COVID-19 mortality [15]. Chen and coworkers also explained the role of temperature, wind speed and relative humidity in COVID-19 severity and spread [19].

Therefore, in the present study, we aim to investigate the correlation between various environmental factors and COVID-19 pandemic in Pakistan. In contrast to other literature, we have focused on considering all the province in Pakistan instead of taking one or two cities or provinces. Weather related parameters include the in-depth investigation of daily average temperature, daily average dew point, daily average humidity, daily average wind speed and daily average pressure. We also focus on the diagnosis of COVID-19 daily new cases in four different provinces of Pakistan.

2. Material and methods

2.1. Data source

In this study, the data of environmental factors such as daily average temperature (°F), daily average dew point (°F), daily average Humidity (%), daily average wind speed (mph), and daily average pressure (Hg) for Punjab, Khyber Pakhtunkhwa, Sindh, and Balochistan are collected from Weather Underground website (https://www.wunderground.com). The information on daily new cases of COVID-19 is obtained from the government of Pakistan official port: (http://www.covid.gov.pk).

2.2. Statistical analysis

Gender and age-related distribution of COVID-19 infected, and deceased cases were expressed by percentages. Descriptive analysis of environmental factors and daily new COVID-19 cases included mean, median, interquartile range, and standard deviation. The Pearson correlation was used to check the relationship between COVID-19 and environmental factors. Association of COVID-19 disease and environmental factors were assessed using the Principal Component Regression (PCR) model among four provinces of Pakistan. 95% confidence interval (CI) was used to predict the COVID-19 disease during the study period between each province. The detail of PCR is presented in next section. *P*-value ≤0.05 was considered as level of significance.

2.3. Multicollinearity and its detection

In multiple regression analysis, when there exists a perfect or exact relationship between the independent variables, this phenomenon is known as multicollinearity or collinearity. When the problem of multicollinearity happened, it is of the challenge to achieve the reliable estimates of regression coefficients. This leads to incorrect inference about the relationship between the dependent variable and the independent variables. The existence of multicollinearity has some significant consequences on the ordinary least square (OLS) estimations of the regression coefficients, such as the high variance of the coefficients which may decrease the accuracy of the estimation and may result in coefficients appearance of the incorrect sign, the estimations of the parameters and their standard errors are highly sensitive to minor variations in data points, and it tends to overestimate variance of predicted values.

The problem of multicollinearity can be detected by several methods, the examination of the correlation matrix of independent variables is the most straightforward method. A high value of the correlation between two independent variables can suggest that the variables are collinear [20].

Variance inflation factor (VIF) is another tool widely used to diagnose multicollinearity problems. VIF measures the magnitude of multicollinearity in an OLS regression analysis. The coefficient of determination (R_j^2) of the multiple regression model is obtained by regressing covariate (X_j) on all other predictor variables in the model. The VIF can be computed as follows.

$$VIF = \frac{1}{1 - R_j^2} j = 1, 2, 3, \ldots p - 1 \tag{1}$$

The VIF provides an index that measures how much variance of the estimated regression coefficient is increased because of multicollinearity. According to practical experience, when any of the VIF values exceed 5 or 10, it is suggested that the corresponding regression coefficients are incorrectly estimated due to multicollinearity [21].

The presence of multicollinearity can also be identified by using the eigenvalues and condition index method. If the independent variables are collinear, then one or more of the eigenvalues would be small (close to zero). Let λ_1, λ_2, $\ldots\lambda_p$ be the eigenvalues of the correlation matrix ($X'X$). The condition number (K) and the condition index (CI) of the correlation matrix are defined as:

$$K = \frac{\lambda_{max}}{\lambda_{min}} and K_j = \frac{\lambda_{max}}{\lambda_j} \tag{2}$$

$$CI = \sqrt{K} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{3}$$

where λ_{max} is the maximum eigenvalue, λ_{min} is the minimum eigenvalue of the correlation matrix, and λ_j is the eigenvalue of the jth independent variable. If K is less than 100, there is no problem of multicollinearity, and if K is between 100 and 1000 indicates moderate to high multicollinearity, and K greater than 1000, there is severe multicollinearity. On the other hand, if $CI = \sqrt{K}$ value is between 10 and 30, there is mild to high multicollinearity, and when it exceeds 30, there is strong multicollinearity. Several researchers have recommended that the CI is the best available multicollinearity diagnostics for all types of regression models [22].

A study revealed that principal component analysis (PCA) is one of the appropriate methods of solving the collinearity among variables. Therefore, this technique produces better estimation and prediction than OLS when predictors are related [21,23,24].

2.4. Principal component regression (PCR)

The PCR is an alternative to classical multiple linear regression

(CMLR), which is based on the PCA. The PCR is widely used to estimate unknown regression coefficients in the CMLR model. The first step is to perform PCA on the original data in PCR; the second step is a dimensional reduction by choosing the number of principal components (PCs) (m) using cross-validation or test set error and finally, perform regression analysis using the first (m) dimension reduced PCs. In PCR, instead of directly regressing the dependent variable to the explanatory variables, the PCs of the explanatory variables are used as regressors. It usually uses only a subset of all the PCs for regression analysis, making PCR a type of regularized process and, therefore, a form of shrinkage estimator. The PCs with higher variances (depending on the eigenvectors corresponding to the higher eigenvalues of the sample variance-covariance matrix of the explanatory variables) are always chosen as regressors. However, the PCs with small variances can also be important in order to predict the outcome, in some cases, even more important.

The significant application of PCR is to solve the problem of multicollinearity that occurs between two or more explanatory variables that are close to being collinear. Initially, Hotelling and Kendall proposed replacing the original explanatory variables with their PCs in a CMLR model [25,26]. PCR can deal with such situations properly by removing some small-variance PCs in the step of regression [27]. Moreover, by generally regressing on only a subset of all PCs, PCR will result in a dimension reduction by considerably lowering the necessary number of parameters that describe the underlying model. This may be especially important in situations with high-dimensional covariates. In addition, by making an adequate selection of the PCs to be used for regression, PCR will lead to an accurate forecast of the outcome based on the assumed model. Consider the CMLR model in matrix notation.

$$Y = X\beta + \varepsilon \tag{4}$$

where Y is the (n × 1) matrix of dependent variable values (daily COVID-19 cases), X is the (n × p) matrix of independent variables (environmental factors), β is the (p × 1) vector of unknown coefficients, and ε is the (n × 1) vector of random errors. The OLS estimate of the regression coefficients are given.

$$\widehat{\beta} = (X'X)^{-1}X'Y \tag{5}$$

$$Var(\widehat{\beta}) = \widehat{\sigma}^2(X'X)^{-1} \tag{6}$$

Remember that as the variables are standardized, $X'X = R$, where R is the correlation matrix of independent variables, which are daily average temperature, daily average dew point, daily average humidity, daily average wind speed, and daily average pressure. To perform PCR, we transform the independent variables to their PCs, which can be mathematically represented as given below.

$$X'X = PDP' = Z'Z \tag{7}$$

where D denotes the diagonal matrix with eigenvalues of XX' on the diagonal, P is orthogonal so that $P'P = PP' = I$, and Z is a new transform data matrix (similar to the structure of X) consisting of PCs. The new set of explanatory variables can be expressed as follows.

$$(Z_1, Z_2, Z_3, \ldots, Z_p) = Z = XP = (X_1, X_2, X_3, \ldots, X_p) \tag{8}$$

The Z is a newly created variable as the weighted average of the variables understudy X. There is nothing new to us as we use transformations like the logarithm and the square root on our data series before running the regression model. As these new variables are PCs, their correlations with each other are zero. Since we have five independent variables in this study, so we will start with variables X1, X2, X3, X4, X5, and we will be ended with Z1, Z2, Z3, Z4, and Z5. These Z's are a linear combination of the original independent variables and known as PCs.

The strong multicollinearity can be identified as very small eigenvalues. To get rid of data on multicollinearity, we exclude the PCs (Z's) that are associated with small eigenvalues. Generally, just only one or two comparatively low eigenvalues can be obtained. For e.g., if only one small eigenvalue were found in a problem with five independent variables, Z5 would be omitted. Then we will regress Y on Z1, Z2, Z3, and Z4, using the OLS method, and multicollinearity will not be a problem anymore. Following that, we should transform our findings back to the X scale in order to achieve an estimate of β. Thus, in terms of PCs, the CMLR model may be reset as:

$$Y = Z\alpha + \varepsilon, where\ Z = XP\ and\ \alpha = P\beta \tag{9}$$

The OLS estimates of α are:

$$\widehat{\alpha} = (Z'Z)^{-1}Z'\ Y = D^{-1}Z'\ Y \tag{10}$$

$$Var(\widehat{\alpha}) = \sigma^2(Z'Z)^{-1} = \sigma^2 D^{-1} \tag{11}$$

Moreover, the smaller eigenvalue of X′X, suggesting that the variance of the associated regression coefficient is high.

3. Results

Overall statistics of infected and deceased COVID-19 patients were stratified by gender and age and expressed in Table 1. Disease was more prevalent among males than females, particularly in age groups 20 to 59 (~57%) years. Males have higher death rate in the age groups of 50–59 (18.4%), 60–69 (22.93%) and 70–79 (12.96%) years.

Further, the summary statistic of daily average temperature, daily average dew point, daily average humidity, daily average wind speed, daily average pressure and daily new COVID-19 cases for Punjab, Khyber Pakhtunkhwa, Sindh, and Balochistan were provided in Table 2. Median number of COVID-19 cases were higher in Sindh, Punjab and Khyber Pakhtunkhwa, respectively. At the same time, the highest daily average temperature was also observed in the same provinces (Sindh, Punjab and Khyber Pakhtunkhwa). Besides that, Sindh has also the highest daily average dew point and daily average humidity, whereas Balochistan has the lowest daily average dew point and daily average humidity and the median number of COVID-19 cases as well in the same duration. It was observed that Balochistan and Sindh have the highest daily average wind speed corresponding to the highest and lowest median number of COVID-19 cases in that duration (Table 2).

In order to assess the relationship between daily average temperature, daily average dew point, daily average humidity, daily average wind speed, daily average pressure and daily new COVID-19 cases, a Pearson correlation was constructed (Table 3). For Punjab province, the daily new COVID-19 cases have significant correlation with daily average temperature, daily average dew point, and daily average wind speed, but daily new COVID-19 cases show significantly higher and moderate positive correlation with daily average dew point ($r = 0.704$), and with daily average temperature ($r = 0.527$). It can be seen from Table 3, the daily new COVID-19 cases were linearly correlated with all environmental factors for Khyber Pakhtunkhwa province, among them COVID-19 has the highest positive correlation with daily average temperature ($r = 0.717$), and daily average dew point ($r = 0.675$). For Sindh province, the daily new COVID-19 cases have only significant correlation with daily average temperature ($r = 0.601$), and daily average dew point (0.512). While in Balochistan province the daily new COVID-19 cases have significant correlation with daily average temperature ($r = 0.695$). Overall, we noted that the daily new cases of COVID-19 in Punjab, Khyber Pakhtunkhwa and in Sindh were positively correlated with daily average temperature, and daily average dew point.

From the above findings, we also observed that there is also multicollinearity problem among environmental factors; therefore, to measure the influence of each environmental variable on COVID-19 we first constructed the principal components (PCs) of the variables understudy. These PCs include all the information contained in the original data set. Further, these PCs were used to model the daily new COVID-19 cases.

Table 1
The overall Percentage of COVID-19 infected and deceased cases by gender and age groups in Pakistan.

Age periods	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80+	Miscellaneous	Total
Overall percentage of infected (%)										
Male	4.44%	16.12%	17.81%	12.68%	10.24%	5.98%	2.13%	0.56%	4.1%	74.06%
Female	2.84%	6.09%	4.98%	3.62%	3.36%	2.03%	0.86%	0.26%	1.91%	25.95%
Overall percentage of death (%)										
Male	0.46%	1.92%	3.76%	8.13%	18.4%	22.93%	12.96%	4.98%	0.38%	73.92%
Female	0.31%	0.92%	1.23%	3.76%	6.9%	6.9%	4.45%	1.38%	0.23%	26.08%

Table 2
Summary of the environmental factors and daily new COVID-19 cases for four provinces.

Provinces	Factors	Time Period	Median	Interquartile range (IQR)	Mean	Standard deviation
Punjab	Daily New COVID-19 cases	17 Mar-16 June	331	816	633	712.3
	Daily average temperature	17 Mar-16 June	81.850	13.025	81.885	8.206
	Daily average dew point	17 Mar-16 June	59.700	7.775	59.985	5.182
	Daily average humidity	17 Mar-16 June	50.10	14.50	50.823	11.50
	Daily average wind speed	17 Mar-16 June	6.800	5.125	7.192	3.386
	Daily average pressure	17 Mar-16 June	29	0.200	28.890	0.666
Khyber Pakhtunkhwa	Daily new COVID-19 cases	16 Mar-16 June	167	236	205.5	204.1
	Daily average temperature	16 Mar-16 June	78	17.050	77.057	9.147
	Daily average dew point	16 Mar-16 June	57.400	6.600	57.701	4.397
	Daily average humidity	16 Mar-16 June	53.70	15.25	54.30	12.57
	Daily average wind speed	16 Mar-16 June	8.000	3.700	8.415	2.453
	Daily average pressure	16 Mar-16 June	28.600	0.200	27.644	3.887
Sindh	Daily new COVID-19 cases	28 Apr-16 June	810.5	853	1058.2	664
	Daily average temperature	28 Apr-16 June	89.050	2.275	88.754	1.426
	Daily average dew point	28 Apr-16 June	73.650	2.875	73.250	2.165
	Daily average humidity	28 Apr-16 June	62.450	5.300	61.794	3.514
	Daily average wind speed	28 Apr-16 June	11.200	3.075	11.310	2.700
	Daily average pressure	28 Apr-16 June	29.600	0.100	29.468	0.452
Balochistan	Daily new COVID-19 cases	13 Mar-16 June	59	119	87.9	102.3
	Daily average temperature	13 Mar-16 June	68.75	17.80	67.91	10.01
	Daily average dew point	13 Mar-16 June	36.800	8.250	36.285	5.970
	Daily average humidity	13 Mar-16 June	31.50	19.77	34.39	12.99
	Daily average wind speed	13 Mar-16 June	12.950	4.875	13.499	3.699
	Daily average pressure	13 Mar-16 June	24.900	0.100	24.794	0.447

Table 4 indicates the loading estimates of environmental factors in each component. The large loading value indicate the strong effect of that variable on that component. It is clear from Table 4 for Punjab province, that the variable daily average temperature, daily average dew point, daily average humidity, and daily average pressure have strong effect in component 1. While component 2 have large variation due to daily average dew point, daily average humidity and daily average wind speed. Similarly, loadings of other provinces can be seen in Table 4. Therefore, variation in these environmental variables may cause to increase COVID-19 disease.

To check the effect of the first two PCs on COVID-19 daily cases using principal component regression (PCR) and multiple regression was used (Table 5). For Punjab province first four components showed the significant effect, while using regression model only daily average dew point is significant. The daily average temperature and daily average wind speed are significant for Khyber Pakhtunkhwa when using a PCR model, but the results were change when using the simple regression model. In the case of Sindh province, the first three components show significant results for the PCR model, whereas the results for the simple regression model were similar to the Punjab. In Balochistan province, the daily average temperature, daily average dew point, and daily average wind speed effects were significant by using PCR model; however, only temperature effect was significant in the case of the regression model.

In summarizing, based on PCR results, average temperature, average dew point and average wind speed were positively related to daily new COVID-19 cases while average humidity showed the negative effect on daily new COVID-19 cases for Punjab province. On the other hand, for Khyber Pakhtunkhwa estimated results indicated the positive association of average temperature and average wind speed with daily new COVID-19 cases. For Sindh, average temperature and average dew point were positively related to daily new COVID-19 cases and average humidity was negatively related to daily new COVID-19 cases. Average temperature, dew point and wind speed was positively related to daily new COVID-19 cases in Baluchistan province (Table 5).

3.1. PCR models for four provinces

$$\text{Daily New COVID}-19\text{ Cases in Punjab} = 633.022 + 217.987 \times \text{daily average temperature} + 458.341 \times \text{daily average dew point} - 64.469 \times \text{daily average humidity} + 130.02 \times \text{daily average wind speed} - 62.876 \times \text{daily average pressure}$$

$$\text{Daily New COVID}-19\text{ Cases in Khyber Pakhtunkhwa} = 205.452 + 91.812 \times \text{daily average temperature} + 104.88 \times \text{daily average dew point} - 54.541 \times \text{daily average humidity} + 50.269 \times \text{daily average wind speed} - 7.245 \times \text{daily average pressure}$$

$$\text{Daily New COVID}-19\text{ Cases in Sindh} = 1058.24 + 307.33 \times \text{daily average temperature} + 361.392 \times \text{daily average dew point} - 57.229 \times \text{daily average humidity} - 47.205 \times \text{daily average wind speed} - 31.391 \times \text{daily average pressure}$$

$$\text{Daily New COVID}-19\text{ Cases in Balochistan} = 87.875 + 56.319 \times \text{daily average temperature} + 18.964 \times \text{daily average dew point} - 34.586 \times \text{daily average humidity} + 25.914 \times \text{daily average wind speed} - 4.242 \times \text{daily average pressure}$$

3.2. Regression models for four provinces

$$\text{Daily New COVID}-19\text{ Cases in Punjab} = -1119.984 - 51.811 \times \text{daily average temperature} + 152.630 \times \text{daily average dew point} - 47.474 \times \text{daily average humidity} + 13.079 \times \text{daily average wind speed} - 29.119 \times \text{daily average pressure}$$

$$\text{Daily New COVID}-19\text{ Cases in Khyber Pakhtunkhwa} = -1080.829 + 4.520 \times \text{daily average temperature} + 21.978 \times \text{daily average dew point} - 3.027 \times \text{daily average humidity} + 8.897 \times \text{daily average wind speed} - 8.704 \times \text{daily average pressure}$$

$$\text{Daily New COVID}-19\text{ Cases in Sindh} = -9112.731 - 64.693 \times \text{daily average temperature} + 402.437 \times \text{daily average dew point} - 169.589 \times \text{daily average humidity} - 55.204 \times \text{daily average wind speed} - 83.540 \times \text{daily average pressure}$$

$$\text{Daily New COVID}-19\text{ Cases in Balochistan} = -33.780 + 4.980 \times \text{daily average temperature} + 3.291 \times \text{daily average dew point} - 1.584 \times \text{daily average humidity} + 3.357 \times \text{daily average wind speed} - 13.182 \times \text{daily average pressure}$$

Additionally, the relationship between the actual daily new COVID-19 cases and predicated new COVID-19 cases using PCR model with a 95% confidence interval were illustrated in Fig. 1. The graphs for Punjab, Khyber Pakhtunkhwa, Sindh and Baluchistan provinces showed the good fit of the under study variable data by PCR model and present how well the predicated COVID-19 cases fall within the confidence interval.

4. Discussion

The PCR model is applied in this study to estimate the impact of environmental factors such as temperature, dew point, humidity, wind speed, and pressure on the transmission of COVID-19 by examining the relationship between the daily confirmed cases of COVID-19 and environmental factors in all the four provinces of Pakistan. Statistical summary displays that disease was more prevalent among males than females, predominantly in age groups 20 to 59 years with higher death rate in the age groups of 50–59 (18.4%), 60–69 (22.93%) and 70–79 (12.96%) years in males. Our findings suggest that COVID-19 is more common among male population and risk of disease increases with increased age, which is concordant with the findings of previous literature [1]. It has been reported that older individuals particularly men tend to develop more severe disease type with greater mortality rate [28] possibly due to higher concentrations of angiotensin-converting enzyme 2 (ACE2) in their blood as compared to women. Since, ACE2 enzymes enables COVID-19 to infect healthy/normal cells, this may help to understand why men are more vulnerable to COVID-19 than women [29].

We have observed that Sindh has highest number of COVID-19 cases followed by Punjab and Khyber Pakhtunkhwa corresponding to the relative temperature, dew point and humidity of these areas. Here we came across one of the most interesting finding that COVID-19 incidence together with death rate was greater in zones with higher temperature ranges and adverse environmental conditions i.e., Sindh as compared to the province with lower average temperature, dew point and humidity such as Balochistan. Furthermore, Balochistan and Sindh have the highest daily average wind speed corresponding to the highest and lowest median number of COVID-19 cases in the selected time period as shown in Table 2.

Further, we check the association between COVID-19 and environmental factors using a correlation matrix. In Punjab province, we found significant positive correlation of COVID-19 with daily average dew point ($r = 0.704$), and with daily average temperature ($r = 0.527$) representing that COVID-19 cases increases with increased temperature and dew point. Results from the Khyber Pakhtunkhwa province were also pretty much alike with Punjab illustrating linear relationship of the daily new COVID-19 cases with all environmental factors and among them COVID-19 has the highest positive correlation with daily average temperature ($r = 0.717$), and daily average dew point ($r = 0.675$). Similarly, in Sindh province, the significant correlation of COVID-19 was found with daily average temperature and daily average dew Point, whereas Balochistan province the daily new COVID-19 cases have significant correlation with daily average temperature. Therefore, it was observed that in Punjab, Khyber Pakhtunkhwa and Sindh provinces COVID-19 have a significant association with daily average temperature, and daily average dew point. These findings are a bit different from the already reported literature might because COVID-19 first arises during January and February in China, Iran and Italy. Whereas, when it comes to Pakistan, weather was almost different and warmer due to seasonal and geographical changes. Another reason might be the continuously changing genome of COVID-19 leading to its better survival in more harsh conditions [30].

It's been so long that researchers have recognized that the rise in average global temperatures is increasing the geographical existence of vector borne diseases for instance dengue fever and malaria, as the animals transmitting them are acclimatizing to more extensive regions [31]. The exact connection of respiratory illnesses such as COVID-19 or influenza with a warming planet is not much clear and needs in-depth investigations. But some of the researchers are worried that climate change can change the relationship between our body's defenses and such devastating pathogens. These variations could embrace the microbial adaptation to a warming world, alterations in how bacteria and viruses interact with their respective animal hosts, and a weakened human immune response [32]. We are blessed with immune system as natural defense system against harmful substances. When any respiratory pathogen like COVID-19 virus enters the body via airways, it damages cells by destroying their normal machinery leading to self-replication. The injured cells are meant to release signaling proteins

Table 3
Pearson correlation matrix between environmental factors and daily new COVID-19 cases for four provinces.

Province	Variables	Daily new cases of COVID-19	Daily average temperature	Daily average dew point	Daily average humidity	Daily average wind speed	Daily average pressure
Punjab	Daily new cases of COVID-19	**1**	**0.527**	**0.704**	−0.091	**0.338**	−0.181
	Daily average temperature	**0.527**	**1**	**0.473**	**−0.756**	0.175	**−0.277**
	Daily average dew point	**0.704**	**0.473**	**1**	0.200	**0.430**	−0.116
	Daily average humidity	−0.091	**−0.756**	0.200	**1**	0.137	**0.225**
	Daily average wind speed	**0.338**	0.175	**0.430**	0.137	**1**	**−0.275**
	Daily average pressure	−0.181	**−0.277**	−0.116	**0.225**	**−0.275**	**1**
Khyber Pakhtunkhwa	Daily new cases of COVID-19	**1**	**0.717**	**0.675**	**−0.481**	**0.425**	**−0.276**
	Daily average temperature	**0.717**	**1**	**0.583**	**−0.871**	**0.304**	**−0.262**
	Daily average dew point	**0.675**	**0.583**	**1**	−0.137	**0.511**	−0.021
	Daily average humidity	**−0.481**	**−0.871**	−0.137	**1**	−0.087	**0.264**
	Daily average wind speed	**0.425**	**0.304**	**0.511**	−0.087	**1**	0.014
	Daily average pressure	**−0.276**	**−0.262**	−0.021	**0.264**	0.014	**1**
Sindh	Daily new cases of COVID-19	**1**	**0.601**	**0.512**	0.054	0.036	−0.070
	Daily average temperature	**0.601**	**1**	**0.317**	**−0.361**	0.010	−0.028
	Daily average dew point	**0.512**	**0.317**	**1**	**0.744**	**0.437**	−0.175
	Daily average humidity	0.054	**−0.361**	**0.744**	**1**	**0.370**	−0.145
	Daily average wind speed	0.036	0.010	**0.437**	**0.370**	**1**	**−0.371**
	Daily average pressure	−0.070	−0.028	−0.175	−0.145	**−0.371**	**1**
Balochistan	Daily new cases of COVID-19	**1**	**0.695**	0.171	**−0.500**	**0.407**	−0.024
	Daily average temperature	**0.695**	**1**	0.074	**−0.752**	**0.369**	0.044
	Daily average dew point	0.171	0.074	**1**	**0.484**	**0.360**	0.058
	Daily average humidity	**−0.500**	**−0.752**	**0.484**	**1**	−0.196	0.025
	Daily average wind speed	**0.407**	**0.369**	**0.360**	−0.196	**1**	0.053
	Daily average pressure	−0.024	0.044	0.058	0.025	0.053	**1**

Values in bold are different from 0 with a significance level alpha = 0.05.

Table 4
Factor loadings of five environmental variables for four provinces.

Punjab					
Variables	PC1	PC2	PC3	PC4	PC5
Daily average temperature	−0.922	−0.237	0.298	−0.013	0.066
Daily average dew point	−0.543	0.652	0.425	−0.312	−0.043
Daily average humidity	0.626	0.751	−0.022	−0.198	0.060
Daily average wind speed	−0.443	0.686	−0.213	0.536	−0.001
Daily average pressure	0.557	−0.057	0.763	0.323	0.000

Khyber Pakhtunkhwa					
Variables	PC1	PC2	PC3	PC4	PC5
Daily average temperature	−0.962	−0.156	0.198	−0.078	0.075
Daily average dew point	−0.683	0.552	−0.106	−0.466	−0.036
Daily average humidity	0.777	0.488	−0.344	−0.193	0.060
Daily average wind speed	−0.514	0.658	−0.248	0.491	0.001
Daily average pressure	0.338	0.551	0.762	0.028	0.003

Sindh					
Variables	PC1	PC2	PC3	PC4	PC5
Daily average temperature	−0.006	0.978	0.198	−0.027	−0.061
Daily average dew point	0.863	0.248	0.410	−0.135	0.089
Daily average humidity	0.832	−0.433	0.290	−0.166	−0.087
Daily average wind speed	0.727	0.097	−0.325	0.597	−0.006
Daily average pressure	−0.460	−0.178	0.777	0.390	−0.001

Balochistan					
Variables	PC1	PC2	PC3	PC4	PC5
Daily average temperature	−0.903	0.192	−0.019	−0.357	0.141
Daily average dew point	0.228	0.896	−0.130	−0.339	−0.112
Daily average humidity	0.930	0.322	−0.018	−0.002	0.177
Daily average wind speed	−0.460	0.704	−0.125	0.526	0.025
Daily average pressure	−0.029	0.219	0.975	0.015	−0.006

Table 5
Estimates model parameters using principal component regression and regression analysis.

Principal component analysis			Regression analysis		
Predictors	Coefficient	P	Predictors	Coefficient	P
Punjab					
Intercept	633.022	<0.0001	Intercept	1119.984	0.733
Daily average temperature	217.987	<0.0001	Daily average temperature	51.811	0.220
Daily average dew point	458.341	<0.0001	Daily average dew point	152.630	0.001
Daily average humidity	−64.469	0.000	Daily average humidity	−47.474	0.082
Daily average wind speed	130.02	0.025	Daily average wind speed	13.079	0.455
Daily average pressure	−62.876	0.065	Daily average pressure	−29.119	0.725
Khyber Pakhtunkhwa					
Intercept	205.452	<0.0001	Intercept	1080.829	0.068
Daily average temperature	91.812	<0.0001	Daily average temperature	4.520	0.655
Daily average dew point	104.88	0.068	Daily average dew point	21.978	0.041
Daily average humidity	−54.541	0.125	Daily average humidity	−3.027	0.612
Daily average wind speed	50.269	0.019	Daily average wind speed	8.897	0.154
Daily average pressure	−7.245	0.831	Daily average pressure	−8.704	0.018
Sindh					
Intercept	1058.240	<0.0001	Intercept	9112.731	0.405
Daily average temperature	307.33	0.019	Daily average temperature	64.693	0.681
Daily average dew point	361.392	<0.0001	Daily average dew point	402.437	0.009
Daily average humidity	−57.229	0.010	Daily average humidity	−169.589	0.064
Daily average wind speed	−47.205	0.086	Daily average wind speed	−55.204	0.080
Daily average pressure	−31.391	0.053	Daily average pressure	−83.540	0.610
Balochistan					
Intercept	87.875	<0.0001	Intercept	33.780	0.936
Daily average temperature	56.319	<0.0001	Daily average temperature	4.980	0.005
Daily average dew point	18.964	0.000	Daily average dew point	3.291	0.181
Daily average humidity	−34.586	0.165	Daily average humidity	−1.584	0.326
Daily average wind speed	25.914	0.018	Daily average wind speed	3.357	0.172
Daily average pressure	−4.242	0.835	Daily average pressure	−13.182	0.433

known as cytokines, which communicates with other body parts to initiate an immune response against the foreign intruders [33]. Another basic defense mechanism in our bodies against pathogens is to elevate the body temperature relative to the pathogen's environment. Due to this change, many microorganisms that are adapted to cool temperatures become impotent to tolerate warmer human body [34]. However, the higher ambient temperatures anticipated with an altering climate could, favor pathogens that will be more tough for our body to fight, leading us to think about the discovery of novel pathogens like COVID-19 and their mode of action. Increased replication rate due to rapid COVID-19 spread might be a reason behind variable rate of environmental factors around the globe [35].

In order to manage the problem of collinearity between the environmental variables, we first developed the PCs of the variables under study. Principal component regression analysis revealed that the average temperature, average dew point and average wind speed were positively related to daily new COVID-19 cases while average humidity showed the negative effect on daily new COVID-19 cases for Punjab province. Whereas, estimated results of Khyber Pakhtunkhwa showed the positive association of average temperature and average wind speed with daily new COVID-19 cases. Similarly, average temperature and average dew point were positively related to daily new COVID-19 cases in Sindh. Furthermore, average temperature, dew point and wind speed were positively linked to daily new COVID-19 cases in Baluchistan province highlighting the role of environmental factors in disease incidence and spread. Many environmental studies involve data collection and use to explore the relationship between the dependent variables and a set of independent variables. When the independent variables are related to each other, then a situation generally known as multicollinearity. Then findings from other scientific methods would become less reliable. In the present research, the usefulness of the PCR to solve multicollinearity problems was investigated with a specific interest in the effect of environmental factors on daily new cases of COVID-19. The findings of this study could stimulate other researchers to investigate alternative approaches and how to apply them to the solution of multicollinearity problems in other related fields of study.

We have compared our findings with the results of one of the recent studies. Tosepu et al., postulated that the average temperature is significantly correlated with COVID-19 among Indonesian Population with the claim that minimum and maximum temperature, humidity and rainfall does not have any role towards COVID-19 spread [36]. In the current study, the daily average temperature of all the four provinces from Pakistan is found to be correlated with COVID-19. The correlation is higher in the present study for all environmental factors for Khyber Pakhtunkhwa province, among them COVID-19 has the highest correlation with daily average temperature ($r = 0.717$). Zhu and Xie have reported positive linear relationship between mean temperature and the number of COVID-19 cases with a threshold of 3 °C [16]. Differences in the results were might be due to different geographical boundaries and seasonal variations. Unlike the other studies, analysis of environmental factors from all the four provinces from Pakistan greatly help towards the generalization of the results. This research provides knowledge and understanding of environmental factors that may decrease or increase the disease's spread, allowing people to prepare better and organize their everyday activities based on weather forecasts.

Moreover, the updated detailed information about COVID-19 cases for each city could help in broadening the findings. Besides, the exact number of people, who have come from other countries and have been quarantined in different areas, could improve the accuracy of the analyses.

5. Conclusions

To the best of our knowledge, this is the first study to investigate the relationship between daily average temperature, daily average dew point, daily average humidity, daily average wind speed, daily average pressure, and daily COVID-19 confirmed cases in Pakistan. The findings of this research are intended to show that environmental factors can affect the dissemination of the virus. We have come to know an interesting result that in Pakistan, disease spread increases with increased

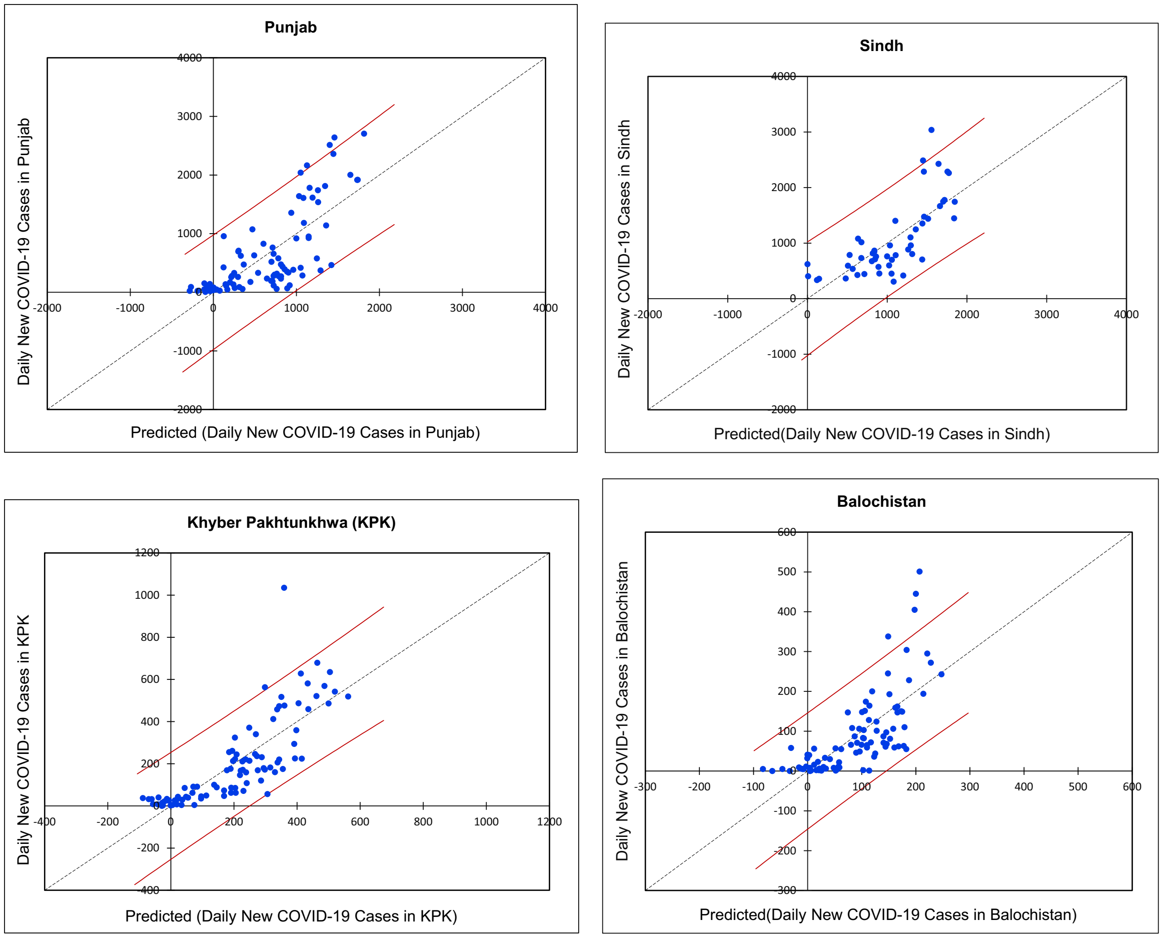

Fig. 1. Relationship between actual daily new cases of COVID-19 and predicated new cases of COVID-19 based on the principal component regression model.

temperature and other environmental factors suggesting survival and resistance of COVID-19 virus in harsher conditions. Here, it should be noted that this study will serve as a guideline to develop general knowledge and understanding of environmental factors that influence the spread of the pandemic, helping policymakers to prepare and handle a catastrophe resulting from COVID-19. It is better to maintain a lock-up and social distance until the vaccine is produced.

Declaration of competing interest

The author reports no conflicts of interest in this work.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No. 11771407), and CAE Advisory Project (No. 2020-ZD-16). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript. The authors would like to thank the funding foundations for their support.

Availability of supporting data

The data of environmental factors are collected from Weather Underground website (https://www.wunderground.com). The information on daily new cases of COVID-19 is obtained from the government of Pakistan official port: (http://www.covid.gov.pk).

References

[1] M. Fawad, S. Mubarik, S.S. Malik, Y. Hao, C. Yu, J. Ren, Trend dynamics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission in 16 cities of Hubei Province, China, Clin. Epidemiol. 12 (2020) 699.

[2] Z. Du, L. Wang, S. Cauchemez, X. Xu, X. Wang, B.J. Cowling, et al., Risk for transportation of coronavirus disease from Wuhan to other cities in China, Emerg. Infect. Dis. 26 (2020) 1049.

[3] World Health Organization, Novel Coronavirus (2019-nCoV) Situation Reports, 2020.

[4] I. Ali, S.A. Shah, N. Siddiqui, Pakistan confirms first two cases of coronavirus, govt says "no need to panic", DAWN COM 26 (February 2020).

[5] A. Waris, A.U. Khan, M. Ali, A. Ali, A. Baset, COVID-19 outbreak: current scenario of Pakistan, New Microbes New Infect. 100681 (2020).

[6] M. Saqlain, M.M. Munir, A. Ahmed, A.H. Tahir, S. Kamran, Is Pakistan prepared to tackle the coronavirus epidemic?, in: Drugs & Therapy Perspectives, 2020, pp. 1–2.

[7] M.A. Akram, COVID-19 pandemic and government policies to control its situation in Pakistan, Acta Sci. Microbiol. 3 (2020) 164–170.

[8] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (2020) 497–506.

[9] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, et al., Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia, N. Engl. J. Med. 382 (2020) 1199–1207.

[10] J.F.-W. Chan, S. Yuan, K.-H. Kok, K.K.-W. To, H. Chu, J. Yang, et al., A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, Lancet 395 (2020) 514–523.

[11] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, P.-R. Hsueh, Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges, Int. J. Antimicrob. Agents 105924 (2020).

[12] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, et al., Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China, Jama 323 (2020) 1061–1069.

[13] S. Herfst, M. Böhringer, B. Karo, P. Lawrence, N.S. Lewis, M.J. Mina, et al., Drivers of airborne human-to-human pathogen transmission, Curr. Opin. Virol. 22 (2017) 22–29.

[14] N. Van Doremalen, T. Bushmaker, V. Munster, Stability of Middle East respiratory syndrome coronavirus (MERS-CoV) under different environmental conditions, Eurosurveillance. 18 (2013) 20590.
[15] Y. Ma, Y. Zhao, J. Liu, X. He, B. Wang, S. Fu, et al., Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China, Sci. Total Environ. 138226 (2020).
[16] Y. Zhu, J. Xie, Association between ambient temperature and COVID-19 infection in 122 cities from China, Sci. Total Environ. 138201 (2020).
[17] M. Şahin, Impact of weather on COVID-19 pandemic in Turkey, Sci. Total Environ. 728 (2020), 138810.
[18] S.K. Pani, N.-H. Lin, S. RavindraBabu, Association of COVID-19 pandemic with meteorological parameters over Singapore, Sci. Total Environ. 740 (2020) 140112.
[19] B. Chen, H. Liang, X. Yuan, Y. Hu, M. Xu, Y. Zhao, et al., Predicting the local COVID-19 outbreak around the world with meteorological conditions: a model-based qualitative study, BMJ Open 10 (2020) e041397.
[20] M. El-Dereny, N. Rashwan, Solving multicollinearity problem using ridge regression models, Int. J. Contemp. Math Sci. 6 (2011) 585–600.
[21] D.C. Montgomery, E.A. Peck, G.G. Vining, Introduction to Linear Regression Analysisx, University of Chicago Press, 2012.
[22] D.N. Gujarati, Basic Econometrics, Forth edition, McGraw-Hill, Singapura, 2003.
[23] H. Artigue, G. Smith, The principal problem with principal components regression, Cogent Math. Stat. 6 (2019) 1622190.
[24] A.S. Gwelo, Principal components to overcome multicollinearity problem, Oradea J. Bus. Econ. 4 (2019) 79–91.
[25] H. Hotelling, The relations of the newer multivariate statistical methods to factor analysis, Br. J. Stat. Psychol. 10 (1957) 69–79.
[26] M.G. Kendall, A Course in Multivariate Analysis, Charles Griffin & Co, London, 1957.
[27] M. Alibuhtto, T. Peiris, Principal Component Regression for Solving Multicollinearity Problem, 2015.
[28] J.-M. Jin, P. Bai, W. He, F. Wu, X.-F. Liu, D.-M. Han, et al., Gender differences in patients with COVID-19: focus on severity and mortality, Front. Public Health 8 (2020) 152.
[29] P. Conti, A. Younes, Coronavirus COV-19/SARS-CoV-2 affects women less than men: clinical response to viral infection, J. Biol. Regul. Homeost. Agents 34 (2020) 71.
[30] M. Sironi, S.E. Hasnain, B. Rosenthal, T. Phan, F. Luciani, M.-A. Shaw, et al., SARS-CoV-2 and COVID-19: a genetic, epidemiological, and evolutionary perspective, Infect. Genet. Evol. 84 (2020) 104384.
[31] J.M. Medlock, S.A. Leach, Effect of climate change on vector-borne disease risk in the UK, Lancet Infect. Dis. 15 (2015) 721–730.
[32] D.R. Brooks, E.P. Hoberg, W.A. Boeger, The Stockholm Paradigm: Climate Change and Emerging Disease, University of Chicago Press, 2019.
[33] T. Acter, N. Uddin, J. Das, A. Akhter, T.R. Choudhury, S. Kim, Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: a global health emergency, Sci. Total Environ. 730 (2020), 138996.
[34] A. Casadevall, Climate change brings the specter of new infectious diseases, J. Clin. Invest. 130 (2020).
[35] N. Masood, S.S. Malik, M.N. Raja, S. Mubarik, C. Yu, Unraveling the epidemiology, geographical distribution, and genomic evolution of potentially lethal coronaviruses (SARS, MERS, and SARS CoV-2), Front. Cell. Infect. Microbiol. 10 (2020) 499.
[36] R. Tosepu, J. Gunawan, D.S. Effendy, L.O.A.I. Ahmad, H. Lestari, H. Bahar, P. Asfian, Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia, Sci. Total Environ. 725 (2020), 138436.