2012 2nd International Conference on Environmental and Agriculture Engineering
IPCBEE vol.37 (2012) © (2012) IACSIT Press, Singapore

Future PM_{10} Concentration Prediction Using Quantile Regression Models

Ahmad Zia Ul-Saufie [1,+], Ahmad Shukri Yahaya [2], Nor Azam Ramli [2] and Hazrul Abdul Hamid[2]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara, MALAYSIA

[2] Clean Air Research Group, School of Civil Engineering,Universiti Sains Malaysia, MALAYSIA

Abstract. Quantile regression is one of the methods for predicting environmental problem. Quantile regression can act as a complement to multiple linear regression (MLR) method because quantile regression provide answers similar to least square regression when the data are linear and have normally distributed errors. Besides that, quantile regression offers more complete view of the statistical landscape and relationships among variables. The aim of this study is to investigate the performance of quantile regression method in predicting future (next day, next 2-day and next 3-day) PM_{10} concentration levels in Seberang Perai, Malaysia and compared the result with multiple linear regression (MLR). Quantile regression (QR) and multiple regression models are examined for Seberang Perai, Pulau Pinang with the same independent variables, enabling a comparative study of the two approaches. Model comparison statistics using Prediction Accuracy (PA), Coefficient of Determination (R^2), Index of Agreement (IA) , Normalised Absolute Error (NAE) and Root Mean Square Error (RMSE) show that QR is better than MLR with average QR method 1.45% better than MLR for next day, 3.3% better for next 2-day and 5.36% better for next 3-day.

Keywords: Quantile Regression, Ordinary Least Square, Performance Indicator, PM_{10}

1. Introduction

Quantile Regression will be used to determine the relationship between dependent variables (x) and explanatory variables. Quantile Regression was introduced by [1] and after three decades, it is gradually popular among researchers as an alternative to MLR when the assumption of ordinary least squares are not met. [2] used this method for modelling the effects of meteorological variables on ozone concentration and the result showed that QR provides more information and accuracy than OLS. QR can provide more information since this method will provide models at all quantiles. QR also can examine the entire distribution of the variable of interest rather than a single measure of the central tendency of its distribution [2]. Typical measures of central tendency are average (mean) values, middle (median) value or the most likely values (mode).

According to [3], OLS have some limitations. First, OLS summarizes the response for fixed values of predictor variable, but cannot extend to non-central locations. Second, model assumption are not always met especially homocedasticity assumption and when the distribution is skewed. Hence the model can be influenced by outlier. Thus QR has the potential to be more useful and accurate because all quantiles can be used to described non-central position of a distribution.

[4] studied the potential of quantile regression to predict ozone concentrations, the result showed QR is more efficient for extreme value data and very useful to forecast higher ozone concentration. [5] found that QR give more different impact at different point of distribution and when the data is skewed, the result is shown to be more accurate than OLS.

+ Corresponding author. Tel.: + 60195900657; fax: +6043822768
E-mail address: ziaulsaufie@gmail.com.

[6] found quantile regression has the lowest residual when compared with MLR, principal component regressions, independent component regression and partial least squares regression during training step. Besides that, [6] also discussed about the criterias selection of the modelling techniques such as complexity, flexibility, accuracy, speed of computation and interpretability.

Quantile regression have some advantages to multiple linear regression such as ([7]) it is distribution free and does not use any properties, does not require independence or a weak degree of dependence and it is robust to outliers.

The aim of this study is to investigate the performance of quantile regression method in predicting future (next day, next 2-day and next 3-day) PM_{10} concentration levels in Seberang Perai, Malaysia. Besides, this study is also to compare the performance between quantile regression and multiple linear regression. This model is useful because it facilitates respective authorities to carry out suitable actions to reduce the impact of air pollution.

2. Methodology

2.1. Area of Study

Seberang Perai is an industrial area in Pulau Pinang, Malaysia. This site is important because historical records showed that it has one of the highest PM_{10} concentrations in Malaysia because it is situated near the industrial area which influenced the PM_{10} concentration reading. Annual average PM_{10} concentration are 2001(61.73µg/m^3), 2002(75.03 µg/m^3), 2003(80.13 µg/m^3), 2004 (92.31 µg/m^3), 2005(78.99 µg/m^3), 2006(49.81 µg/m^3) and 2007 (45.45 µg/m^3). This study used hourly observations from January 2004 until December 2007 that was transformed into daily data by taking the average PM_{10} concentration level for each day. Relative humidity (RH), wind speed (ws), nitrogen dioxide (NO_2), temperature (T), carbon monoxide (CO), sulphur dioxide (SO_2) and previous PM_{10} were used as independent variables. On average, wind speed in the area was 6.523 m/s, T (28.185^{o}C), RH (75.315%), SO_2 (0.0061 ppm), NO_2 (0.01334 ppm), CO (0.4967 ppm) and PM_{10} (67.24 µg/m^3).

2.2. Quantile Regression

[1] introduced quantile regression and [8] described the estimation of the coefficient of a quantile regression model. Given a random variable y with right continuous distribution, $F_y = \Pr(Y \leq y)$. The quantile regression $Q(\tau)$with $\tau \epsilon (0,1)$ is defined as follows:

$$Q(\tau) = \inf\{y: F(y) \geq \tau\}$$

The quantile was also formulated ([2], [4]) as the solution to minimize problem:

$$\hat{Q}_y(\tau) = arg \min_a \left\{ \sum_{i:y_i \geq a} \tau |y_i - a| + \sum_{i:y_i < a} (1-\tau)|y_i - a| \right\} = arg \min_a \sum_i \rho_\tau (y_i - a)$$

From equation 2, the quantile regression coefficients are obtained by solving with respect to $\beta(\tau)$:

$$\hat{\beta}(\tau) = \underset{\beta(\tau) \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ \sum_{i:y_i \geq \acute{x}\beta(\tau)} \tau |y_i - x_\iota \beta(\acute{\tau})| + \sum_{i:y_i < \acute{x}\beta(\tau)} (1-\tau)|y_i - x_\iota \acute{\beta}(\tau)| \right\}$$

2.3. Performance Indicator

Performance indicators are used to evaluate the goodness of fit for the QR and MLR for future PM_{10} concentration prediction in Seberang Prai, Pulau Pinang. Performance indicators were used to determine the best method in predicting PM_{10} concentration are normalized absolute error (NAE), root mean square error

(RMSE), index of agreement (IA), prediction accuracy (PA), and coefficient of determination (R^2). The equations used were reported by [9].

3. Result and Discussion

The air pollution data for 2004 until 2007 at Seberang Perai is summarized in Table 1. Since the mean of PM_{10} concentration is 67.24 µg/m^{3}and median (50th percentile) is 57.87 µg/m^3, it showed the data is skewed to the right (have extreme event). From this table, it also showed that ws, RH, temperature, and NO_2 had almost equal values of mean and median. But PM_{10}, SO_2 and CO is more skewed to the right side. From the first inference of study, it can be concluded that quantile regression is more suitable than MLR because QR can minimize influence of outlier data.

Table 1: Quantile Values of Variables

Quantile	PM_{10}	ws	RH	T	SO_2	NO_2	CO
Mean	67.24	6.523	75.315	28.185	0.006	0.013	0.4967
0.1	36.17	5.267	66.610	26.513	0.002	0.009	0.296
0.2	40.63	5.605	69.258	27.150	0.002	0.010	0.342
0.3	44.93	5.918	71.750	27.580	0.003	0.011	0.389
0.4	49.53	6.179	73.833	27.921	0.004	0.012	0.430
0.5 (med)	57.87	6.440	75.375	28.238	0.005	0.013	0.473
0.6	73.73	6.663	77.195	28.547	0.006	0.014	0.521
0.7	83.91	6.992	79.000	28.896	0.007	0.015	0.564
0.8	93.59	7.334	81.275	29.324	0.009	0.016	0.627
0.9	106.32	7.916	84.125	29.838	0.013	0.018	0.725

One of the advantages of QR is to provide readily interpretable results. Table 2 shows all coefficient of QR PM_{10} concentrations models for next day in Seberang Perai, Penang. The higher the PM_{10} concentration quantiles, the higher the value of the constant in the model such as at the 0.30 quantile is 3.29 µg/m^{3}and over 20 µg/m^3 at the 0.7 and above quantile.

The quantile regression approach shows that the effects of SO_2, NO_2, CO and previous PM_{10} concentration are consistent for all quantile. SO_2 had positive correlation with PM_{10} in the area because most SO_2 came from diesel fueled vehicle motor emissions and industrial activities. NO_2 and CO have negative correlation because this area uses less petrol fueled vehicle. But, relationship between meteorological (RH, WS and T) parameters and PM_{10} concentration for next day are more complex which is reflected in the sign, size and significance of the estimated coefficient.

Table 2: Coefficient of Quantile Regression Models for Next Day.

Quantile	Constant	ws	RH	T	SO_2	NO_2	CO	PM_{10-1}
0.1	2.123	0.414	0.039	0.056	250.860	-541.931	-6.675	0.809
0.2	2.030	0.344	0.038	0.131	132.676	-753.027	-3.762	0.875
0.3	3.291	0.088	0.026	0.417	84.238	-647.600	-4.431	0.905
0.4	16.571	-0.019	-0.0004	-0.151	76.492	-521.200	-7.905	0.925
0.5	19.521	0.100	-0.001	-0.275	54.120	-507.894	-7.012	0.958
0.6	8.708	0.080	0.064	0.007	73.444	-532.813	-7.188	0.972
0.7	21.191	0.008	0.001	-0.256	116.073	-578.236	-5.901	1.013
0.8	28.598	-0.146	-0.006	-0.388	241.418	-733.318	-7.211	1.059
0.9	28.995	-0.333	0.059	-0.526	102.964	-526.941	-11.122	1.133

Performance indicators were used to select the best quantile for predicting PM_{10} concentration for next day at Seberang Perai as shown in Table 3. From five performance indicators applied, PA and R^2 show that 0.4 quantile gave better fit than others quantiles. Only NAE, RMSE and IA show that 0.5 quantile is the best quantile for PM_{10} concentration models. Therefore, 0.5 quantile is used to represent PM_{10} concentration models in Seberang Perai station.

Table 3: Performance Indicators for Next Day PM_{10} Concentration Prediction

Quantile	**NAE**	**PA**	**R^2**	**RMSE**	**IA**
0.1	0.199540	0.927044	0.858209	17.107463	0.899953
0.2	0.159383	0.926801	0.857759	14.260407	0.931728
0.3	0.137367	0.926794	0.857746	12.549789	0.948058
0.4	0.126366	**0.927127**	**0.858362**	11.597187	0.955750
0.5	**0.122344**	0.927032	0.858187	**11.091921**	**0.960640**
0.6	0.126188	0.927117	0.858344	11.294835	0.960117
0.7	0.139011	0.927024	0.858172	12.278250	0.954978
0.8	0.164617	0.927116	0.858342	14.209117	0.943222
0.9	0.218045	0.926767	0.857697	18.171985	0.915483

Repeating the procedure revealed the best quantiles for next 2-day and next 3-day. Table 4 shows the best model using quantile regression for next day, next 2-day and next 3-day. The best quantile for next day is at 0.5, next 2-day at 0.6 and next 3-day at 0.6. This is because the data for next 2-day and next 3-day is more skewed to the right than the next day QR model. The different RH sign for next day and next 2 and next 3-day model is because RH influences PM_{10} when the sun rises at around 0900 to 1900 hours. The quantile regression for the next day gave the negative sign at the quantile 0.5 because the RH and PM_{10} is inversely related at the quantile. However, the quantile regression for next 2-day and next 3-day showed the positive correlation at quantile 0.6 because at this quantile, the correlation between RH and PM_{10} can be considered as positive value correlation. That is one of the advantages using quantile regression because the model can give more information at every qauntile and interpretable of result can be done at each quantile. This scenario was also happened for wind speed sign due to the strong wind in this site, which can transport and dilute the PM_{10} at 0800to 1700.

Table 4: Model Summary of PM_{10} Concentration Using Quantile Regression

Days	Models
Next day (0.5)	$PM_{10,t+1}$= 19.52 + 0.1ws - 0.001RH - 0.28T + 54.12SO_2 - 507.89NO_2 - 7.01CO + 0.96PM_{10}
Next 2-day (0.6)	$PM_{10,t+2}$ = 42.2 - 0.1ws + 0.005RH - 0.8T + 188.0SO_2 - 1019.7NO_2 - 12.7CO + 1.0PM_{10}
Next 3-day (0.6)	$PM_{10,t+3}$ = 63.4 - 1.0ws + 0.1RH - 1.3T + 402.5SO_2 - 1329.1NO_2 - 14.4CO + 0.9PM_{10}

Multiple linear regression analysis based on the ordinary least square (OLS) method have been developed for comparing performance between quantile regression (QR) and multiple linear regression (MLR). Table 5 showed the model for predicting PM_{10} concentration using MLR and quantile regression. The performance indicators reflected greater accuracy in next day PM_{10} concentration prediction compared to the next 2-day and next 3-day predictions. However, the result showed that quantile regression and MLR could predict future PM_{10} concentration accurately until the next 3-day.

The result also showed, quantile regression models give more accurate and less error compared with MLR. It happens because of the influence of outlier data for all the models. The result showed QR give better results from next day until next 3-day such as NAE (QR is 3.28% better than MLR for next day, 5.92% for next 2-day and 9.04% for next 3-day and R^2 also showed that QR is better than MLR for next day in 0.82%, 3.20% for next 2-day and 6.21% for next 3-day. It can be concluded that QR can predict better than MLR until next 3-day.

Table 5: Performance Indicators Between Quantile Regression and MLR Models

	Method	NAE	RMSE	PA	R^2	IA
Next day	**Quantile (0.5)**	**0.122**	**11.092**	**0.927**	**0.858**	**0.961**
	MLR (OLS)	0.126	11.374	0.923	0.851	0.959
Next 2-day	**Quantile (0.6)**	**0.152**	**14.200**	**0.880**	**0.772**	**0.935**
	MLR (OLS)	0.161	14.815	0.865	0.748	0. 923
Next 3-day	**Quantile (0.6)**	**0.166**	**15.744**	**0.848**	**0.718**	**0.911**
	MLR (OLS)	0.181	16.799	0.823	0.676	0.895

4. Conclusion

The result shows that the quantile regression model is a good alternative to the multiple linear regression method. Quantile regression give more accurate result as compared to multiple linear regression such as average of performance indicators for QR is 1.45% better than MLR for next day, 3.3% better for next 2-day and 5.36% for next 3-day. Similar conclusions were found by a previous study [6]. However, by applying this model as the average hourly data to daily data input will create problem for all the parameters that influenced PM10 during the daytime like ws and RH. This will lead to the weakness of this model. This model is hoped to be useful for helping relevant government authorities to carry out suitable action to reduce the impact of air pollution in Seberang Perai, Malaysia.

5. Acknowledgements

This study is funded by Universiti Sains Malaysia under the Grant 304/PAWAM/60311017. Thank you to Universiti Sains Malaysia and Universiti Teknologi MARA for providing financial support to carry out this study and also sincere appreciation to the Department of Environment Malaysia for its support.

6. References

[1] R. Koenker and G. Bassett. *Regressiom Quantile.* Econometrica. 1978, 46, pp. 33-50.

[2] D. Baur, M. Saisana and N. Schulze. *Modelling the effect of meteorological variables on ozone concentration – a quantile regression approach,* Atmospheric Environment, 2004, 38, pp. 4689-4699.

[3] Hao, Lingxin, and Daniel Q. Naiman, *Quantile Regression*, Sage Publications, 2007.

[4] S.I.V Sousa, J.C.M Pires, F.G. Martins, M.C. Pereira and M.C.M Alvim-Ferraz, *Potentialities of quantile regression to predict ozone concentrations,* Environmetrics, 2009, 20, pp. 147-158.

[5] J. Mata, and J. Machado. *Firm start-up: a conditional quantile approach*. European Economic Review, 1996, 40, pp 1305-1323

[6] J.C.M Pires, F.G. Martins, S.I.V Sousa, M.C.M. Alvim-Ferraz, M.C. Pereira. *Prediction of the Daily Mean PM_{10} Concentrations Using Linear Models,* American Journal of Environmental Sciences, 2008, 4(5), pp 445-453.

[7] A.A. Kudryavtsev. *Using Quantile Regression for rate-making*, Insurance, Mathematics and economics, 2009, 45, pp 296-304.

[8] R. Koenker and K.F. Hallock, *Quantile Regression an introduction*, Journal of Economic Perspektives, 2001, 15, pp 143-156.

[9] H.C. Lu, *Estimating the Emission Source Reduction of PM_{10} in Central Taiwan*. Journal of Chemosphere, 2003, 54, pp 805-814.