Estimation of Beijing Air Quality Index Using Baidu Search Entries

Fengyuan Pan

***Abstract*—Protecting the environment while sustaining economic growth is a tough task for every country in the world, especially for China. China has required major cities to publicise their Air Pollution Index since 2000 (changed to Air Quality Index in 2012). Since then, the AQI has become one of the critical indicators for the central government to assess the local governments' performance. Comparing official AQI data from the US Embassy and 35 Beijing air quality monitoring stations, result reveals a significant manipulation of AQI data (to just below the Blue Sky threshold of 100). This research aims to find a way to predict the true AQI values through search entries in Baidu – the largest search engine in China. This would remove the need to rely on the data reported by the air quality monitoring stations, which seems to be unreliable. 73 search entries relating to air pollution and haze were collected from Baidu to run a LASSO (least absolute shrinkage and selection operator) analysis. To justify the LASSO analysis and find out the shrinkage factor, cross-validation method was used. After the LASSO analysis and cross-validation process, 33 predictors remained to predict AQI from search entries with R^2 0.69. These results indicate that search entries can be an alternative way to predict AQI with 69% prediction accuracy. In addition, due to limited time, there are only 73 search entries included in the dataset. For future research, a much higher prediction accuracy would be expected if more than 500 search entries included.**

***Index Terms*—Air quality index prediction, search entries, justification of air quality index, lasso, cross-validation.**

I. Introduction

China is one of the worst worldwide offenders in terms of air pollution: half of its population experience long-term exposure to unhealthy $PM_{2.5}$ concentrations [1]. Whilst experiencing benefits from steady GDP growth over the last 30 years, China now possesses eight of the world top 30 most polluted areas [2]. With a regional decentralised authoritarian regime, it has been argued that local governments in China have been incentivised to ignore environmental pollution to achieve higher GDPs [3]. However, as environmental problems become more serious, the central government is now implementing environmental measures in the performance of local governments. To move towards air quality improvement, especially in large cities such as Beijing, China has brought in two regulations.

Firstly, the air pollution indexes (APIs) from 37 large cities have been publicised since 2000, aiming to monitor and evaluate the performance of local governments [4]. The API was replaced by air quality index (AQI) in 2012, which moved to include three more pollutants ($PM_{2.5}$, O_3, CO), in addition to those covered by the API (SO_2, NO_2, PM_{10}) [4]. These measures of publicising AQI data have vastly increased public awareness on air pollution.

Secondly, starting from 2007, cities with 85% (increased from 80% in 2003) 'Blue Sky Days' in a calendar year are awarded as 'National Environmental Protection Model Cities'. The blue-sky days are defined as a day with API equal or smaller than 100, and the award of the model city is related to the promotion of government officials [4]. It has been argued that local governments manipulate and under-report the AQI data to achieve the standards of a 'National Environmental Protection Model City' [5], [6], [7]. AQI data reported by the Beijing monitoring station has been proved to be manipulated, compare to unmanipulated reports provided by the US Embassy.

In this paper, AQI data was collected from the US Embassy and thirty-five air quality monitoring stations in Beijing between 2014 to 2017. The manipulation of AQI figures is evident in Figure 1, which reveals the under-reporting AQI to just below AQI100 (the indicator of 'Blue Sky Day') from the Beijing monitoring stations. Figure 1 is a histogram of the numbers of days that Beijing air quality monitoring stations reported specific AQI data over the period 2014 to 2017. The red line represents AQI 100, and it can be clearly seen that there are more days reported in AQI 90-100 then AQI 100-110. This reveals that the local government intended to manipulate AQI data to just below 100 to achieve the 'Blue Sky Days' standard. Hence, there needs to be a more reliable method of reporting AQI data, rather than using the official AQI data reported by Beijing monitoring stations.

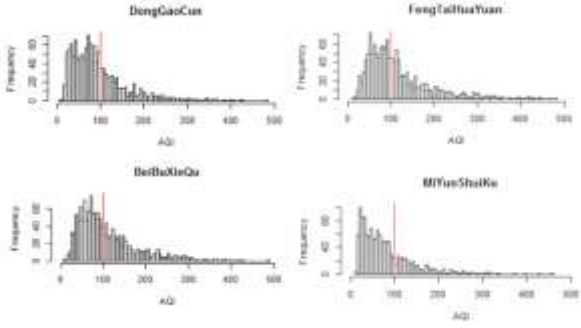

Fig. 1. AQI data from Beijing monitoring stations (2014-2017)

This research proposes a way to predict the Beijing AQI using search entries in Baidu – the largest search engine in China. The air quality index from the US Embassy was regarded as unmanipulated data and used as the dependent variable in the regression. The frequency of 73 air pollution-related search entries, such as 'haze', 'mask', and 'Beijing air quality index', were analysed with LASSO (least absolute shrinkage and selection operator) and cross-validation method. The LASSO was used to filtrate

Manuscript received April 28, 2018; revised June 12, 2018.

Fengyuan Pan is with the University College London, UK. (Email: fengyuan.pan.15@ucl.ac.uk).

large set of predictors with a penalised term or a shrinkage factor, and cross-validation method was used to determine the value of shrinkage factor. The computing R^2 of the model after LASSO analysis was 0.69, indicating a strong prediction power of search entry for Beijing AQI.

II. Data and Method

This section will expound the data and research method in this paper.

A. Search entry from Baidu

There are, in total, 73 search entry terms collected from Baidu Index webpage (http://index.baidu.com) considered in this study. Baidu is the biggest search engine in China, and is regarded as Chinese version of Google. Table I shows a brief overview of some of the main search entries and the appendices report the full list of search entries collected. The data for these search entries was collected for the Beijing area only, not for the whole of China.

Search entries have a number of advantages over the data reported by monitoring stations: they are decentralised so are not prey to single acts of manipulation, their data is transparent and easily available, and they can be measured from any location (to name but a few).

There are three main categories of search entries collected. The first category is entries that are related to the weather and air quality in Beijing, such as 'Beijing PM2.5', 'Beijing Air Quality Index', 'Haze in Beijing', 'Beijing Dust Storm'. The second category is commodities that people tend to use relating to the air quality in Beijing. For example, 'anti-haze mask', 'dustproof mask', 'air purification machine' and other related terms. The third category are terms which relate to haze or air quality as a whole. These entries include 'Copenhagen climate conference', 'cough', 'class suspension' and other entries that are caused by or largely related to air quality. Although these entries do not have as strong semantic links to air quality as the first two categories, it is reasonable to assume that they would be searched at a higher frequency when the general public have a higher awareness of air quality.

TABLE I: List of Search Entry

Variables	Average	Standard deviation	Minimum	Maximum
$PM_{2.5}$	2774.504	3666.613	571	75559
Haze	904.028	1763.435	163	28384
Beijing Air Quality Index	2563.187	2721.901	486	56984
Anti-haze mask	129.8615	146.9468	0	2038

B. Air quality data

This research used air quality data from 35 China-based air quality monitoring stations in Beijing and one monitoring station set up by the US Embassy. These stations measured the daily Air Quality Index between 2014 and 2017. The data was collected from the Beijing Municipal Environmental Protection Bureau (http://www.bjepb.gov.cn/) and the US Embassy (http://www.stateair.net/web/post/1/1.html). The data was reported hourly by these two stations, so this was manipulated in this study to give a daily average that could be compared to Baidu search entries.

The AQI data from Nongzhanguan and the US Embassy were the main data that the data analysis was conducted on. The data from the other Beijing air quality monitoring stations were only used to detect the AQI discontinuity and data manipulation over the period. The AQI data from the US Embassy is regarded as un-manipulated AQI data and is used as the dependent variables in our LASSO regression. This is because several studies have been able to attest to its veracity [8]. Although data reported by US Embassy is correct, the only monitoring station set up by US Embassy is not representative enough for the whole Beijing area. Hence, there is a need for justification of AQI monitored by other air quality monitoring station.

C. LASSO

When considering a good regression model for a large number of predictors, prediction accuracy is the key factor [9]. In terms of prediction accuracy, the aim is to find a model that provides low bias and low variance [9]. To achieve this, some coefficients may be shrunk or set to zero [10]. In addition, to increase the prediction power, a subset of key predictors that provides the most powerful predictive power will be extracted from the large predictor set (ibid.).

By using the ordinary least square (OLS) for estimation, the results are usually deducted from minimising the residual squared error [9]. However, when considering a regression with a large number of predictors, OLS is not suitable for two reasons. Firstly, the OLS estimates usually have large variance, although it aims to minimise this [10]. Therefore, the prediction accuracy is adversely affected. Secondly, the OLS estimate does not set any coefficient to zero, meaning it is not easily interpretable if there are a large number of predictors [10], which is the case in this study.

Consequently, when using the 73 search entries to predict the AQI in Beijing, LASSO was used. Lasso is a linear regression that aims to minimise the sum of squared errors with a penalised term [10]. It is a regression method that uses variable selection and regularisation processes to perform more accurate and interpretable statistical model to predict relationships between two variables [10]. To achieve a higher prediction accuracy and interpretability, the LASSO model shrinks some coefficients and set the others to 0.

LASSO has been used for prediction in a wide range of areas. For example, it has been used as an effective method for risk prediction in healthcare, where a large number of predictors are used to predict clinical outcomes [11], [12]. In addition, LASSO has been used in the financial markets to predict stock price movements [13], [14].

III. Data Analysis

A short analysis was performed to see the correlation between search entries and the AQI data monitored by US Embassy. Table II shows the results of a simple regression analysis between the AQI values reported by the US and four of the main search entries identified. There is certainly a positive correlation present. All four regressions have a high t-value, showing the regressions are statistically significant. As a result, preliminary results seem to show that entries are effective and efficient predictors for the AQI data. If a preliminary analysis of 4 search terms shows a statistically

significant correlation, the power of 73 search terms should certainly be investigated.

TABLE II: SIMPLE REGRESSION ANALYSES BETWEEN SEARCH ENTRY KEYWORDS AND US EMBASSY DATA

	'PM2.5'	'Haze'	'Beijing Air Quality Index'	'Anti-haze mask'
Coefficient	0.009 (20.57)	0.015 (15.63)	0.014 (23.85)	0.192 (16.54)
Standard error	0.000	0.001	0.001	0.012
R^2	0.242	0.156	0.301	0.171
Observations	1321	1321	1321	1321

Table III shows the results of that analysis, finding that that the frequency of the four search entries on separate days were indeed positively correlated to each other. The high correlation makes the variance for the slope coefficient large, leading to a less statistically significant coefficient. To reduce the variance of the model, a larger sample size could be used but this would not be an efficient use of time. To overcome this problem, conventionally, one regressor of the pairs with a high correlation coefficient is dropped in order to reduce the variance. Hence, entries with high correlation coefficient, for example, either $PM_{2.5}$ or Beijing Air Quality Index will be omitted from the list of predictor search terms since their correlation coefficient is an unusually high 0.912.

TABLE III: CORRELATION COEFFICIENT BETWEEN THE INITIAL FOUR SEARCH ENTRY KEYWORDS

	PM2.5	Haze	Beijing Air Quality Index	Anti-haze mask
PM2.5	1			
Haze	0.774	1		
Beijing Air Quality Index	0.912	0.595	1	
Anti-haze mask	0.736	0.931	0.564	1

A. LASSO Analysis and Cross-validation

The second step of data analysis was to run a LASSO analysis to find the most powerful prediction model for the use of search entries to predict AQI. Generally, LASSO aims to find a shrinkage factor or penalised term to filtrate explanatory variables.

Before the results of the LASSO regression are given, the theory behind the analysis will be quickly expounded below.

Suppose that there is a set of data (x^i, y_i) with N cases where $i = 1,2,3 \ldots . N$. Each cases would have a single outcome y_i , and covariate $x^j = \left(x_{i1}, x_{i2}, \ldots . x_{ip}\right)^T$. The x_{ij} are standardised, meaning $\sum_i \frac{x_{ij}}{N} = 0$ and $\sum_i \frac{{x_{ij}}^2}{N} = 1$.

With $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2 \ldots . \hat{\beta}_p)^T$, the lasso estimates $(\hat{\alpha}, \hat{\beta})$ is defined as:

$(\hat{\alpha}, \hat{\beta}) = argmin\{\sum_{i=1}^{N} (y_i - \alpha - \sum \beta_j x_{ij})^2\}$, subject to $\sum_j |\hat{\beta}_j| \leq s$ (1)

Where $s \geq 0$ is a tuning parameter.

When s approaches infinity, the regression approaches OLS, since none of the predictors' coefficient are shrunk or set to 0. On the other hand, when s approaches 0, y will regress on a constant since all predictors will be set to 0.

1) Validation and cross-validation

Conventionally, errors and residuals are used to evaluate a model. However, this approach does not give an indication of how the model would fit unknown data. For example, when we use samples for OLS regression, higher numbers of variables lead to higher R^2 values. This can be a problem, since the R^2 will increase if more variables are added to the right-hand side, no matter whether these variables are related to the left-hand side or not. Hence, when using errors and residuals in OLS, models with larger numbers of variables are incentivised as they will give higher R^2 values and greater in-sample fit. However, this leads to the issue of over-fitting, where regression results are manipulated to fit the existing data (and usually focus on a limited set of data points). In this case, regression results do not guarantee the prediction power of out-of-sample datapoints. Previous research on justifying AQI data has suffered from the problem of over-fitting [15].

To overcome the over-fitting problem and assess how the model performs on new data, validation and cross-validation methods are adopted in this study. The validation method uses a randomly selected two-thirds portion of the sample called the 'training set', on which LASSO regression analysis is run and a model is developed. The rest of the data, called the 'validation set', is used to determine whether the model developed from the 'training set' holds up on new data values.

In this research, the parameter s was chosen based on the training set using cross-validation method. The chosen method for cross-validation is k-fold cross-validation. This method broke the training set into 10 equal size subsamples. Of the 10 samples, a single subsample will be used as validation data to test the model with shrinkage factor s and the remaining nine subsamples are used as training data [15]. The cross-validation process was then repeated 10 times with each of the 10 subsamples used exactly once as the validation data [15]. The shrinkage factor was determined to be the value which minimised the 10-fold cross-validation error.

2) LASSO analysis

After dividing the sample into training and validation sets, a LASSO regression was run in the training set. The US Embassy data was used as the independent variable in the regression to compute the coefficient and prediction errors. The dependent variable was the search frequency of 73 key search entries.

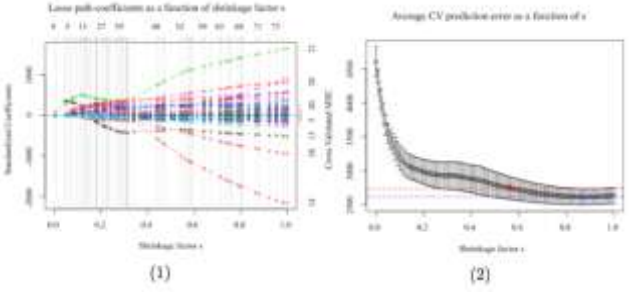

Fig. 2. LASSO regression.

Fig. 2 (1) shows the LASSO analysis between the 73 search entries and the AQI data reported by the US embassy. During the LASSO analysis, several breaks occurred, which allows the determination of the shrinkage factor s and the coefficient of the non-zero predictors. The figure reveals 73 breaks on the upper side of the graph. The coefficient of each predictor is standardised with the equation $Z = \frac{x-\mu}{\delta}$ where μ is the average and δ is standard deviation. The reason for

standardising the coefficients is to overcome the problem of different dimensionality between predictors, which allows easier evaluation and comparison between different types of random variable. Standardisation can weaken the economic interpretability of data. However, this is deemed not to be important in this research, as the predictive power of LASSO is what is required to answer the question at hand.

The coloured line in Figure 2 (1) represents the LASSO regression of each specific regressor (search entries). It is important here to quickly note that, as expounded in the explanation of LASSO, the shrinkage factor *s* is a relative index and so with a smaller *s*, each coefficient will receive more shrinkage. This means that there will be fewer predictors included in the model if more coefficients are set to 0, and the mean sqaure error (MSE) will be increase as a result (as shown in Figure 2.2). The right-hand side of the graph indicates the position of the regressors in the dataset. Some key search entries are important to note and correspond to the numbers at the right-hand side as follows: 21 - 'Beijing Air Quality Index', 5 - 'Controlling factor of haze', 14 - '$PM_{2.5}$', 38 - 'Air purifying machine'.

As previously mentioned, the shrinkage factor, *s*, is the value which minimises cross-validation errors. Figure 2.2 shows the average mean squared error (MSE) based on 10-fold sets for different value of *s*. The average MSE was the cross-validation error calculated with the 10 subsamples using the following equation (The $\hat{y}_i$ is the predicted value of y_i based on the model):

$$CV(s) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The vertical bar depicts one standard error. The blue point of Fig. 2.2 represents the model with lowest cross validation MSE, where the shrinkage factor s = 0.889. However, in this research the model with s = 0.576 will be chosen (the red point in Figure 2.2). The *s* of the red point is at the value that is one standard error away from the lowest MSE or lowest point on the curve. The reason for choosing the *s* at the red point is that if we have a set of models that are similar in explanatory power, we would choose the simplest model with the smallest number of predictors. The rationale for choosing the red point is demonstrated in mathematical equations as follows.

The shrinkage factor *s* of the red point set 23 predictors to 0, leaving 50 predictors left after LASSO analysis corresponding to s = 0.576. Some predictors and their coefficients are listed in Table 4, and the full list of predictors corresponding to s = 0.576 are shown in the appendix. The 23 predictors are set to 0 because they have a significantly lower explanatory value than the remaining 50 predictors. Hence, to increase the interpretability and prediction accuracy of the model, the LASSO analysis extracted 50 predictors with the highest predictive power from the original 73 search entries.

TABLE IV: EXAMPLE OF THE COEFFICIENT OF PREDICTOR SEARCH TERMS AFTER LASSO REGRESSION WITH SHRINKAGE FACTOR S=0.576

Searching entry	**Coefficient**
Pollution	0.2434
Haze weather	0.1002
Beijing Air Quality Index	0.0113
Human body	0.0106
3m mask	0.0035

As mentioned in the previous section, predictors with a high correlation coefficient increase the variance of the slope coefficient. Therefore, one predictor in pairs with high correlation coefficients (larger than 0.9) are selected to be dropped. For example, 'anti-haze mask' and '3m mask' have a correlation coefficient of 0.909. The '3m mask' search entry is thus dropped from the regression model, and 'anti-haze mask' remains. Another example is 'Beijing Air Quality Index' and '$PM_{2.5}$', which have a correlation coefficient of 0.912. One of these search entries will be omitted from the regression model.

To examine whether the model with 33 search entries has a strong predicting power on unknown data, the R^2 value of the model was computed. The result was an R^2 of 0.69, meaning 69% of the variation in AQI could be explained by the model computed from LASSO analysis. This indicates that the model proposed for predicting AQI through the search entries from Baidu has a moderate prediction power when out-sample data is used.

IV. DISCUSSION AND CONCLUSION

To conclude, the result shows that the US Embassy has a positive relationship with AQI monitored by other Beijing air quality monitoring station and different search entries. LASSO analysis was run to filtrate the explanatory variables, showing a high prediction accuracy. First, 23 search entries were set to zero with the penalised term s = 0.576. The remaining 50 search entries were then re-analysed to deduct variables with high correlation coefficients, one of which could then be discarded and so reduce the variance of the model. The prediction power was shown to be better than the ordinary least square method since the predictors left after the LASSO analysis were the ones with highest explanatory power. Moreover, the final model with 33 search entries showed an R^2 of 0.69. This drives our main conclusion from the analysis - frequency of search entries from Baidu can use as an effective estimator for Beijiing AQI with a 69% accuracy.

Although this paper achieves its research aims and addresses the research question, there are a few limitations (which provide fertile areas for further research). Primarily, the research only focuses on the Air Quality Index in Beijing only. This may lead to the problem that the model fits the Beijing area and does not fit other areas in China. Hence, future research should aim to collect data from major cities in China, such as Shanghai, Chengdu, Xian, where air pollution is also serious. Moreover, the data included in this research covers only the period 2014 to 2017. Further research could assess whether the relationship between AQI and search frequency is still statistically significant if a longer time period is considered.

Lastly, there are a few aspects to improve the research results in future with regards to the search entries used in the research. First, more search entries should be included to increase the reliability of the model. If over 1000 entries were included in the model, the prediction power would be massively increased and would lead to a higher R^2 for out-sample data. Second, if it was possible to access the regional data of search entries within city, such as entries in the region of Nongzhanguan, it would be possible to justify

the data from each of the monitoring stations. Moreover, besides using search entries from Baidu Index, more predictors could be included. For example, including the traffic index for each region and combining these predictors could allow for large-scale meta-analysis on conclusions made.

REFERENCES

[1] V. Li, Y. Han, J. Lam, Y. Zhu, and J. Bacon-Shone, "Air pollution and environmental injustice: Are the socially deprived exposed to more PM 2.5 pollution in Hong Kong?" *Environmental Science & Policy*, vol. 80, pp.53-61.

[2] Cbsnews. (2017). The most polluted cities in the world, ranked. [Online] Available: https://www.cbsnews.com/pictures/the-most-polluted-cities-in-the-world-ranked/31/

[3] C. G. Xu, "The fundamental institutions of China's reforms and development," *Journal of Economic Literature*, vol. 49, no. 4, pp. 1076-1151.

[4] Minisitry of Environmental Protection of the People's Republic of China (MEP) (2000) Annual Environmental Statement. (Chinese). [Online] Available at:http://www.zhb.gov.cn/hjzl/zghjzkgb/lnzghjzkgb/index_1.shtml

[5] S. Andrews, "Inconsistencies in air quality metrics: 'Blue Sky' days and PM10concentrations in Beijing," *Environmental Research Letters*, vol.3, no.3, p. 034009.

[6] Y. Chen, G. Jin, N. Kumar, and G. Shi, "Gaming in air pollution data? Lessons from China," *The B.E. Journal of Economic Analysis & Policy*, vol.12, no.3.

[7] D. Ghanem and J. Zhang, "'Effortless perfection:' Do Chinese cities manipulate air pollution data?" *Journal of Environmental Economics and Management*, vol. 68, no. 2, pp.203-225.

[8] Stateair.net. (2018). U.S. Department of state air quality monitoring program. [Online] Available: http://www.stateair.net/web/post/1/1.html

[9] F.Hayashi, "Econometrics," Princeton: Princeton University Press.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267-288.

[11] J. Musoro, A. Zwinderman, M. Puhan, G. ter Riet, and R. Geskus, "Validtion of prediction models based on lasso regression with multiply imputed data," *BMC Medical Research Methodology*, vol.14, no.1.

[12] M. Pavlou, G. Ambler, S. Seaman, O. Guttmann, P. Elliott, M. King, and R. Omar, "How to develop a more accurate risk prediction model when there are few events," *BMJ*, p.h3868.

[13] S. Roy, D. Mittal, A. Basu, and A. Abraham, "Stock market forecasting using LASSO linear regression model," *Advances in Intelligent Systems and Computing*, pp.371-381.

[14] O. Kohannim, "Discovery and replication of gene influences on brain structure using LASSO regression," Frontiers in Neuroscience, 6.

[15] L. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *International Journal of Forecasting*, vol. 16, no. 4, pp.437-450.

[16] S. Andrews, "Inconsistencies in air quality metrics: 'Blue Sky' days and PM10concentrations in Beijing," Environmental Research Letters, vol. 3, no.3, p.034009.

[17] Baidu (2018). About Baidu. [Online] Available: http://www.baidu.com/

Fengyuan Pan is with University College London, London, UK BSci Information for Business (Obtained at 2018)