Over the last three decades, air pollution has become a major environmental challenge in many of the fast growing cities in China, including Beijing. Given that any long-term exposure to high-levels of air pollution has devastating health consequences, accurately monitoring and reporting air pollution information to the public is critical for ensuring public health and safety and facilitating rigorous air pollution and health-related scientific research. Recent statistical research examining China’s air quality data has posed questions regarding data accuracy, especially data reported during the Blue Sky Day (BSD) period (2000 – 2012), though the accuracy of publicly available air quality data in China has improved gradually over the recent years (2013 – 2017). To the best of our understanding, no attempt has been made to re-estimate the air quality data during the BSD period. In this paper, we put forward a machine-learning model to re-estimate the official air quality data during the BSD period of 2008 – 2012, based on the PM2.5 data of the Beijing US Embassy, and the proxy data covering Aerosol Optical Depth (AOD) and meteorology. Results have shown that the average re-estimated daily air quality values are respectively 64% and 61% higher than the official values, for air quality index (AQI) and AQI equivalent PM2.5, during the BSD period of 2008 to 2012. Moreover, the re-estimated BSD air quality data exhibit reduced statistical discontinuity and irregularity, based on our validation tests. The results suggest that the proposed data re-estimation methodology has the potential to provide more justifiable historical air quality data for evidence-based environmental decision-making in China.
How **BLUE** is the Sky? Estimating the Air Quality Data in Beijing During the Blue Sky Day Period (2008-2012) by the Bayesian LSTM Approach

EPRG Working Paper 1912

Cambridge Working Paper in Economics 1929

Yang Han, Victor OK Li, Jacqueline CK Lam, and Michael Pollitt

**Abstract** Over the last three decades, air pollution has become a major environmental challenge in many of the fast growing cities in China, including Beijing. Given that any long-term exposure to high-levels of air pollution has devastating health consequences, accurately monitoring and reporting air pollution information to the public is critical for ensuring public health and safety and facilitating rigorous air pollution and health-related scientific research. Recent statistical research examining China’s air quality data has posed questions regarding data accuracy, especially data reported during the Blue Sky Day (BSD) period (2000 – 2012), though the accuracy of publicly available air quality data in China has improved gradually over the recent years (2013 – 2017). To the best of our understanding, no attempt has been made to re-estimate the air quality data during the BSD period. In this paper, we put forward a machine-learning model to re-estimate the official air quality data during the BSD period of 2008 – 2012, based on the PM$_{2.5}$ data of the Beijing US Embassy, and the proxy data covering Aerosol Optical Depth (AOD) and meteorology. Results have shown that the average re-estimated daily air quality values are respectively 64% and 61% higher than the official values, for air quality index (AQI) and AQI equivalent PM$_{2.5}$, during the BSD period of 2008 to 2012. Moreover, the re-estimated BSD air quality data exhibit reduced statistical discontinuity and irregularity, based on our validation tests. The results suggest that the proposed data re-estimation methodology has the potential to provide more justifiable historical air quality data for evidence-based environmental decision-making in China.

**Keywords** Blue Sky Day (BSD), Air Quality, Beijing, Data Irregularity, Bayesian LSTM, Data Estimation

**JEL Classification** C53, C63, Q53

Contact vhan@eee.hku.hk ; vli@eee.hku.hk ; jcklam@eee.hku.hk

Publication March 2019

Financial Support This research is supported in part by the Theme-based Research Scheme of the Research Grants Council of Hong Kong, under Grant No. T41-709/17-N.

www.eprg.group.cam.ac.uk
How **BLUE** is the Sky? Estimating the Air Quality Data in Beijing During the Blue Sky Day Period (2008-2012) by the Bayesian LSTM Approach

Yang Han¹,²*, Victor OK Li¹,²*, Jacqueline CK Lam¹,²*, and Michael Pollitt³

¹Department of Electrical and Electronic Engineering, The University of Hong Kong

³Energy Policy Research Group, Judge Business School, The University of Cambridge

*Corresponding Authors
² The authors have contributed equally significantly

Abstract

Over the last three decades, air pollution has become a major environmental challenge in many of the fast growing cities in China, including Beijing. Given that any long-term exposure to high-levels of air pollution has devastating health consequences, accurately monitoring and reporting air pollution information to the public is critical for ensuring public health and safety and facilitating rigorous air pollution and health-related scientific research. Recent statistical research examining China’s air quality data has posed questions regarding data accuracy, especially data reported during the Blue Sky Day (BSD) period (2000 – 2012), though the accuracy of publicly available air quality data in China has improved gradually over the recent years (2013 – 2017). To the best of our understanding, no attempt has been made to re-estimate the air quality data during the BSD period. In this paper, we put forward a machine-learning model to re-estimate the official air quality data during the BSD period of 2008 – 2012, based on the PM$_{2.5}$ data of the Beijing US Embassy, and the proxy data covering Aerosol Optical Depth (AOD) and meteorology. Results have shown that the average re-estimated daily air quality values are respectively 64% and 61% higher than the official values, for air quality index (AQI) and AQI equivalent PM$_{2.5}$, during the BSD period of 2008 to 2012. Moreover, the re-estimated BSD air quality data exhibit reduced statistical discontinuity and irregularity, based on our validation tests. The results suggest that the
proposed data re-estimation methodology has the potential to provide more justifiable historical air quality data for evidence-based environmental decision-making in China.

**Keywords:** Blue Sky Day (BSD), Air Quality, Beijing, Bayesian LSTM, Data Estimation

1. Introduction

Over the past decades, rapid socio-economic development has resulted in serious degradation in air qualities in many fast growing cities in China, such as Beijing. Accurately monitoring and reporting air qualities in China can help alert the public on how bad the air is in China and when one should avoid the bad air, while providing accurate air quality information to facilitate health-related scientific studies. Existing studies have shown that bad air carries a clear negative impact on physical and mental health (Pui et al., 2014; Zhang et al., 2017). However, it is difficult to obtain high-quality historical air quality data in China, and recent statistical studies (Chen et al., 2012; Ghanem and Zhang, 2014; Stoerk, 2016) have noted statistical irregularities in the official air quality data released to the public during the Blue Sky Day (BSD) period (2000–2012) when the number of BSDs was used for air quality evaluation. Air quality monitoring in Beijing dates back to the 1980s. The number of monitoring stations has increased from 12 in 2000 to 35 starting in 2012. Starting from June 2000, Beijing has released daily city-level air quality index (AQI)\(^1\) according to the National Ambient Air Quality Standard (NAAQS) introduced in 1996. Meanwhile, annual BSD information was published by the Beijing Environmental Protection Bureau (EPB) to facilitate public understanding of air quality trends from 1999 to 2012. A BSD is defined as a day when the AQI value falls below 100 (Andrews, 2008). The number of BSDs increased from 100 in 1998 to 286 in 2011 in Beijing (Beijing EPB, 1999; Beijing EPB, 2012). Since April 2008, the US Embassy in Beijing has started to report hourly PM\(_{2.5}\) based on its own sensors. Shortly after, Beijing EPB announced in June 2012 that the number of BSDs will no longer be used for air quality evaluation after 2012 (China Daily, 2012). In January 2013, following the new NAAQS introduced in 2012, Beijing EPB has officially launched a new air quality monitoring system. Since then, PM\(_{2.5}\) has been fully monitored by an automatic monitoring network in Beijing, with the hourly

---

\(^1\) In China, air pollution index (API) was used as the official name for the air quality index before 2013. We use AQI to denote the air quality index throughout this paper for consistency.
AQI and concentrations of six pollutants recorded from individual monitoring stations released to the public in real time.

BSD was first reported in Beijing when “Defending the Blue Sky” project was launched in 1998, and had become widely reported in major cities in China since then (Andrews, 2008). It served a policy-relevant metric for media reporting and evaluation of environmental performance of local governments. During the 11th Five-Year Plan period (2006 – 2010), Chinese key cities (including Beijing) were ranked by their environmental improvement. A city would receive a full score on air quality performance if the BSDs had exceeded 85% of the year (Chen et al., 2012). As environmental performance was a key factor affecting the promotion of Chinese officials (Zheng et al., 2014), Stoerk (2016) highlighted that local officials could be over-reporting the number of BSDs to the central government. A number of studies raised questions about the accuracy of air quality reporting in China during the BSD period. Andrews (2008) observed the station-level air quality data during 2001 – 2007 and noted the inconsistency between the official AQI values and the average AQI values based on the data collected from individual monitoring stations. Chen et al. (2012) raised the concern over the quality of air quality data via self-reporting by the local governments (Chen et al., 2012). Chen et al. (2012) conducted a statistical discontinuity test to examine discontinuity in AQI distribution during 2000 – 2009 and identified a discontinuity at the BSD threshold (AQI=100), with discontinuities increasing in magnitude across the years after 2003. Ghanem and Zhang (2014) adopted a panel matching approach to identify the circumstances under which statistical irregularities of air pollution data would occur for Beijing during 2001 – 2010, and suggested that such irregularities usually occurred during the days when they were least detectable, for instance, during the days of high visibility. Furthermore, Stoerk (2016) conducted a statistical regularity test on the Beijing US Embassy data and the official Beijing air quality data before and after the BSD period (2008 – 2013). He showed that the US Embassy data exhibited a high regularity over time, while the official Beijing data displayed patterns of irregularity during 2008 – 2012. Based on his finding, the irregularity of air quality data in Beijing had come to an end by 2012.

Existing statistical tests have examined the potential irregularity of air quality data in China. Recently, machine learning approaches, especially deep artificial neural networks, have achieved state-of-the-art performance in many tasks related to air quality modelling (Freeman et al., 2018; Han et al., 2018; Li et al., 2017a; Li et al., 2017b; Ong et al., 2016). A deep learning approach that attempts to tackle the
irregularity of air quality data in China may benefit future public health-related scientific research. Moreover, incorporating Bayesian methods into deep learning can reduce the overfitting of model parameters due to data sparsity and noise, while providing an uncertainty/trustworthiness measure for the prediction (Gal, 2016). Therefore, this study presents a Bayesian deep learning method to tackle air quality data irregularity during the recent BSD period (2008 – 2012) in China, using Beijing as an example. A Bayesian Long Short-Term Memory (LSTM) network model is constructed based on the relationship between the official city-level AQI values and the proxy data combined with the Beijing US Embassy daily PM\textsubscript{2.5} data after 2012. Then, based on the proxy data and the Beijing US Embassy daily PM\textsubscript{2.5} data reported during 2008 – 2012, we re-estimate the daily AQI values in Beijing during 2008 – 2012. Our result shows that across the five-year period, the re-estimated daily AQI generated from the proposed model is higher than the official daily AQI by 47 – 65 (54% – 70% for percentage increase) on average, while the re-estimated daily AQI equivalent PM\textsubscript{2.5} is higher than the official daily AQI equivalent PM\textsubscript{2.5} by 39 – 48 µg/m\textsuperscript{3} (56% – 67% for percentage increase) on average.

2. Data and Methodology

This study proposes a machine learning framework to re-estimate air quality data during the BSD period (2008 – 2012) in China. Our proposed framework consists of five components, as shown in Figure 1, namely, data collection, data pre-processing, model training, re-estimation of air quality data, and statistical test for air quality data validation.
Figure 1. Overall framework of our AI-driven BSD air quality re-estimation model

2.1 Data Collection

1) Air Quality Data: The Chinese official air quality data can be retrieved from the official website of the Ministry of Environmental Protection (MEP), China (China MEP, 2017). We collected the daily city-level AQI data from 5 June 2000 to 31 December
2012, and the daily city-level AQI data from 1 January 2014 to 30 June 2017\(^2\). The PM\(_{2.5}\) data from Beijing US Embassy can be downloaded from the official US Embassy website (US Department of State, 2017). We downloaded the hourly PM\(_{2.5}\) concentration data from 9 April 2008 to 30 June 2017, from the US Embassy, Beijing. We assumed that the PM\(_{2.5}\) data obtained from US Embassy, Beijing is valid since 2008. Moreover, previous studies showed that both the Chinese official data and the US Embassy data can fit the statistical regularity tests from 2013 onwards, which is likely due to the implementation of MEP’s new air quality standards starting from 2013, including PM\(_{2.5}\) monitoring and updated AQI calculation (Stoerk, 2016). We also performed a direct comparison between the US Embassy PM\(_{2.5}\) and the official PM\(_{2.5}\) observed at the nearest air pollution monitoring station from 2014 to 2017, and we found that they are highly correlated (Adjusted R\(^2\) = 97%; see Figure 2). Therefore, we assumed that the collected official air quality data can be used as the ground truths for city-level air pollution concentration in Beijing from 2014 onwards.

\(^2\) Beijing’s AQI data during 2013 – 2014 is not available on the China MEP’s official website.
2) **Proxy Data:** Previous studies showed that AOD and meteorology data can be used in the statistical modelling of air quality (Liu et al., 2012). We downloaded daily city-level AOD data from the US NASA’s AERONET database (US NASA, 2017) from 26 March 2008 to 21 May 2017. Eight features were selected based on data availability during the period of study. In addition, meteorology data, including temperature, pressure, humidity, visibility, precipitation, and wind speed, measured at the Beijing Capital International Airport, from 1 January 2008 to 31 December 2017, were collected from a weather service website (Weather Underground, 2018).

### 2.2 Data Pre-processing

Hourly US Embassy PM$_{2.5}$ values are aggregated to daily means. The input is a vector representing the historical data including PM$_{2.5}$, AOD, and meteorology. To reflect time trends, month and day of week are also included in the input vector. The output is a real value representing the corresponding daily city-level air quality (AQI). At first, each feature in the historical data is normalized into the range of zero to one. Then, for each feature, missing values in the time-series are imputed. More specifically, forward linear temporal interpolation of observed daily values is first used to fill in the gaps of less or equal than three days in time series which are caused by missing values. For the remaining missing values, mean values at the same month in the same year are used.

### 2.3 Model Training

The pre-processed data is fed into a Bayesian deep learning model for training. In this study, we focus on Bayesian Recurrent Neural Network (RNN), which is capable of modeling time series data (Fortunato et al., 2017). A Bayesian RNN model with network structure $f$ and parameters $\theta$ is denoted as $f_\theta$. During the post-BSD period (2014 – 2017), each observation of air quality and other covariates at day $t$ consists of the features $x_t$, including US Embassy PM$_{2.5}$ values and proxy values. The model input consists of the observations over the past $L + 1$ days (including current day $t$): $X_{t-L:t} = \{x_{t-L}, ..., x_t\}$, and the corresponding factual outcome $y_t$, i.e., daily city-level AQI.

---

3 The historical hourly PM$_{2.5}$ data monitored at the nearest official station to US Embassy, Beijing was downloaded on 10 August 2018, from an unofficial Chinese website: http://beijingair.sinaapp.com. We have verified this data using the official hourly air quality data we collected from the website of Beijing’s Environmental Monitoring Center in 2017.
Bayesian RNN model $f_{\theta}$ aims to find the optimal posterior distribution of the network weight parameters $\theta$, given the observed pairs $(X_{t-L:t}, y_t)$. We use LSTM as the recurrent unit in the network. A Bayesian fully connected linear layer is used to predict $y_t$, based on the final hidden state of Bayesian LSTM. The proposed model structure is shown in Figure 3. Conceptually, our proposed model is as follows:

$$h_t = \text{Bayesian-LSTM}(x_t, h_{t-1})$$

$$y_t = \text{Bayesian-LINEAR}(h_t)$$

In the network model, each weight parameter is a random variable with a Gaussian prior, and the weight at each time step has the same distribution. A diagonal Gaussian distribution is used as the variational posterior, and Bayes by Backprop is adopted to update the weight parameters of the network while minimizing the loss in terms of Mean Absolute Error (MAE) and Kullback–Leibler (KL) complexity cost (Blundell et al., 2015; Fortunato et al., 2017).

Figure 3. Architecture of the proposed Bayesian LSTM neural network structure

2.4 Re-estimation of Air Quality Data
1) AQI Re-estimation: AQI values during the recent BSD period (2008 – 2012) are predicted based on the fitted model \( f_\theta \) after training. First, a sample from the posterior of the network weight parameters is drawn randomly to obtain a model \( f_\theta_i \). Next, for each day during 2008 – 2012, daily city-level AQI values are re-estimated using this model with the corresponding US Embassy PM\(_{2.5}\) data and the proxy data. This is repeated \( N \) times, so that we can calculate the mean and the variance of AQI re-estimations, to reflect the uncertainties of the model parameters (Kendall and Gal, 2017). Given a BSD period \( T \), the final estimation of average AQI can be calculated as follows:

\[
\text{AQI}_i = E_{t \in T}[f_\theta_i(X_{t-L,t})]
\]

\[
\mu_{\text{AQI}} = \frac{1}{N} \sum_{i=1}^{N} \text{AQI}_i
\]

\[
\sigma_{\text{AQI}}^2 = \frac{1}{N} \sum_{i=1}^{N} (\mu_{\text{AQI}} - \text{AQI}_i)^2 + \text{data uncertainty,}
\]

where \( X_{t-L,t} \) is the input vector at day \( t \), and data uncertainty refers to the irreducible noise inherent in the data, which could be estimated by the residual sum of squares on an independent validation dataset (Zhu and Laptev, 2017). Finally, 95% Confidence Interval (CI) of AQI is calculated as follows:

\[
[\mu_{\text{AQI}} - z_\alpha \sigma_{\text{AQI}}, \mu_{\text{AQI}} + z_\alpha \sigma_{\text{AQI}}],
\]

where \( \alpha \) is set to 5%, \( z_\alpha \) is the critical value derived from the corresponding normal distribution.

2) AQI to PM\(_{2.5}\) Conversion: Before 2013, the concentrations of PM\(_{10}\), SO\(_2\), and NO\(_2\) were used by Beijing’s EPB for AQI calculation. Starting from 2013, following an update in AQI calculation, PM\(_{2.5}\), CO, and O\(_3\) are also included. For each pollutant, an individual AQI (IAQI) is calculated based on a linear interpolation of break points set by NAAQS (see Table 1). AQI is the maximum value of the IAQIs, and the pollutant with the highest value of IAQI is denoted as the primary pollutant. Primary pollutant is reported if AQI is greater than 50.

Based on the historical AQI observations, previous studies showed that particulate pollution is the dominant air pollution in Beijing (Stoerk, 2016). Therefore,
to compare our re-estimated AQI values with the observed AQI values, we assumed that PM$_{10}$ is the primary pollutant during 2008 – 2012 and PM$_{2.5}$ is the primary pollutant from 2013 onwards. Then, we converted the daily city-level observed AQI to the daily city-level PM$_{10}$ concentration, and the daily city-level re-estimated AQI to the daily city-level PM$_{2.5}$ concentration, based on the AQI calculation formula (Chen et al., 2015) as follows:

\[
C = \frac{AQI - IAQI_{Lo}}{IAQI_{Hi} - IAQI_{Lo}} (BP_{Hi} - BP_{Lo}) + BP_{Lo},
\]

where $C$ is the PM$_{2.5}$ or PM$_{10}$ concentrations, $IAQI_{Hi}$ and $IAQI_{Lo}$ are the nearby high and low values of AQI for PM$_{2.5}$ or PM$_{10}$ pollutant, $BP_{Hi}$ and $BP_{Lo}$ are the concentrations that correspond to $IAQI_{Hi}$ and $IAQI_{Lo}$ (see Table 1). Finally, we converted the daily city-level PM$_{10}$ concentration to the daily city-level PM$_{2.5}$ concentration using a seasonal adjusted ratio. Previous studies showed that the ratio between PM$_{2.5}$ and PM$_{10}$ in Beijing tends to be smaller during the spring and the summer (ranging from 0.4 to 0.6; Lv et al., 2016) and larger during the winter (ranging from 0.5 and 0.7; Sun et al., 2004). Therefore, we used the average ratio PM$_{2.5}$/PM$_{10}$ = 0.6 for the autumn and the winter, and the average ratio PM$_{2.5}$/PM$_{10}$ = 0.5 for the spring and the summer.

Table 1. IAQIs and their corresponding break points

<table>
<thead>
<tr>
<th>IAQI</th>
<th>Before 2013</th>
<th>After 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Daily SO$_2$ Conc. ($\mu$g/m$^3$)</td>
<td>Daily NO$_2$ Conc. ($\mu$g/m$^3$)</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>100</td>
<td>150</td>
<td>120</td>
</tr>
<tr>
<td>150$^4$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200</td>
<td>800</td>
<td>280</td>
</tr>
<tr>
<td>300</td>
<td>1600</td>
<td>565</td>
</tr>
<tr>
<td>400</td>
<td>2100</td>
<td>750</td>
</tr>
</tbody>
</table>

$^4$ This break point was not used for AQI calculation before 2013.
2.5 Statistical Validation for Air Quality Data

To test how accurate our re-estimated daily AQI values derived from our Bayesian deep learning model is, we undertook two statistical tests. We examined the discontinuity/irregularity before and after the re-estimation of daily AQI values across the period 2008 – 2012. Further, we also compared the discontinuity/irregularity of our re-estimated daily AQI distribution, with that of the US Embassy daily AQI distribution, across the period 2008 – 2012.

1) Discontinuity Test: Previous studies showed that there is a statistically significant discontinuity at the BSD threshold/cutoff (Ghanem and Zhang, 2014; Stoerk, 2016). In general, the proposed discontinuity parameter is an estimator of the log difference in height between the left and right limits of the density of the test variable at the cut-off (McCrary, 2008). In this study, we followed the discontinuity test proposed by these studies. The following procedures were taken to derive the discontinuity estimate. First, a first-step histogram was calculated to discretize the test variable. Second, two separate local linear regressions, which were weighted regressions using the bin midpoints to explain the height of bins, were derived on two sides of the cut-off to calculate the discontinuity. Third, $t$-statistic and $p$-value were used to infer the statistical significance of the discontinuity, as it was proven that this estimator is asymptotically normal. The larger the discontinuity estimate ($t$-statistic), the larger the statistical significance in terms of discontinuity at the BSD threshold/cutoff in the air quality data.

2) Benford’s Law: Benford’s Law is an observation about the frequency distribution of the first significant digits, which can be applied to detect irregularity in numeric data (Benford, 1938). Previous studies showed that air quality data could generally fit Benford’s Law (Stoerk, 2016). Following this study, we used the chi-squared statistic to compare the observed frequency distribution for the first two digits of the air quality data and the theoretical frequency distribution indicated by Benford’s Law. The larger the chi-squared statistic, the larger the statistical significance in terms of numeric irregularity in the air quality data.

3. Results

3.1 Model Evaluation
The proposed model was developed based on DeepMind Sonnet (Fortunato et al., 2017). In our experiment, a linear regression model was selected as the baseline:

\[
\text{City-level AQI} = \alpha + \beta \times \text{US EMBASSY PM}_{2.5} + \epsilon,
\]

where \( \alpha \) is the intercept, \( \beta \) is the regression coefficient, and \( \epsilon \) is the error term. To evaluate our proposed model, we used an 80/10/10 random split of the available data as the training set, the validation set, and the test set. We used Mean Absolute Percentage Error (MAPE) for model evaluation and comparison. We fine-tuned the hyper-parameters and chose the model with the lowest error rate of the validation set as the final model for daily AQI re-estimation from 2008 to 2012. We set the sample size \( N \) to 100 to obtain a reasonable estimation of the mean and the variance of the re-estimated AQI values.

On the test set, the MAPE of the final fitted model was 12% when using the mean of the posterior over the network weight parameters for prediction, while the MAPE of the baseline model was 23%. This suggested that our proposed model can better predict daily city-level AQI with an accuracy of 88%. In addition, we also evaluated the models using \( R^2 \), which measures the variance of AQI that can be explained by the input data. \( R^2 \) was 92% and 86% for the proposed model and the baseline model, respectively.

### 3.2 Results of Re-estimated Air Quality Data

We used the final fitted model to re-estimate daily city-level AQI values during 2008 – 2012. Figure 4 showed the trends of observed daily AQI values and the re-estimated daily AQI values which were aggregated into monthly means.
During this recent BSD period, the average of observed official daily city-level AQI values is 89, and the average of re-estimated daily city-level AQI values is 146 (95% CI: 139 to 154). Moreover, we also converted AQI to PM$_{2.5}$ for further comparison (see Figure 5). During this recent BSD period, the average of observed official daily city-level AQI equivalent PM$_{2.5}$ values is 70 µg/m$^3$, and the average of re-estimated daily city-level AQI equivalent PM$_{2.5}$ values is 113 µg/m$^3$ (95% CI: 106 µg/m$^3$ to 120 µg/m$^3$).

**Figure 4.** Observed monthly city-level AQI values and re-estimated monthly city-level AQI values, 2008 - 2012 (gaps due to missing data)
Figure 5. Observed monthly city-level AQI equivalent PM$_{2.5}$ concentrations and re-estimated monthly city-level AQI equivalent PM$_{2.5}$ concentrations, 2008 - 2012 (gaps due to missing data)

In general, the city-level AQI values and AQI equivalent PM$_{2.5}$ values after re-estimation were larger than that before re-estimation (see Figure 6). Table 2 also showed the average daily city-level AQI values and AQI equivalent PM$_{2.5}$ values before and after re-estimation across different years from 2008 to 2012.
Figure 6. Absolute difference between observed monthly city-level AQI/PM$_{2.5}$ values and re-estimated monthly city-level AQI/PM$_{2.5}$ values, 2008 - 2012 (gaps due to missing data)

Table 2. Average daily city-level AQI/PM$_{2.5}$ value before and after re-estimation

<table>
<thead>
<tr>
<th>Period</th>
<th>AQI (Before)</th>
<th>AQI (After)</th>
<th>Increase / Percentage Increase</th>
<th>PM$_{2.5}$ (Before)</th>
<th>PM$_{2.5}$ (After)</th>
<th>Increase (µg/m$^3$) / Percentage Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>87</td>
<td>134</td>
<td>47</td>
<td>63</td>
<td>102</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>54%</td>
<td></td>
<td></td>
<td>62%</td>
</tr>
<tr>
<td>2009</td>
<td>84</td>
<td>140</td>
<td>56</td>
<td>64</td>
<td>107</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>67%</td>
<td></td>
<td></td>
<td>67%</td>
</tr>
<tr>
<td>2010</td>
<td>91</td>
<td>150</td>
<td>59</td>
<td>73</td>
<td>116</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>65%</td>
<td></td>
<td></td>
<td>59%</td>
</tr>
<tr>
<td>2011</td>
<td>93</td>
<td>158</td>
<td>65</td>
<td>74</td>
<td>122</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>70%</td>
<td></td>
<td></td>
<td>65%</td>
</tr>
<tr>
<td>2012</td>
<td>92</td>
<td>146</td>
<td>54</td>
<td>72</td>
<td>112</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59%</td>
<td></td>
<td></td>
<td>56%</td>
</tr>
<tr>
<td>2008–2012</td>
<td>89</td>
<td>146</td>
<td>57</td>
<td>70</td>
<td>113</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64%</td>
<td></td>
<td></td>
<td>61%</td>
</tr>
</tbody>
</table>

3.3 Statistical Validation for Re-estimated Air Quality Data

Figure 7 showed the distribution of AQI values across different years from 2008 to 2012. Irregular air quality distribution could yield a larger discontinuity estimate and a smaller $p$-value at some specific cut-off points. A discontinuity point at the BSD
threshold AQI=100 was identified in official air quality data in China in two previous studies (Chen et al., 2012; Ghanem and Zhang, 2014). In these two studies, the discontinuity estimates could range from 0.41 to 0.96, and the p-value could be less than 0.01. Moreover, as demonstrated by another study (Stoerk, 2016), irregular air quality distribution could yield a larger chi-squared statistic in terms of Benford’s Law.
Figure 7. Daily US Embassy AQI distribution 2008 - 2012 and daily city-level AQI distribution before and after Bayesian LSTM re-estimation 2008 - 2012\(^5\)

\(^5\) US Embassy PM\(_{2.5}\) was converted to AQI using the AQI equation in Section 2.4. Probability density was the Gaussian kernel density estimation from the AQI histogram.
In Table 3, our statistical tests showed that (a) the discontinuity estimate of the US Embassy AQI distribution was insignificant, (b) the discontinuity estimate of daily official city-level AQI distribution before the re-estimation was significant, and was reduced significantly after the re-estimation, (c) the chi-squared statistic of Benford’s Law for the US Embassy AQI distribution is lower, and (d) the chi-squared statistic of Benford’s Law for daily official city-level AQI distribution before the re-estimation was larger, and was reduced significantly after the re-estimation. These tests suggested that the daily official city-level AQI distribution based on our re-estimation model follows better the natural air quality distribution.

Table 3. Statistical tests for the distribution of US Embassy AQI, official AQI and re-estimated AQI from 2008 to 2012

<table>
<thead>
<tr>
<th>Period</th>
<th>US Embassy AQI</th>
<th>Official AQI</th>
<th>Re-estimated AQI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discontinuity Estimate at AQI=100 (p-value in parentheses)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>0.14 (0.70)</td>
<td>0.88 (&lt; 0.01)</td>
<td>0.09 (0.77)</td>
</tr>
<tr>
<td>2009</td>
<td>0.15 (0.57)</td>
<td>0.43 (0.20)</td>
<td>0.10 (0.71)</td>
</tr>
<tr>
<td>2010</td>
<td>0.07 (0.80)</td>
<td>1.02 (&lt; 0.01)</td>
<td>0.22 (0.29)</td>
</tr>
<tr>
<td>2011</td>
<td>0.04 (0.85)</td>
<td>1.12 (&lt; 0.01)</td>
<td>0.39 (0.27)</td>
</tr>
<tr>
<td>2012</td>
<td>0.03 (0.90)</td>
<td>0.33 (0.33)</td>
<td>0.25 (0.36)</td>
</tr>
<tr>
<td>2008 – 2012</td>
<td>0.03 (0.78)</td>
<td>0.49 (&lt; 0.01)</td>
<td>0.13 (0.29)</td>
</tr>
<tr>
<td></td>
<td><strong>Chi-squared Statistic of Benford’s Law</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>134</td>
<td>175</td>
<td>185</td>
</tr>
<tr>
<td>2009</td>
<td>144</td>
<td>359</td>
<td>224</td>
</tr>
<tr>
<td>2010</td>
<td>131</td>
<td>372</td>
<td>149</td>
</tr>
<tr>
<td>2011</td>
<td>140</td>
<td>443</td>
<td>175</td>
</tr>
<tr>
<td>2012</td>
<td>133</td>
<td>380</td>
<td>173</td>
</tr>
<tr>
<td>2008 – 2012</td>
<td>377</td>
<td>1198</td>
<td>583</td>
</tr>
</tbody>
</table>

4. Conclusion

Existing studies examining and verifying the statistical irregularity of official air quality data collected during the BSD period in China have motivated our follow-up study to identify ways of resolving this irregularity. To the best of our knowledge, our study is
the first one to re-estimate the irregular air quality data in Beijing, China during the BSD period of 2008 – 2012, using a data-driven Bayesian deep learning approach, with the US Embassy PM$_{2.5}$ data and proxy data, including AOD and meteorology data across 2008 – 2017, as the input data. Our results have shown that the Bayesian LSTM air quality re-estimation model achieves an accuracy of 88%, with exhibited reduced discontinuity and irregularity across the five-year BSD period. During 2008 – 2012, the re-estimated AQI was higher than the official AQI by 64% on average, and the re-estimated AQI equivalent PM$_{2.5}$ was higher than the official AQI equivalent PM$_{2.5}$ by 61% on average, suggesting that the official air quality values reported during the BSD period may be lower than their natural values. The use of reliable and consistent air quality data has significant implications for evidence-based environmental research/decision-making in China. Our proposed data re-estimation methodology offers a means to fix the data irregularity challenge of historical air quality data in Beijing, during the period of 2008 to 2012, where the re-estimated air quality dataset can be used to more justifiably inform the health impacts of air pollution and the effects of air pollution control regulations in Beijing during this period.

**Acknowledgement**

We gratefully acknowledge the provision of PM$_{2.5}$ data by Beijing US Embassy, China, and AOD data by NASA, USA. This research is supported in part by the Theme-based Research Scheme of the Research Grants Council of Hong Kong, under Grant No. T41-709/17-N.

**References**


