UNIVERSITY OF CAMBRIDGE | Energy Policy Research Group

How BLUE is the Sky? Estimating 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

For more than three decades, air pollution has become a major environmental challenge in many of the fast growing cities in China, including Beijing. Given that 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, while facilitating rigorous air pollution and health-related scientific studies. A Blue Sky Day (BSD) is defined as a day when the air quality index (AQI) value falls below 100. Recent statistical studies examining China's air quality data have posed questions regarding data accuracy, especially the AQI values reported during the BSD period (2000 – 2012) when the number of BSDs was used for air quality evaluation, even though the accuracy of publicly available air quality data in China has improved substantially over the recent years (2013 – 2017). Until now, no attempts have been made to re-estimate the air quality data during the BSD period. In view of this, we propose a machine-learning model to re-estimate the official air quality data during the recent BSD period, from 2008 – 2012, utilizing the $PM_{2.5}$ data reported by the Beijing US Embassy and proxy data including Aerosol Optical Depth (AOD) and meteorology.

Our proposed framework consists of five components, namely, data collection, data pre-processing, model training, re-estimation of air quality data, and statistical test for air quality data validation. First, we collected historical data, including air quality data and proxy data, from 2008 to 2017. Second, we performed data normalization and missing data interpolation on the historical data. Third, the pre-processed data was fed into a Bayesian deep learning model for training. More specifically, a Bayesian Long Short-Term Memory (LSTM) network model was constructed based on the relationship between the official city-level AQI values and the proxy data combined with the Beijing US Embassy daily $PM_{2.5}$ data after 2012.

Next, based on the proxy data and the Beijing US Embassy daily $PM_{2.5}$ data reported during 2008 – 2012, we re-estimated the daily AQI values in Beijing during 2008 – 2012. We also converted the daily AQI values to AQI equivalent $PM_{2.5}$ concentrations for comparison. Finally, to test how accurate our re-estimated daily AQI values derived from our Bayesian deep learning model is, we undertook two statistical tests for air quality data validation. We examined the statistical discontinuity/irregularity before and after the re-estimation of daily AQI values across the period 2008 – 2012.

Our results have shown that the Bayesian LSTM air quality re-estimation model achieves an accuracy of 88%, with exhibited reduced statistical 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 precisely inform the health impacts of air pollution and the effects of air pollution control regulations in Beijing during this period.

Contact yhan@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 and by the Seed Fund for Basic Research from the University of Hong Kong, under Grant No. 201611159182.