# A Bayesian LSTM Model to Evaluate the Effects of Air Pollution Control Regulations in Beijing, China

Yang Han[a], Jacqueline CK Lam[a, b*], Victor OK Li[a, b], and David Reiner[b]

[a]Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong
[b]Energy Policy Research Group, Judge Business School, The University of Cambridge, Trumpington Street, Cambridge, United Kingdom

**Acknowledgement**
This research is supported in part by the General Research Fund of the Research Grants Council of Hong Kong, under Grant No. 17620920. We gratefully acknowledge the $PM_{2.5}$ data provided by Beijing US Embassy, China; and AOD data provided by NASA, USA. We also gratefully acknowledge the valuable comments and suggestions of our reviewers, and participants who attended our first HKU-Cambridge AI-WiSe workshop, titled "AI for Social Good", held in the Department of Computer Science and Technology, Cambridge, the United Kingdom, in May 2019, during which an earlier version of this work was presented.

**Declarations of interest**: none



* Corresponding author at: Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong. Tel.: (852) 3917 4843; E-mail address: jcklam@eee.hku.hk

# A Bayesian LSTM Model to Evaluate the Effects of Air Pollution Control Regulations in Beijing, China

**Abstract**

Rapid socio-economic development and urbanization have resulted in serious deterioration in air-quality in many world cities, including Beijing, China. This study attempts to examine the effectiveness of air pollution control regulations implemented in Beijing during 2008 – 2019 through a data-driven regulatory intervention analysis. Our proposed Bayesian deep learning model utilizes proxy data including Aerosol Optical Depth (AOD) and meteorology as well as socio-economic data, while accounting for confounding effects via propensity score estimation. Our results show that air pollution control regulatory measures implemented in China and Beijing during 2008 – 2019 reduced $PM_{2.5}$ pollution in Beijing by 11% on average. After the introduction of Action Plan for Clean Air in China and Beijing in late 2013, as compared to the hypothetical $PM_{2.5}$ concentration (without any regulatory interventions), the estimated $PM_{2.5}$ reduction increased dramatically from 15% in 2015 to 44% in 2018. Our results suggest that Beijing's air quality has improved gradually over the past decade, though the annual $PM_{2.5}$ pollution still exceeds the WHO threshold. In this regard, the air pollution control regulations introduced in Beijing and China tend to become more effective after 2015, suggesting a 2-year time lag before the stringent air pollution control regulations starting from 2013 takes any strong positive effects. Moreover, as compared to the air pollution control regulations introduced before 2013, newly introduced policy-making governance, which couples the policy-makings of the local jurisdictions with that of the central government, and the new policy measures that tackle the vested interests of the local stakeholders in Beijing and its nearby cities, alongside with the stringent local and national air pollution control regulations and plans, should help reduce air pollution and promote healthy living in Beijing over the longer term.

**Keywords:** air pollution control regulations, effects of regulatory interventions, Bayesian LSTM, propensity score, counterfactual analysis, causal inference

**Highlights**

- Aerosol optical depth, meteorology, and socio-economic data are collected
- A Bayesian deep-learning approach is proposed for regulatory intervention analysis

- Confounding effects are addressed by the propensity score estimation
- Air pollution controls reduced $PM_{2.5}$ in Beijing by 11% during 2008 – 2019

## 1. Introduction

Over the past few decades, rapid socio-economic development and urbanization have resulted in serious deterioration in air quality in Beijing, China. Air pollutants, especially $PM_{2.5}$ (particulates smaller than 2.5 micrometers in diameter), can lead to extremely detrimental health consequences, such as cancer, stroke, asthma, or heart disease (Pope III and Dockery, 2006; Pui et al., 2014). To provide in a timely manner, the critical health advice for Beijing's citizens based on scientific evidence, the introduction of real-time air pollution monitoring and reporting system in China has become increasingly crucial. Since April 2008, the US Embassy in Beijing has been publishing hourly $PM_{2.5}$ readings based on its own monitors installed in the embassy building. In January 2013, Beijing officially launched a new air quality monitoring system. Since then, $PM_{2.5}$ has been fully monitored by Beijing's automatic monitoring network, with hourly air pollution concentrations released by Beijing's Environmental Monitoring Center. A number of air pollution control regulations have been introduced by the government in China to control air pollution, with increasing stringency over the last two decades. Using Beijing as a case study, this study proposes a data-driven regulatory intervention analysis framework to study the causal relationship between air pollution control regulations and city-level $PM_{2.5}$ pollution concentration, based on available monitored air pollution data, proxy data including AOD and meteorology, and socio-economic data. The effects of air pollution control regulations in Beijing during 2008 – 2019 are evaluated. Our current work is an extension of an earlier work, which evaluates the effectiveness of air pollution control regulations in Beijing, China, during 2013 – 2017 (Han et al., 2018), by (1) adding socio-economic statistics in the input data, (2) taking account of the effective periods of air pollution control regulations, (3) extending the period of study to 2008 – 2019, and (4) reducing the confounding biases via propensity score estimation. The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 discusses our collected data and proposed the methods for regulatory intervention analysis. Section 4 presents our experimental results, followed by discussions on limitations of study and future directions. Section 5 highlights the policy implications. Section 6 concludes our study.

## 2. Related Work

## 2.1 Machine-Learning for Causal Inference and Policy Evaluation

Examining the causal effects from observational datasets has been a subject of serious attention in the fields of social science, policy, or medical science (Athey, 2017; Imbens and Rubin, 2015). Numerous theories attempted to account for the cause of an outcome/event, with the counter-factual framework being widely recognized and adopted (Rubin, 2005). Under such framework, the causal relationship between X and Y can be re-formulated as a counter-factual question. For example, in order to test whether there is a causal relationship between X and Y, a question such as "If X had not occurred, what would Y be?" is raised. However, it remains a difficult challenge to determine the counter-factual outcome. For any unit at a given time point, only the factual (instead of the counter-factual) outcome of a specific intervention could be observed (Rubin, 2005). Confounding factors were considered as the major barrier for determining with confidence whether causal relationships exist across a set of examined variables in a big dataset (Pearl, 2018). A proper and rigorous solution is to resort to the randomized control trials (RCT) and perform statistical adjustments to reduce the confounding biases. RCT is the "gold standard" for evaluating the causal effects while controlling for any confounding variables. However, in many cases, it was an infeasible task due to high research cost and ethical constraints (Stolberg et al., 2004). Hence, traditional statistical techniques, such as matching, re-weighting, and propensity score, were proposed to reduce confounding bias in observational studies (Athey and Imbens, 2017). However, these traditional methods were usually based on low-dimensional linear modelling, and failed to capture the complex non-linear relationships identified from high dimensional datasets (Hartford et al., 2017).

Advances in deep-learning have given rise to the remarkable success in overcoming many computational challenges that involve non-linear modelling of high dimensional data, such as natural language processing and computer vision (LeCun et al., 2015). However, these deep-learning models were mostly trained on datasets that carried noisy and unrepresentative big data (Caliskan et al., 2017), and often failed to account for the confounding effects when making causal inference (Marcus, 2018). As a result, spurious causations might occur and biased decisions made (Osoba and Welser IV, 2017). For supervised machine-learning algorithms, a fundamental shift from correlation analysis to causality analysis is needed to fully understand the causal relationship between an intervention (treatment or policy change) and an outcome. Recently, there has been a growing interest in using deep-learning models for causal inference and policy evaluation, based on techniques such as autoencoder (Atan et al., 2018) or variational autoencoder (VAE) (Louizos et al., 2017), propensity dropout (Alaa et al., 2017), propensity score estimation (Shi et al., 2019), domain adaptation (Shalit et al., 2017), multi-task learning (Alaa and van der Schaar, 2017), and generative adversarial network (GAN) (Yoon et al., 2018). These techniques aimed to improve the generalization ability of the models beyond observational data, and to reduce the confounding biases in high dimensional data. Moreover, since it is difficult to take into account all important confounders in counterfactual modelling, some other techniques, including the instrument variable (IV) method, were applied to the deep-learning models to control for any unobserved confounders, with additional assumptions taken (Hartford et al., 2017).

### 2.2 Evaluation of Air Pollution Regulatory Interventions

Many studies examined the effect of regulatory interventions on pollution concentrations in both the Chinese and the international context. Two major approaches, namely, (1) the environmental engineering approach and (2) the environmental economic approach, were adopted in these studies (Li et al., 2017d). The first approach provided an *ex ante* evaluation of policy impacts, by forecasting air qualities under different policy scenarios or constructing hypothetical air qualities in the absence of policy regulations, using physical and statistical modelling (Liu et al., 2012). The second approach performed an *ex post* evaluation of the causal effects of policy interventions, using experimental/quasi-experimental design and observational data, and methods such as difference-in-differences estimation (Chen et al., 2013), regression discontinuity design (RDD) (Li et al., 2017d), and panel data regression (Zheng et al., 2015). However, both approaches had drawbacks. The first one was often constrained by high computational costs, complex process modelling, and high uncertainties in emission inventories (Li et al., 2017d; Liu et al., 2010). The second one often failed to model the complex relationship between air pollution and other confounders such as meteorology and time trends, account for the uncertainties in input data and model parameters, and establish the causal relationship only after controlling for the confounders (Ferraro, 2009; Henneman et al., 2017).

Rapid development in machine learning made the adoption of data-driven regulatory analysis possible, with applications in resource allocation and causal inference (Athey, 2017). Recently, deep-learning approaches achieved state-of-the-art performance in air pollution estimation and forecasting (Li et al., 2017b; Li et al., 2017c; Ong et al., 2016), including $PM_{2.5}$ estimation, utilizing satellite-based Aerosol Optical Depth (AOD) as proxy data (Li et al., 2017a). However, in studies such as Li et al. 2017a, the temporal correlation between $PM_{2.5}$ pollution concentration and AOD is yet to be fully exploited by the neural network structure. Moreover, deep learning can still suffer from limited data source and low data quality when compared to other machine-learning techniques. Incorporating the Bayesian approach into deep learning can reduce network overfitting due to data sparsity and noise, and provide uncertainty measure for the prediction (Gal, 2016). However, a data-driven approach is yet to be applied to accurately estimate the counter-factual effects of air pollution regulatory interventions on air pollution outcomes, while accounting for the confounding biases.

## 3. Data and Method

This study proposes a machine-learning framework to provide counter-factual inference and evaluate the effects of air pollution regulatory interventions. The problem setup is similar to previous studies where potential outcome frameworks were adopted for causal inference (Alaa and van der Schaar, 2017; Atan et al., 2018). Differing from previous studies, our work focusses more specifically on evaluating the aggregate effect of multiple air pollution regulatory interventions. For each daily observation, there is a corresponding regulatory intervention state, which falls into two potential outcomes: the first potential outcome is a regulatory state where all regulatory interventions have been implemented as planned, whilst the second potential outcome is a regulatory state where no regulatory intervention has been implemented. Our goal is to learn how each feature-intervention pair is mapped to its corresponding factual outcomes, based on the observational air pollution samples collected during the period of study. Once the mapping model is trained, given an observed sample of air pollution outcomes after a group of regulatory interventions has been implemented, the counter-factual outcomes can be estimated for the scenario when the equivalent regulatory interventions are not implemented. Moreover, to estimate the causal effects of regulatory interventions, we follow the un-confoundedness assumption made in the potential outcome framework (Wooldridge, 2000). We assume that all important confounders that can potentially affect the regulatory interventions and the air quality outcomes have been taken into account in our model, and the confounding biases can be addressed via the propensity score estimation. Our proposed Bayesian deep learning policy intervention framework consists of four components, covering, data collection, data pre-processing, model training, and regulatory intervention analysis (see Figure 1).

[Insert Figure 1 about here]

### 3.1 Data Collection

We collected data consisting of air quality, AOD, meteorology, socio-economic, and air pollution regulatory measures from 2008 to 2019 (see Table 1 for a summary of data sources).

[Insert Table 1 about here]

### 3.1.1 Air Quality Data

We collected hourly $PM_{2.5}$ concentration data recorded at the US Embassy, Beijing from 9 April 2008 to 31 December 2019 (US Department of State, 2020). Only $PM_{2.5}$ observations with the quality control label "Valid" were included. Hourly $PM_{2.5}$ data were aggregated to daily means. Given that official $PM_{2.5}$ concentration data had not been available until 2013, air pollution observations at the US Embassy, Beijing for city-level $PM_{2.5}$ pollution concentrations in Beijing during the study period were used as the ground truths. Existing studies showed that Beijing's city-level $PM_{2.5}$ concentrations was highly correlated with $PM_{2.5}$ concentrations observed at the US Embassy, Beijing. Hence, it was reasonable to assume that the readings reported by the US Embassy can be used to represent the level of air-quality throughout the city in Beijing (Wang et al., 2013). To further examine the representativeness of the US Embassy $PM_{2.5}$ data, we collected official hourly station-level $PM_{2.5}$ concentration data from 1 January 2014 to 31 December 2019 using the data source provided in Zhang et al. (2019), and examined the correlation between the daily average $PM_{2.5}$ concentrations measured at the US Embassy, Beijing and the daily city-level average $PM_{2.5}$ concentrations measured at the 35 official stations in Beijing during 2014 – 2019. Result showed that the two measurements were highly correlated ($R^2$=96.2%; see Figure 2).

[Insert Figure 2 about here]

### 3.1.2 Proxy Data (AOD and Meteorology)

Previous studies showed that AOD and meteorology data can be incorporated into the statistical modelling to examine the effects of regulatory interventions on air pollution concentrations in Beijing (Liu et al., 2012). Our study had incorporated the AOD data into our statistical modelling. AOD observations at the city level were collected from the NASA MODIS satellite database from 26 March 2008 to 21 March 2019 (US NASA, 2020). Five features were selected based on data availability during the period of study, including AOD at 1020 nm, AOD at 870 nm, AOD at 675 nm, AOD at 440 nm, and precipitable water. AOD data points observed each day were aggregated into daily means. In addition, hourly city-level meteorology data, including temperature, relative humidity, wind speed, wind bearing, and visibility, across the period from 1 January 2008 to 31 December 2019, were collected from a weather data application program interface (API), based on the official data sources (Apple Inc., 2020). Hourly meteorology data were aggregated to daily means.

#### 3.1.3 Socio-economic Data

Previous studies showed that socio-economic data can be used as control variables to model statistically the effects of regulatory interventions on air pollution concentrations at the provincial-level in China (Zheng et al., 2015). In this study, we collected the yearly socio-economic statistics including the percentage of GDP generated from the secondary sector, the population density, and the number of vehicles, during the period of 2008 to 2019 (Beijing Municipal Bureau of Statistics, 2020; Beijing Transport Institute, 2020).

#### 3.1.4 Regulatory Measures Data

We identified major air pollution control regulations at the city- or the national-level during the period of 2008 to 2019 (DieselNet, n.d.; Lam et al., 2019; Zhang et al., 2016). These regulations were directly responsible for air pollution prevention and control in Beijing/China, with a strong focus on the energy and transportation sectors, including emission controls on the coal-fired power plants and the industrial facilities and vehicles, emission standards on cars and light trucks, optimization of energy structures and traffic systems, technological innovations of clean environment, emergency plans for high pollution episodes, and legal responsibilities. Some were updated during the period of study, including, the Air Pollution Prevention and Control Law in China (see Figure 3).

[Insert Figure 3 about here]

### 3.2 Data Preprocessing

The daily air quality data and the daily proxy data were combined to generate a tabular dataset, ranging from 9 April 2008 to 21 March 2019. The dataset was pre-processed for model training, validation, and evaluation. The data pre-processing procedure was listed as follows. First, a random 80/10/10 split of the data was used as the training set, the validation set, and the test set. Second, each dataset was converted into the input/output pairs. The input data consisted of two parts: a vector representing the historical daily proxy data (including AOD and meteorology) of the current day and the previous days, and a binary vector representing the current status of regulatory interventions. To account for the socio-economic variation, the corresponding yearly statistics were included in the input vector. To account for the unobserved time trends and recurrent effects, the month and the day of the week were included as the categorical features in the input vector. The output was a continuous value representing the corresponding city-level daily $PM_{2.5}$ concentration (i.e., $PM_{2.5}$ concentration observed at the US Embassy, Beijing). Then, missing data was filled in via iterative imputation (Buuren and Groothuis-Oudshoorn, 2010). To avoid any potential information leakages, the iterative imputer was constructed based on the training set, which was subsequently used to impute the validation and test datasets. Next, each input feature (except for the time trend) in the training set was standardized according to its mean and standard deviation; which were then used to standardize the corresponding feature of the validation and the test datasets. Finally, the data pre-processing procedure was repeated five times. Eventually, five datasets for model training, validation, and test were constructed from five random data splits.

### 3.3 Model Training

The pre-processed data was fed into a Bayesian deep learning model for training. During the period of study, the covariate data at day $t$ was denoted as $x_t$. The input data for day $t$ consisted of the observations over the past $L+1$ days (including the current day $t$) and the time trend: $X_t = \{x_{t-L}, \dots, x_t, \text{Month}_t, \text{Day of week}_t\}$. The regulatory status vector at day $t$ consisted of the status of $K$ regulatory interventions $I_t = \{I_t^1, \dots, I_t^k\}$, e.g., {Regulation 1 is implemented, Regulation 2 is not implemented, …, Regulation $K$ is not implemented} (see Figure 3 for the effective periods). We used zero or one to indicate the status of a particular regulatory intervention $I_t^k$, namely, one for "is implemented" and zero for "is not implemented". The output $y_t$ was the observed city-level air quality (i.e. $PM_{2.5}$ concentrations observed at the US Embassy, Beijing). The proposed framework had two potential outputs, the first one corresponded to $I_t$, where all regulatory interventions are implemented as planned, while the second one corresponded to a regulatory state where no regulatory interventions are implemented. A Bayesian deep-learning model with network structure $f$ and parameters $\theta$ was denoted as $f_\theta$. During the study period of length $T$, given the input $X_t$, the regulatory intervention status $I_t$, and the output $y_t = f(X_t,\ I_t)$, we aimed to estimate the counter-factual output $\widetilde{y_t} = f(X_t,\ \mathbf{0})$. The model $f_\theta$ aimed to find the optimal posterior distribution of the network weight parameters $\theta$, given the observed tuples $\{(X_t, I_t, y_t)\}_{t=1}^{t=T}$. To better address the confounding effects, a shared representation layer was used (1) to predict air quality based on the covariate and the regulatory intervention status and (2) to predict regulatory intervention status from the covariates (i.e., the propensity score estimation). By incorporating the propensity score estimation model into the proposed framework, the input features relevant for confounding effects could be distilled automatically (Shi et al., 2019). More specifically, we focussed on the Bayesian RNN, which is a particular type of Bayesian deep learning model capable of modelling time-series data (Fortunato et al., 2017). We used LSTM as the recurrent unit of the network. A Bayesian embedding layer was used to map the time trend vector into a vector of continuous values (Yi et al., 2018). Another Bayesian embedding layer was used to map the regulatory status vector to a vector of continuous values (Pham et al., 2017). Two Bayesian fully connected linear layers were utilized. One Bayesian fully connected linear layer was used to predict $y_t$, while the other Bayesian fully connected linear layer followed by a sigmoid function, was used to predict $I_t$ (Shi et al., 2019). Both of them were based on the shared representation, which consisted of three parts, the final hidden state of Bayesian LSTM ($h_t$), the embedded time trend ($e_t^1$), and the embedded regulatory intervention status ($e_t^2$). Conceptually, our proposed model was as follows:

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

$$e_t^1 = \text{Bayesian-Embedding}(\text{Month}_t, \text{Day of week}_t) \quad (2)$$

$$e_t^2 = \text{Bayesian-Embedding}(I_t) \quad (3)$$

$$y_t = \text{Bayesian-Linear}(h_t, e_t^1, e_t^2) \quad (4)$$

$$I_t = \text{Sigmoid}(\text{Bayesian-Linear-Propensity-Score}(h_t, e_t^1, e_t^2)) \quad (5)$$

To train our proposed model, we followed the work done by Blundell et al. (2015); Fortunato et al. (2017). In the network, each weight parameter was a random variable with a Gaussian mixture prior, and the weight at each time step had the same distribution. A diagonal Gaussian distribution was used as the variational posterior distribution, which is often computationally tractable and numerically stable, assuming that the network weights were uncorrelated. The loss function of the proposed model consisted of three components. The first part was the Mean Squared Error (MSE) loss calculated by the predicted and the observed air quality values, which is the most commonly used loss for predicting continuous values. The second part was the Binary Cross Entropy (BCE) loss calculated by the predicted and the observed regulatory intervention status, which is often used for multi-label classification. The BCE loss enforced the learned shared representation layer to account for the propensity score estimation, in order to address the confounding effects. The third part was the Kullback–Leibler (KL) divergence between the posterior and the prior distribution, which is a regularization term to penalize model overfitting. Bayes by Backprop was adopted to update the weight parameters of the network while minimizing the loss function, given the observed inputs (see Algorithm 1). The proposed model was trained via the shuffled mini batches, using a stochastic gradient descent (SGD) optimizer.

---

**Algorithm 1.** Bayesian LSTM Model Training via Bayes by Backprop

**Require:** training data $D = \{(X_t, I_t, y_t)\}_{t=1}^{t=T}$, epoch size $E$, batch size $B$, and learning rate $\alpha$

**For epoch from 1 to E**

**Repeat**

1. Sample a mini batch of size $B$ from the training data $D$ without replacement
2. Sample $\varepsilon \sim \text{Gaussian}(0, I)$, where $I$ is the identity matrix
3. Set network parameters $\theta = \mu + \sigma\varepsilon$, where $\mu$ and $\sigma$ are the mean and standard deviation, respectively
4. Compute the gradients of MSE loss plus BCE loss with respect to $\theta$ using normal back-propagation: $g_\theta^L$
5. Compute the gradients of $F(\mu, \sigma, \theta) = \log \text{Gaussian}(\mu, \sigma^2) - \log p(\theta)$ with respect to $\mu, \sigma, \theta$: $g_\mu^F, g_\sigma^F, g_\theta^F$, where $p(\theta)$ is the Gaussian mixture prior
6. Update $\mu = \mu - \alpha \frac{g_\theta^L + g_\theta^F + g_\mu^F}{B}$
7. Update $\sigma = \sigma - \alpha \frac{g_\theta^L \varepsilon + g_\theta^F \varepsilon + g_\sigma^F}{B}$

**Until all mini-baches are sampled**

**End**

**Return** fitted network model $f_\theta$

---

During the model training, the tuning hyper-parameters took into account the number of lagged observations (0 or 7; 0 indicated that no lagged observations were used, while 7 indicated that the past one week data was used for prediction), the embedding dimension of the regulatory intervention status vector (3 or 5), the number of hidden units used in the neural network (128 or 256), the batch size (32 or 64). For each data split (including the training, the validation, and the test dataset), the best hyper-parameters were selected based on the validation MSE. Moreover, the fixed hyper-parameters included the number of training epochs (30), the learning rate (0.01), the number of recurrent layers (1), the embedding dimension of the time trend vector (3; based on a configuration used by Yi et al. (2018)), the prior distribution of the Bayesian deep-learning model ($\pi$ = 0.25, $-\log\sigma_1$ = 0, and $-\log\sigma_2$ = 6; based on a configuration used by Blundell et al. (2015)).

### 3.4 Regulatory Intervention Analysis

After the model training, counter-factual outcomes, in the absence of regulatory interventions, were predicted to quantify the net effects of regulatory intervention, based on the fitted model $f_\theta$. More specifically, for each data split $j$, the regulatory intervention analysis was performed according to the following steps. First, a random sample was drawn from the posterior of the network weight parameters to obtain a model $f_{\theta_{i,j}}$. Next, the corresponding regulatory status vector was constructed with the hypothesis that no regulatory intervention was implemented and represented by a vector of zeros. Such hypothetical regulatory intervention status vector, after combining with the covariate data $X_t$, were used to re-estimate $PM_{2.5}$ concentration using model $f_{\theta_{i,j}}$. This was repeated $N$ times, such that the mean of $PM_{2.5}$ re-estimations could be calculated to account for the uncertainties of the model parameters (Kendall and Gal, 2017). During the study period of length $T$ (2008 – 2019 or a particular year such as 2017), the final estimation of $PM_{2.5}$ concentrations on day $t$ and the average regulatory effect (ARE) were calculated by the following equations:

$$\tilde{y}_t^{i,j} = \mathrm{E}[f_{\theta_{i,j}}(X_t, \mathbf{0})] \quad (6)$$

$$\mathrm{ARE}_{i,j} = \tilde{y} - y = \mathrm{E}_{t\in T}\left[\tilde{y}_t^{i,j}\right] - \mathrm{E}_{t\in T}[y_t] \quad (7)$$

where $\theta_{i,j}$ was the $i$th sample from the network weights posterior trained on the $j$th data split, $i$ ranged from 1 to $N$, $j$ ranged from 1 to 5, $T$ was in the *ex post* evaluation period, $y_t$ and $\widetilde{y_t}$ were the observed and counter-factual air quality, respectively, and $X_t$ is the covariate data. The number of posterior samples $N$ was set to 100, in order to obtain a reasonable estimation of the re-estimated air quality values. Note that the regulatory intervention analysis was performed for five times, using the models trained across different data splits. Finally, given that the air quality values may not follow a Gaussian distribution (see Figure 2), the final estimation of ARE with 95% confidence interval (CI) was calculated based on bootstrapping. More specifically, a list of ARE values of length $5 * N$ was resampled from $\{\mathrm{ARE}_{1,1}, \mathrm{ARE}_{2,1}, \ldots, \mathrm{ARE}_{N,5}\}$ with replacement, the resampled mean was subsequently calculated. This was repeated 10,000 times, and the 250 percentiles and the 9,750 percentiles of the resampled means were selected as the lower and the upper bound of the ARE during the study period, respectively.

# 4. Results

## 4.1 Baseline Selection and Model Evaluation

Previous research suggested that non-linear relationship might exist between $PM_{2.5}$ pollution concentration and other covariates data (Han et al., 2018). Hence, in our experiment, two non-linear machine-learning models, namely, Support Vector Regression (SVR) and Random Forest (RF), were selected as the baseline models. We used Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) for model evaluation and comparison. For Bayesian LSTM, we fine-tuned the hyper-parameters as listed in Section 3.3. For SVR, we fine-tuned three hyper-parameters, including the lagged observations (0 or 7), the kernel function (polynomial function or radial basis function) and the penalty parameter of the error term (0.1, 1, or 10). For RF, we fine-tuned four hyper-parameters, including the lagged observations (0 or 7), the number of estimators (10 or 100), the maximum depth of the tree (1, 16, or 32), and the maximum number of features ($n$, sqrt($n$), or $\log_2(n)$, where $n$ is the number of features). Finally, the models with the lowest MSE on the validation set were selected as the final models for further analysis.

The performance of the proposed model and the baseline models are shown in Table 2. Results have clearly revealed that the Bayesian LSTM model outperforms the baseline models. On the test set, the mean MAE of the proposed model is 20.3, while the mean MAE of the SVR and RF model are 22.1 and 22.4, respectively. The mean MAPE of the proposed model is 36.8%, while the MAPE of the SVR and RF model are 38.8% and 46.9%, respectively. Moreover, the standard deviation of the proposed model's performance is also the lowest as compared to the baseline models. This suggests that our proposed model can give a much better prediction of the out-of-sample data as compared to traditional machine-learning techniques, across different training/validation/test data splits. Note that the absolute/relative error rates of the proposed model remain high, partly due to the fact that some features inputs (which were irrelevant to the causal relationships according to the propensity score estimation) were considered as noise for air quality estimation. However, the causal estimation of ARE can be improved through such a trade-off between predictive accuracy and propensity score estimation (Shi et al., 2019).

[Insert Table 2 about here]

## 4.2 Regulatory Intervention Analysis

We used the final fitted Bayesian LSTM models across different data splits to estimate/simulate the average counter-factual air quality, based on the assumption that all regulatory interventions were not implemented, in order to examine the ARE of all regulatory interventions during 2008 – 2019. Figure 4 shows that the observed monthly average daily air quality and simulated monthly average daily air quality without any regulatory interventions during 2008 – 2019. The average of observed daily $PM_{2.5}$ concentration was 86 μg/m$^3$ during 2008 – 2019. Had the same set of regulatory interventions not been implemented before 2008, the hypothetical average daily $PM_{2.5}$ pollution would be 97 μg/m$^3$ (95% CI: 96 μg/m$^3$ to 99 μg/m$^3$). The average intervention effect of all regulatory interventions was 11 μg/m$^3$ (95% CI: 10 μg/m$^3$ to 13 μg/m$^3$). This implies that the aggregate effect of all air pollution regulatory interventions implemented during this period can lead to a 11% reduction in $PM_{2.5}$ pollution concentration on average. Based on Eq. (7), the relative reduction was calculated as ARE / $\tilde{y}$, where $\tilde{y}$ is the hypothetical average daily $PM_{2.5}$ pollution.

[Insert Figure 4 about here]

Table 3 shows the observed yearly average daily air quality and simulated yearly average daily air quality without any regulatory interventions during the period of study. Results have shown that the estimated $PM_{2.5}$ reduction due to the implementation of the set of air pollution regulatory interventions implemented during the 2008 – 2019 period on average was not as significant as expected on average (11%), even when a series of stringent air pollution control regulations and plans have been introduced during the period (see Figure 2). However, after the introduction of Action Plan for Clean Air in China and Beijing in late 2013, the estimated $PM_{2.5}$ reduction increased dramatically from 2% in 2014 to 15% in 2015. After 2015, the estimated $PM_{2.5}$ reduction increased up to 44% in 2018, and dropped to 37%[1] in 2019.

[Insert Table 3 about here]

### 4.3 Limitations of Study and Future Work

There are some limitations in our current study. First, the interpretability of the proposed Bayesian deep-learning framework for policy evaluation can be improved. Although the

[1] This only covers the data in the first quarter of 2019. We expect the average improvement would be changed (would likely be increased) when the full year data is incorporated into our model.

confounding variables have been addressed in the proposed model by incorporating a propensity score estimation layer, it remains difficult to understand which variables have contributed most to the confounding biases and the causal relationships, and when/where the proposed model works better as compared to the traditional statistical methods for policy evaluation (such as propensity score estimation using a logistic linear regression model). Future work can focus on an interpretable machine-learning framework for policy evaluation. Second, our study only examines the aggregate effect of air pollution control regulations and plans during the period of study. More sophisticated analysis is needed to understand the individual effect of a particular regulatory intervention on air quality, and over a particular sector. Third, the proxy data is still very limited. Additional data, such as satellite images and industrial outputs published by the government's statistical bureau, can be included in the regulatory analysis to improve the accuracy of policy evaluation. Finally, this study only uses a single-point $PM_{2.5}$ monitor data. Given that the air quality can vary across different parts of Beijing, in future work, more fine-grained air quality data obtained from the 35 official stations can be used to evaluate the effects of air pollution regulatory interventions since 2013.

## 5. Policy Implications

Evaluating the effects of air pollution control regulations has significant implications for environmental policy-makings in China and the rest of the world. We have identified two major policy implications with regard to our proposed Bayesian deep-learning policy intervention study methodology and results. First, our proposed data-driven regulatory analysis methodology can estimate the aggregate effects of air pollution control regulations and plans with reduced confounding biases and higher accuracies, when compared to other machine-learning techniques. Hence, our model can provide the needed evidence to support evidence-based air pollution policy-makings. For instance, the governments can perform *ex post* evaluation on air pollution control regulations to test the effectiveness of the regulations they implemented based on our model. Second, though the annual $PM_{2.5}$ pollution concentration in Beijing remains far beyond the WHO threshold ($10\mu g/m^3$), our results suggest that Beijing's air quality has been improved gradually over the past decade (11% improvement on average; see Table 3). The air pollution control regulations implemented during 2008 – 2019 tend to be more effective after 2015, i.e., after the air pollution control laws in Beijing/China have been further revised and stringent air pollution control action plans have been implemented in Beijing/China since 2013 (see Figure 3 and Table 3). This suggests that there is a 2-year time lag before the stringent air pollution control regulations in Beijing/China taken any strong positive effects. As compared to the regulatory interventions introduced before 2013, policy-makings that coordinate that of the local jurisdictions and the central governments (such as the guidelines on air quality monitoring and law enforcement introduced by provincial authorities, effective in 2016), and laws and policies that tackle the vested interests of the local stakeholders in Beijing and neighbouring cities (such as the joint action plan for air pollution control in Beijing-Tianjin-Hebei Region, effective in 2013), alongside with the stringent air pollution control regulations and plans, can help reduce air pollution and promote healthy living in Beijing over the longer term (Lam et al., 2019).

## 6. Conclusion

This study extends our previous work on modelling the effects of air pollution control regulations (Han et al., 2018), to investigate the effectiveness of existing and newly introduced air pollution control regulations in Beijing, China during 2008 – 2019, using a Bayesian deep-learning approach. Our approach can model the complex relationship between $PM_{2.5}$ pollution concentrations and other confounding factors that potentially affect $PM_{2.5}$ pollution concentrations, better address the confounding effects in policy evaluation, and predict the hypothetical $PM_{2.5}$ pollution concentrations in the absence of any regulatory interventions (MAE=20.3; MAPE=36.8%). Results of our novel Bayesian deep learning regulatory intervention analysis show that the $PM_{2.5}$ pollution concentrations in Beijing were reduced by 11% on average, due to the aggregate effects of all regulatory interventions implemented during the period of 2008 – 2019. Moreover, after the introduction of Action Plan for Clean Air in China and Beijing in late 2013, as compared to the hypothetical $PM_{2.5}$ concentration (without any regulatory interventions), the estimated $PM_{2.5}$ reduction increased dramatically from 15% in 2015 to 44% in 2018. In the future, more relevant data should be collected, and more advanced machine-learning methods can be used to improve the interpretability of our proposed model and provide more fine-grained estimation of the regulatory effects in China and elsewhere.

## Acknowledgement

This research is supported in part by the General Research Fund of the Research Grants Council of Hong Kong, under Grant No. 17620920. We gratefully acknowledge the $PM_{2.5}$ data provided by Beijing US Embassy, China; and AOD data provided by NASA, USA. We also gratefully acknowledge the valuable comments and suggestions of our reviewers, and participants who attended our first HKU-Cambridge AI-WiSe workshop, titled “AI for Social Good”, held in

the Department of Computer Science and Technology, Cambridge, the United Kingdom, in May 2019, during which an earlier version of this work was presented.

**Declarations of interest:** none

## References

Alaa, A.M., van der Schaar, M., 2017. Bayesian inference of individualized treatment effects using multi-task Gaussian processes, Advances in Neural Information Processing Systems, pp. 3424-3432.

Alaa, A.M., Weisz, M., Van Der Schaar, M., 2017. Deep counterfactual networks with propensity-dropout. arXiv preprint arXiv:1706.05966.

Apple Inc., 2020. Dark Sky API, https://darksky.net/dev

Atan, O., Jordon, J., van der Schaar, M., 2018. Deep-treat: Learning optimal personalized treatments from observational data using neural networks, Thirty-Second AAAI Conference on Artificial Intelligence.

Athey, S., 2017. Beyond prediction: Using big data for policy problems. Science 355, 483-485.

Athey, S., Imbens, G.W., 2017. The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives 31, 3-32.

Beijing Municipal Bureau of Statistics, 2020. Beijing Statistical Year Book [Webpage; in Chinese], http://nj.tjj.beijing.gov.cn/nj/main/2019-tjnj/zk/indexch.htm

Beijing Transport Institute, 2020. Annual Report for Transport Development in Beijing [PDF; in Chinese], http://www.bjtrc.org.cn/List/index/cid/7/p/1.html

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural networks. arXiv preprint arXiv:1505.05424.

Buuren, S.v., Groothuis-Oudshoorn, K., 2010. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 1-68.

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356, 183-186.

Chen, Y., Jin, G.Z., Kumar, N., Shi, G., 2013. The promise of Beijing: Evaluating the impact of the 2008 Olympic Games on air quality. Journal of Environmental Economics and Management 66, 424-443.

DieselNet, n.d. China: Cars and Light Trucks, https://dieselnet.com/standards/cn/ld.php

Ferraro, P.J., 2009. Counterfactual thinking and impact evaluation in environmental policy. New Directions for Evaluation 2009, 75-84.

Fortunato, M., Blundell, C., Vinyals, O., 2017. Bayesian recurrent neural networks. arXiv preprint arXiv:1704.02798.

Gal, Y., 2016. Uncertainty in deep learning. PhD thesis, University of Cambridge.

Han, Y., Lam, J.C.K., Li, V.O.K., 2018. A Bayesian LSTM Model to Evaluate the Effects of Air Pollution Control Regulations in China, 2018 IEEE International Conference on Big Data (Big Data). IEEE, pp. 4465-4468.

Hartford, J., Lewis, G., Leyton-Brown, K., Taddy, M., 2017. Deep IV: A flexible approach for counterfactual prediction, Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 1414-1423.

Henneman, L.R., Liu, C., Mulholland, J.A., Russell, A.G., 2017. Evaluating the effectiveness of air quality regulations: A review of accountability studies and frameworks. Journal of the Air & Waste Management Association 67, 144-172.

Imbens, G.W., Rubin, D.B., 2015. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision?, Advances in Neural Information Processing Systems, pp. 5574-5584.

Lam, J.C.K., Han, Y., Wang, S., Li, V.O.K., Pollitt, M., Warde, P., 2019. A comparative study of air pollution trends in historical London and contemporary Beijing, In Search of Good Energy Policy. Cambridge University Press.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436-444.

Li, T., Shen, H., Yuan, Q., Zhang, X., Zhang, L., 2017a. Estimating ground‐level PM2.5 by fusing satellite and station observations: A geo‐intelligent deep learning approach. Geophysical Research Letters 44, 11,985-911,993.

Li, V.O.K., Lam, J.C.K., Chen, Y., Gu, J., 2017b. Deep learning model to estimate air pollution using M-BP to fill in missing proxy urban data, GLOBECOM 2017-2017 IEEE Global Communications Conference. IEEE, pp. 1-6.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T., 2017c. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental Pollution 231, 997-1004.

Li, X., Qiao, Y., Zhu, J., Shi, L., Wang, Y., 2017d. The “APEC blue” endeavor: Causal effects of air pollution regulation on air quality in China. Journal of Cleaner Production 168, 1381-1388.

Liu, X.-H., Zhang, Y., Cheng, S.-H., Xing, J., Zhang, Q., Streets, D.G., Jang, C., Wang, W.-X., Hao, J.-M., 2010. Understanding of regional air pollution over China using CMAQ, part I performance evaluation and seasonal variation. Atmospheric Environment 44, 2415-2426.

Liu, Y., He, K., Li, S., Wang, Z., Christiani, D.C., Koutrakis, P., 2012. A statistical model to evaluate the effectiveness of PM2.5 emissions control during the Beijing 2008 Olympic Games. Environment International 44, 100-105.

Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R., Welling, M., 2017. Causal effect inference with deep latent-variable models, Advances in Neural Information Processing Systems, pp. 6446-6456.

Marcus, G., 2018. Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.

Ong, B.T., Sugiura, K., Zettsu, K., 2016. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5. Neural Computing and Applications 27, 1553-1566.

Osoba, O.A., Welser IV, W., 2017. An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation.

Pearl, J., 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:1801.04016.

Pham, T., Tran, T., Phung, D., Venkatesh, S., 2017. Predicting healthcare trajectories from medical records: A deep learning approach. Journal of Biomedical Informatics 69, 218-229.

Pope III, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association 56, 709-742.

Pui, D.Y., Chen, S.-C., Zuo, Z., 2014. PM2.5 in China: Measurements, sources, visibility and health effects, and mitigation. Particuology 13, 1-26.

Rubin, D.B., 2005. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association 100, 322-331.

Shalit, U., Johansson, F.D., Sontag, D., 2017. Estimating individual treatment effect: generalization bounds and algorithms, Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 3076-3085.

Shi, C., Blei, D., Veitch, V., 2019. Adapting neural networks for the estimation of treatment effects, Advances in Neural Information Processing Systems, pp. 2507-2517.

Stolberg, H.O., Norman, G., Trop, I., 2004. Randomized controlled trials. American Journal of Roentgenology 183, 1539-1544.

US Department of State, 2020. Beijing US Embassy Air Quality Data [CSV file], https://www.airnow.gov/international/us-embassies-and-consulates/#China$Beijing

US NASA, 2020. AERONET Data Download Tool [CSV file], https://aeronet.gsfc.nasa.gov/cgi-bin/webtool_aod_v3?stage=3&place_code=10&region=Asia&state=China&site=Beijing&submit=Get+Download+Form

US NOOA, 2020. Hourly/Sub-Hourly Observational Data Map, https://gis.ncdc.noaa.gov/maps/ncei/cdo/hourly

Wang, J.-F., Hu, M.-G., Xu, C.-D., Christakos, G., Zhao, Y., 2013. Estimation of citywide air pollution in Beijing. PloS One 8, e53400.

Wooldridge, J.M., 2000. Econometric analysis of cross section and panel data. MIT Press.

Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y., 2018. Deep distributed fusion network for air quality prediction, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 965-973.

Yoon, J., Jordon, J., van der Schaar, M., 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. International Conference on Learning Representations.

Zhang, H., Wang, S., Hao, J., Wang, X., Wang, S., Chai, F., Li, M., 2016. Air pollution and control action in Beijing. Journal of Cleaner Production 112, 1519-1527.

Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., Xu, X., Wang, J., He, H., Liu, W., 2019. Drivers of improved PM2. 5 air quality in China from 2013 to 2017. Proceedings of the National Academy of Sciences 116, 24463-24469.

Zheng, S., Yi, H., Li, H., 2015. The impacts of provincial energy and environmental policies on air pollution control in China. Renewable and Sustainable Energy Reviews 49, 386-394.

**Table 1.** Data source

| **Data** | **Resolution** | **Variable** | **Source** |
|---|---|---|---|
| Air quality | Hourly station-level (aggregated into daily means) | $PM_{2.5}$ concentrations (ug/m$^3$) | US Department of State (2020) |
| Meteorology | Hourly city-level (aggregated into daily means) | Temperature, relative humidity, air pressure, wind speed, wind bearing, and visibility | Apple Inc. (2020)[1] |

| | | | |
|---|---|---|---|
| AOD | All city-level observations per day (aggregated into daily means) | AOD at 1020 nm, AOD at 870 nm, AOD at 675 nm, AOD at 440 nm, and precipitable water (cm) | US NASA (2020) |
| Socio-economic | Yearly | Population density (population per km$^2$), percentage of GDP generated from the secondary sector, and the number of vehicles | Beijing Municipal Bureau of Statistics (2020), Beijing Transport Institute (2020) |
| **Notes**<br>1. The weather data application program interface (API) no longer accepts new signups (Apple Inc., 2020). The historical meteorology data in Beijing can also be downloaded from the US's National Climatic Data Center (US NOOA, 2020). | | | |

**Table 2.** Comparison of the performance between Bayesian deep-learning and other baseline air pollution regulatory intervention models based on the test set

| Model | MAE[1] | MAPE[1] |
|---|---:|---:|
| SVR | 22.1 (1.6) | 38.8% (3.4%) |
| RF | 22.4 (1.7) | 46.9% (4.9%) |
| Bayesian LSTM | **20.3 (0.6)** | **36.8% (1.8%)** |
| **Notes**<br>1. Standard deviation is shown in parenthesis. | | |

**Table 3.** Annual $PM_{2.5}$ reduction due to local and national air pollution control regulations implemented in Beijing, China

| Year | Observed $PM_{2.5}$ (µg/m$^3$) | Simulated $PM_{2.5}$ (µg/m$^3$) | $PM_{2.5}$ reduction (µg/m$^3$) | Relative reduction of $PM_{2.5}$ |
|---|---:|---:|---:|---:|
| 2008 | 92 | 96 | 4 | 4% |
| 2009 | 102 | 99 | -3 | -3% |
| 2010 | 104 | 100 | -4 | -4% |
| 2011 | 98 | 99 | 1 | 1% |
| 2012 | 91 | 98 | 7 | 7% |

| | | | | |
|---|---|---|---|---|
| 2013 | 101 | 101 | 0 | 0% |
| 2014 | 98 | 100 | 2 | 2% |
| 2015 | 82 | 97 | 15 | 15% |
| 2016 | 73 | 96 | 23 | 24% |
| 2017 | 59 | 94 | 35 | 37% |
| 2018 | 51 | 91 | 40 | 44% |
| 2019 | 58 | 92 | 34 | 37% |
| 2008 – 2019 | 86 | 97 | 11 | 11% |

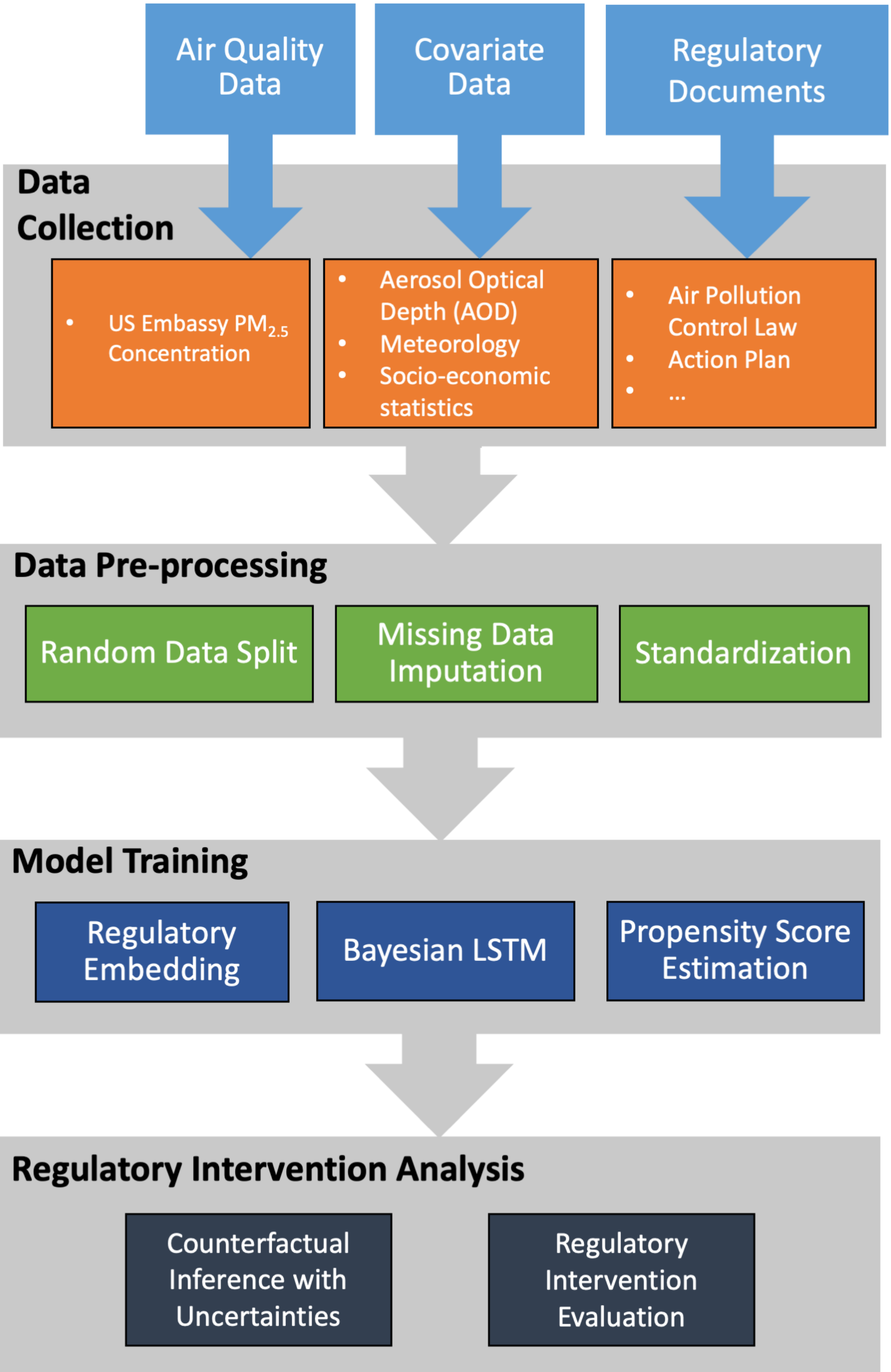


**Figure 1.** The overall framework of our proposed Bayesian deep-learning regulatory intervention analysis

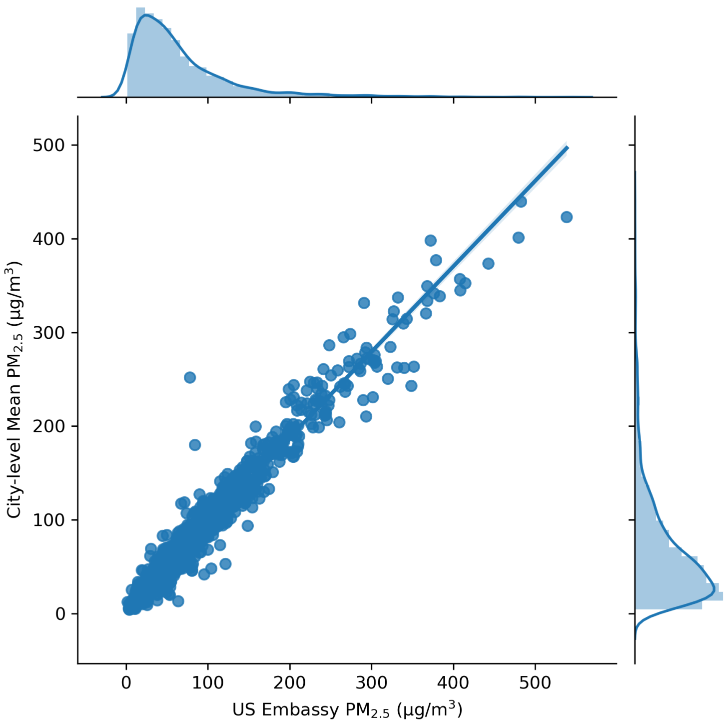


**Figure 2.** Correlation between the daily $PM_{2.5}$ concentrations monitored at the US Embassy, Beijing and the daily city-level average $PM_{2.5}$ concentrations monitored at the 35 official stations in Beijing, 2014 – 2019

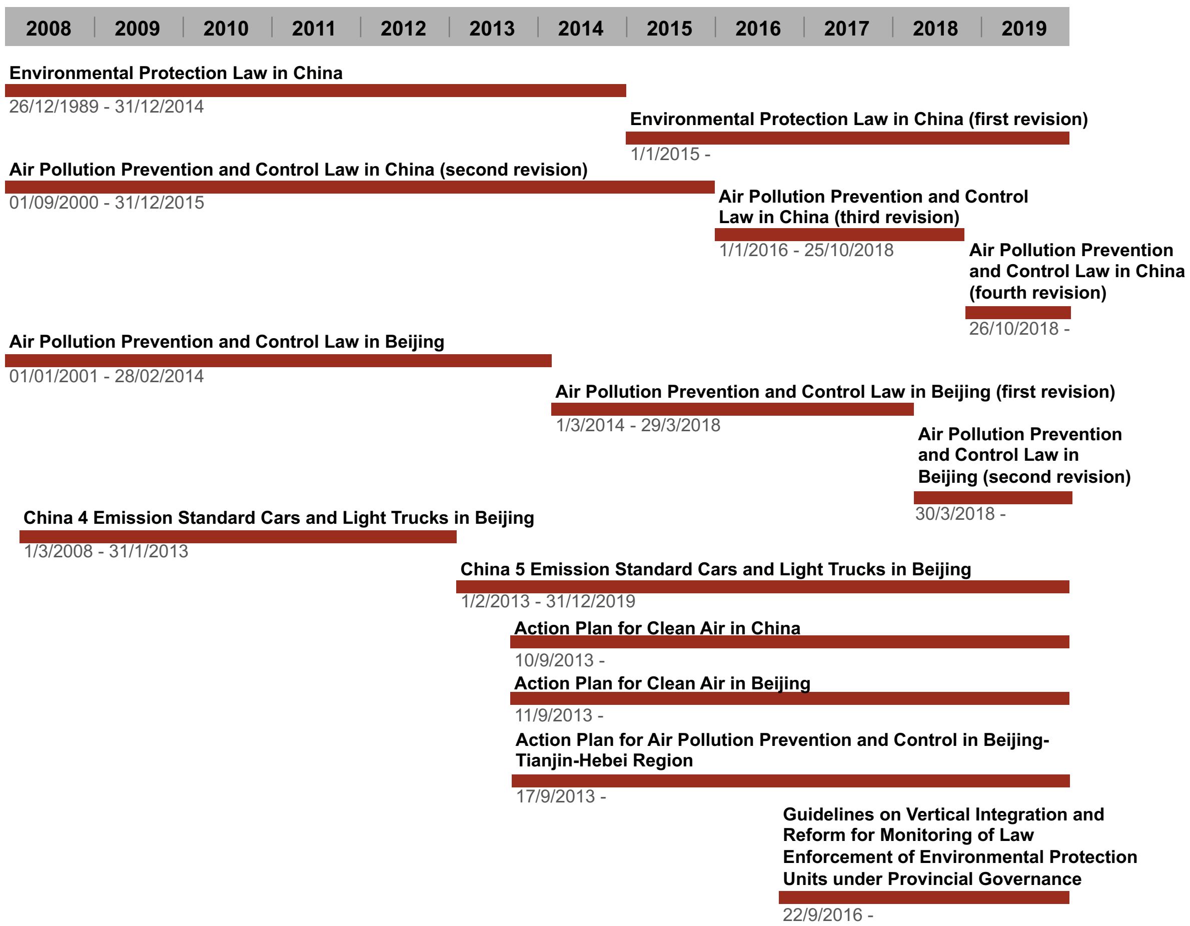


**Figure 3.** Timeline of air pollution control regulations implemented in Beijing/China during 2008 – 2019

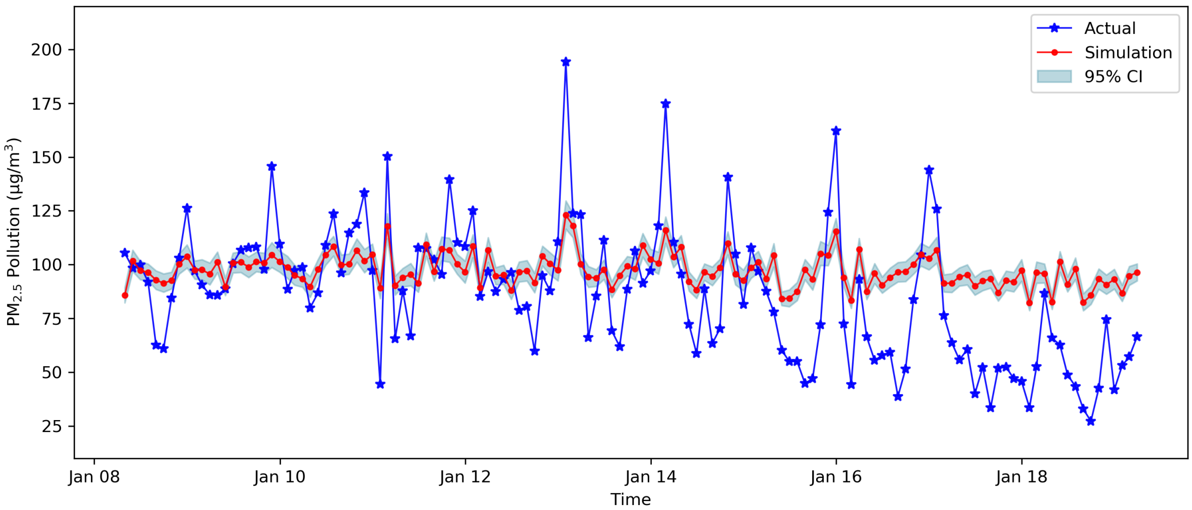


**Figure 4.** The monthly trend of observed $PM_{2.5}$ pollution concentrations (with regulatory interventions) and simulated $PM_{2.5}$ pollution concentrations (without any regulatory interventions) during 2008 – 2019