big data and cognitive computing

MDPI

Article

A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization

Dixian Zhu [1,*], Changjie Cai [2], Tianbao Yang [1] and Xun Zhou [3]

[1] Department of Computer Science, University of Iowa, Iowa City, IA 52242, USA; tianbao-yang@uiowa.edu
[2] Department of Occupational and Environmental Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA; changjie-cai@ouhsc.edu
[3] Department of Management Sciences, University of Iowa, Iowa City, IA 52242, USA; xun-zhou@uiowa.edu
* Correspondence: dixian-zhu@uiowa.edu

Received: 28 December 2017; Accepted: 19 February 2018; Published: 24 February 2018

Abstract: In this paper, we tackle air quality forecasting by using machine learning approaches to predict the hourly concentration of air pollutants (e.g., ozone, particle matter ($PM_{2.5}$) and sulfur dioxide). Machine learning, as one of the most popular techniques, is able to efficiently train a model on big data by using large-scale optimization algorithms. Although there exist some works applying machine learning to air quality prediction, most of the prior studies are restricted to several-year data and simply train standard regression models (linear or nonlinear) to predict the hourly air pollution concentration. In this work, we propose refined models to predict the hourly air pollution concentration on the basis of meteorological data of previous days by formulating the prediction over 24 h as a multi-task learning (MTL) problem. This enables us to select a good model with different regularization techniques. We propose a useful regularization by enforcing the prediction models of consecutive hours to be close to each other and compare it with several typical regularizations for MTL, including standard Frobenius norm regularization, nuclear norm regularization, and $\ell_{2,1}$-norm regularization. Our experiments have showed that the proposed parameter-reducing formulations and consecutive-hour-related regularizations achieve better performance than existing standard regression models and existing regularizations.

Keywords: air pollutant prediction; multi-task learning; regularization; analytical solution

1. Introduction

Adverse health impacts from exposure to outdoor air pollutants are complicated functions of pollutant compositions and concentrations [1]. Major outdoor air pollutants in cities include ozone (O_3), particle matter (PM), sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen oxides (NO_x), volatile organic compounds (VOCs), pesticides, and metals, among others [2,3]. Increased mortality and morbidity rates have been found in association with increased air pollutants (such as O_3, PM and SO_2) concentrations [3–5]. According to the report from the American Lung Association [6], a 10 parts per billion (ppb) increase in the O_3 mixing ratio might cause over 3700 premature deaths annually in the United States (U.S.). Chicago, as for many other megacities in U.S., has struggled with air pollution as a result of industrialization and urbanization. Although O_3 precursor (such as VOCs, NO_x, and CO) emissions have significantly decreased since the late 1970s, O_3 levels in Chicago have not been in compliance with standards set by the Environmental Protection Agency (EPA) to protect public health [7]. Particle size is critical in determining the particle deposition location in the human respiratory system [8]. $PM_{2.5}$, referring to particles with a diameter less than or equal to 2.5 μm, has been an increasing concern, as these particles can be deposited into the lung gas-exchange region, the alveoli [9]. The U.S. EPA revised the annual standard of $PM_{2.5}$ by lowering the concentration to 12 $\mu g/m^3$ to provide improved protection against health effects associated with long- and short-term

exposure [10]. SO_2, as an important precursor of new particle formation and particle growth, has also been found to be associated with respiratory diseases in many countries [11–15]. Therefore, we selected O_3, $PM_{2.5}$ and SO_2 for testing in this study.

Meteorological conditions, including regional and synoptic meteorology, are critical in determining the air pollutant concentrations [16–21]. According to the study by Holloway et al. [22], the O_3 concentration over Chicago was found to be most sensitive to air temperature, wind speed and direction, relative humidity, incoming solar radiation, and cloud cover. For example, a lower ambient temperature and incoming solar radiation slow down photochemical reactions and lead to less secondary air pollutants, such as O_3 [23]. Increasing wind speed could either increase or decrease the air pollutant concentrations. For instance, when the wind speed was low (weak dispersion/ventilation), the pollutants associated with traffic were found at the highest concentrations [24,25]. However, strong wind speeds might form dust storms by blowing up the particles on the ground [26]. High humidity is usually associated with high concentrations of certain air pollutants (such as PM, CO and SO_2) but with low concentrations of other air pollutants (such as NO_2 and O_3) because of various formation and removal mechanisms [25]. In addition, high humidity can be an indicator of precipitation events, which result in strong wet deposition leading to low concentrations of air pollutants [27]. Because various particle compositions and their interactions with light were found to be the most important factors in attenuating visibility [28,29], low visibility could be an indicator of high PM concentrations. Cloud can scatter and absorb solar radiation, which is significant for the formation of some air pollutants (e.g., O_3) [23,30]. Therefore, these important meteorological variables were selected to predict air pollutant concentrations in this study.

Statistical models have been applied for air pollution prediction on the basis of meteorological data [31–35]. However, existing studies on statistical modeling have mostly been restricted to simply utilizing standard classification or regression models, which have neglected the nature of the problem itself or ignored the correlation between sub-models in different time slots. On the other hand, machine learning approaches have been developing for over 60 years and have achieved tremendous success in a variety of areas [36–41]. There exist various new tools and techniques invented by the machine learning community, which allow for more refined modeling of a specific problem. In particular, model regularization is a fundamental technique for improving the generalization performance of a predictive model. Accordingly, many efficient optimization algorithms have been developed for solving various machine learning formulations with different regularizations.

In this study, we focus on refined modeling for predicting hourly air pollutant concentrations on the basis of historical metrological data and air pollution data. A striking difference between this work and the previous works is that we emphasize how to regularize the model in order to improve its generalization performance and how to learn a complex regularized model from big data with advanced optimization algorithms. We collected 10 years worth of meteorological and air pollution data from the Chicago area. The air pollutant data was from the EPA [42,43], and the meteorological data was from MesoWest [44]. From their databases, we fetched consecutive hourly measurements of various meteorological variables and pollutants reported by two air quality monitoring stations and two air pollutant monitoring sites in the Chicago area. Each record of hourly measurements included meteorological variables such as solar radiation, wind direction and speed, temperature, and atmospheric pressure; as well as air pollutants, including $PM_{2.5}$, O_3, and SO_2. We used two methods for model regularization: (i) explicitly controlling the number of parameters in the model; (ii) explicitly enforcing a certain structure in the model parameters. For controlling the number of parameters in the model, we compared three different model formulations, which can be considered in a unified multi-task learning (MTL) framework with a diagonal- or full-matrix model. For enforcing the model matrix into a certain structure, we have considered the relationship between prediction models of different hours and compared three different regularizations with standard Frobenius norm regularization. The experimental results show that the model with the intermediate size and the proposed regularization, which enforces the prediction models of two consecutive hours to be

close, achieved the best results and was far better than standard regression models. We have also developed efficient optimization algorithms for solving different formulations and demonstrated their effectiveness through experiments.

The rest of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we describe the data collection and preprocessing. In Section 4, we describe the proposed solutions, including formulations, regularizations and optimizations. In Section 5, we present the experimental studies and the results. In Section 6, we give conclusions and indicate future work.

2. Related Work

Many previous works have been proposed to apply machine learning algorithms to air quality predictions. Some researchers have aimed to predict targets into discretized levels. Kalapanidas et al. [32] elaborated effects on air pollution only from meteorological features such as temperature, wind, precipitation, solar radiation, and humidity and classified air pollution into different levels (low, med, high, and alarm) by using a lazy learning approach, the case-based reasoning (CBR) system. Athanasiadis et al. [45] employed the σ-fuzzy lattice neurocomputing classifier to predict and categorize O_3 concentrations into three levels (low, mid, and high) on the basis of meteorological features and other pollutants such as SO_2, NO, NO_2, and so on. Kurt and Oktay [33] modeled geographic connections into a neural network model and predicted daily concentration levels of SO_2, CO, and PM_{10} 3 days in advance. However, the process of converting regression tasks to classification tasks is problematic, as it ignores the magnitude of the numeric data and consequently is inaccurate.

Other researchers have worked on predicting concentrations of pollutants. Corani [46] worked on training neural network models to predict hourly O_3 and PM_{10} concentrations on the basis of data from the previous day. Mainly compared were the performances of feed-forward neural networks (FFNNs) and pruned neural networks (PNNs). Further efforts have been made on FFNNs: Fu et al. [47] applied a rolling mechanism and gray model to improve traditional FFNN models. Jiang et al. [48] explored multiple models (physical and chemical model, regression model, and multiple layer perceptron) on the air pollutant prediction task, and their results show that statistical models are competitive with the classical physical and chemical models. Ni, X. Y. et al. [49] compared multiple statistical models on the basis of $PM_{2.5}$ data around Beijing, and their results implied that linear regression models can in some cases be better than the other models.

MTL focuses on learning multiple tasks that have commonalities [50] that can improve the efficiency and accuracy of the models. It has achieved tremendous successes in many fields, such as natural language processing [37], image recognition [38], bioinformatics [39,40], marketing prediction [41], and so on. A variety of regularizations can be utilized to enhance the commonalities of the related tasks, including the $\ell_{2,1}$-norm [51], nuclear norm [52], spectral norm [53], Frobenius norm [54], and so on. However, most of the former machine learning works on air pollutant prediction did not consider the similarities between the models and only focused on improving the model performance for a single task, that is, improving prediction performance for each hour either separately or identically.

Therefore, we decided to use meteorological and pollutant data to perform predictions of hourly concentrations on the basis of linear models. In this work, we focused on three different prediction model formulations and used the MTL framework with different regularizations. To the best of our knowledge, this is the first work that has utilized MTL for the air pollutant prediction task. We exploited analytical approaches and optimization techniques to obtain the optimal solutions. The model's evaluation metric was the root-mean-squared error (RMSE).

3. Data Collection and Preprocessing

3.1. Data Collection

We collected air pollutant data from two air quality monitoring sites and meteorological data from two weather stations from 2006 to 2015 (summarized in Table 1). The air pollutant data in this

study included the concentrations of O_3, $PM_{2.5}$ and SO_2. We downloaded the air pollutant data from the U.S. EPA's Air Quality System (AQS) database (https://www.epa.gov/outdoor-air-quality-data), which has been widely used for model evaluation [42,43]. We selected the meteorological variables that would affect the air pollutant concentrations, including air temperature, relative humidity, wind speed and direction, wind gust, precipitation accumulation, visibility, dew point, wind cardinal direction, pressure, and weather conditions. We downloaded the meteorological data from MesoWest (http://mesowest.utah.edu/), a project within the Department of Meteorology at the University of Utah, which has been aggregating meteorological data since 2002 [44].

The locations of the two air quality monitoring sites and two weather stations are shown in Figure 1. The Alsip Village (AV) air quality monitoring site is also located in a suburban residential area, which is in southern Cook County, Illinois (AQS ID: 17-031-0001; latitude/longitude: 41.670992/−87.732457. The Lemont Village (LV) air quality monitoring site is located in a suburban residential area, which is in southwestern Cook County, Illinois (AQS ID: 17-031-1601; latitude/longitude: 41.66812/−87.99057. The weather station situated in Lansing Municipal Airport (LMA) is the closest meteorological site (MesoWest ID: KIGQ; latitude/longitude: 41.54125/−87.52822) to the AV air quality monitoring site. The weather station positioned at Lewis University (LU) is the closest meteorological site (MesoWest ID: KLOT; latitude/longitude: 41.60307/−88.10164) to the LV air quality monitoring site.

Table 1. Summary of measurement sites and observed variables.

Measurement Sites	Variables
Alsip Village (AV)	Ozone concentration and $PM_{2.5}$ concentration
Lemont Village (LV)	Ozone concentration and sulfur dioxide concentration
Lansing Municipal Airport (LMA)	Temperature, relative humidity, wind speed and direction, wind gust, precipitation accumulation, visibility, dew point, wind cardinal direction, pressure, and weather conditions
Lewis University (LU)	The same as for LMA site

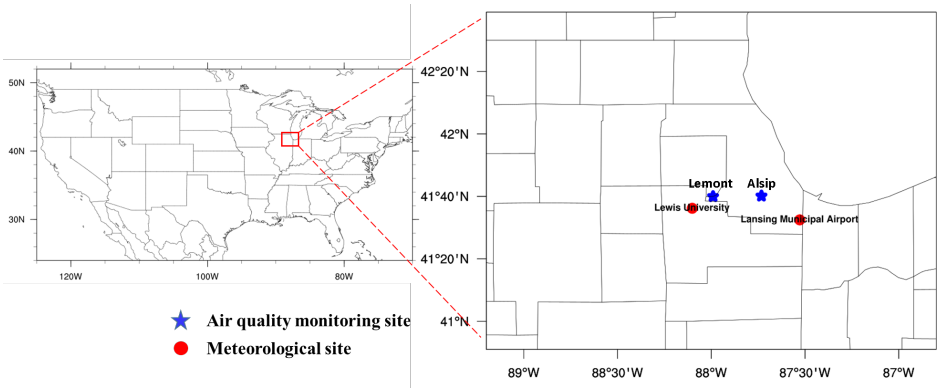


Figure 1. Locations of measurement sites. *Blue stars* denote the two air quality monitoring sites. *Red circles* denote the two meteorological sites.

3.2. Preprocessing

We paired the collected meteorological data and air pollutant data on the basis of time to obtain the required data format for applying the machine learning methods. In particular, for each variable, we formed one value for each hour. However, the original data may have contained multiple records or missing values at some hours. To preprocess the data, we calculated the hourly mean value of each numeric variable if there were multiple observed records within an hour and chose the category with the highest frequency per hour for each categorical variable if there were multiple values. Missing

values existed for some variables, which was not tolerable for applying the machine learning methods used in this study. Therefore, we imputed the missing values by using the closest-neighbor values for four continuous variables and one categorical variable: wind gust, pressure, altimeter reading, precipitation, and weather conditions. We deleted the days that still had missing values after imputing. We applied dummy coding for two categorical variables, the cardinal wind direction (16 values, e.g., N, S, E, W, etc.) and weather conditions (31 values, e.g., sunny, rainy, windy, etc.). Then, we added the weekday and weekend as two boolean features. Finally, we obtained 60 features in total (9 numerical meteorological features, 16 dummy codings for wind direction, 31 dummy codings for weather conditions, 2 boolean features for weekday/weekend, 1 numerical feature for pollutants, and 1 bias term). We applied normalization for all the features and pollutant targets to make their values fall in the range $[0, 1]$.

4. Machine Learning Approaches for Air Pollution Prediction

In this section, we describe the proposed approaches for predicting the ambient concentration of air pollutants.

4.1. A General Formulation

Our goal is to predict the concentration of air pollutants of the next day on the basis of the historical meteorological and air pollutant data. In this work, we have focused on using the former day's data to predict the next day's hourly pollutants. In particular, we let $(\mathbf{x}_i; y_i)$ denote the ith training data, where $y_i \in \mathbb{R}^{24\times 1}$ denotes the concentration of a certain air pollutant on a day, and $\mathbf{x}_i = (\mathbf{u}_i; \mathbf{v}_i)$ denotes the observed data on the previous day that include two components, where a semicolon ";" represents the column layout. The first component $\mathbf{u}_i = (\mathbf{u}_{i,1}; \ldots; \mathbf{u}_{i,D}) \in \mathbb{R}^{24\cdot D\times 1}$ includes all meteorological data over 24 h for the previous day, where $\mathbf{u}_{i,j} \in \mathbb{R}^{24\times 1}$ denotes the jth meteorological feature of the 24 h and D is the number of meteorological features; the second component $\mathbf{v}_i \in \mathbb{R}^{24\times 1}$ includes the hourly concentration of the same air pollutant on the previous day. The general formulation can be expressed as

$$\min_W \frac{1}{n}\sum_{i=1}^{n} \|f(W, \mathbf{x}_i) - y_i\|_2^2 + \varphi(W) \tag{1}$$

where W denotes the parameters of the model, $f(W, \mathbf{x}_i)$ denotes the prediction of the air pollutant concentration, and $\varphi(\cdot)$ denotes a regularization function of the model parameters W.

Next, we introduce two levels of model regularization. The first level is to explicitly control the number of model parameters. The second level is to explicitly impose a certain regularization on the model parameter. For the first level, we consider three models that are described below:

- **Baseline Model**. The first model is a baseline model that has been considered in existing studies and has the fewest number of parameters. In particular, the prediction of the air pollutant concentration is given by

$$f_k(W, \mathbf{x}_i) = \sum_{j=1}^{D} \mathbf{e}_k^\top \mathbf{u}_{i,j} \cdot w_j + \mathbf{e}_k^\top \mathbf{v}_i \cdot w_{D+1} + w_0, \quad k = 1, \ldots, 24$$

  where $\mathbf{e}_k \in \mathbb{R}^{24\times 1}$ is a basis vector with 1 at only the kth position and 0 at other positions; $w_0, w_1, \ldots, w_D, w_{D+1} \in \mathbb{R}$ are the model parameters, where w_0 is the bias term. We denote this model by $W = (w_0, w_1, \ldots, w_{D+1})^\top$. It is notable that this model predicts the hourly concentration on the basis of the same hourly historical data of the previous day and that it has $D + 2$ parameters. This simple model assumes that all 24 h share the same model parameter.

- **Heavy Model**. The second model takes all the data of the previous day into account when predicting the concentration of every hour of the second day. In particular, for the kth hour, the prediction is given by

$$f_k(W, \mathbf{x}_i) = \sum_{j=1}^{D} \mathbf{u}_{i,j}{}^{\top} \mathbf{w}_{k,j} + \mathbf{v}_i^{\top} \mathbf{w}_{k,D+1} + w_{k,0}, \quad k = 1, \ldots, 24$$

where $\mathbf{w}_{k,j} \in \mathbb{R}^{24\times 1}, j = 1, \ldots, D+1$ and $w_{k,0} \in \mathbb{R}$. This model is defined by

$$W = \begin{bmatrix} w_{1,0} & w_{2,0} & \cdots & w_{24,0} \\ \mathbf{w}_{1,1} & \mathbf{w}_{2,1} & \cdots & \mathbf{w}_{24,1} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{w}_{1,D+1} & \mathbf{w}_{2,D+1} & \cdots & \mathbf{w}_{24,D+1} \end{bmatrix}$$

We note that each column of W corresponds to the prediction model for each hour. There are a total of $24 \times (24 \times (D+1)+1)$ parameters. It is notable that the baseline model is a special case by enforcing all columns of W to be the same and because each $\mathbf{w}_{k,j}$ has only one non-zero element at the kth position.

- **Light Model**. The third model is between the baseline model and the heavy model. It considers the 24 h pattern of the air pollutants in the previous day and the same hourly meteorological data of the previous day to predict the concentration at a particular hour. The prediction is given by

$$f_k(W, \mathbf{x}_i) = \sum_{j=1}^{D} \mathbf{e}_k^{\top} \mathbf{u}_{i,j} \cdot w_{k,j} + \mathbf{v}_i^{\top} \mathbf{w}_{k,D+1} + w_{k,0}, \quad k = 1, \ldots, 24$$

where $w_{k,j} \in \mathbb{R}, j = 1, \ldots, D$ and $\mathbf{w}_{k,D+1} \in \mathbb{R}^{24\times 1}$. This model is defined by

$$W = \begin{bmatrix} w_{1,0} & w_{2,0} & \cdots & w_{24,0} \\ w_{1,1} & w_{2,1} & \cdots & w_{24,1} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{w}_{1,D+1} & \mathbf{w}_{2,D+1} & \cdots & \mathbf{w}_{24,D+1} \end{bmatrix}$$

It is also notable that each column corresponds to the predictive model for one hour and that W has a total of $24 \times (D+1) + 24 \times 24 \times 1$ parameters.

4.2. Regularization of Model Parameters

In this section, we describe different regularizations for the model parameter matrices W in the heavy and light models. We consider the problem using MTL, in which predicting the concentration of air pollutants over one hour is one task. In the literature, a number of regularizations have been proposed by considering the relationship between different tasks. We first describe three baseline regularizations in the literature and then present the proposed regularization that takes the dimension of time into consideration for modeling the relationship between models at different times.

- **Frobenius norm regularization**. Frobenius norm regularization is a generalization of standard Euclidean norm regularization to the matrix case, for which

$$\varphi(W) = \lambda ||W||_F^2$$

where $\lambda > 0$ is a regularization parameter.

- **$\ell_{2,1}$-norm regularization**. $\ell_{2,1}$-norm regularization has been used for feature selection in MTL. The norm is formed by first computing the ℓ_2-norm of each row of the W matrix (across different tasks) and then computing the ℓ_1-norm of the resulting vector. In particular, for $W \in \mathbb{R}^{d\times K}$,

$$\|W\|_{2,1} = \sum_{j=1}^{d} \|W_{j,*}\|_2$$

where $W_{j,*}$ denotes the jth row of W. We consider a $\ell_{2,1}$-norm regularizer $\varphi(W) = \lambda\|W\|_{2,1}$.

- **Nuclear norm regularization**. The nuclear norm is defined as the sum of singular values of a matrix, which is a standard regularization for enforcing a matrix to have a low rank. The motivation for using a low-rank matrix is that models for consecutive hours are highly correlated, which could render the matrix W to be low rank. We denote by $\|W\|_*$ the nuclear norm of a matrix W; the regularization is $\varphi(W) = \lambda\|W\|_*$.
- **Consecutive close (CC) regularization**. Finally, we propose a useful regularization for the considered problem that explicitly enforces the predictive models for two consecutive hours to be close to each other. The intuition is that usually the concentrations of air pollutants for two consecutive hours are close to each other. We denote the model by $W = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ and by $Cons(W) = [(\mathbf{w}_1 - \mathbf{w}_2), (\mathbf{w}_2 - \mathbf{w}_3), ..., (\mathbf{w}_{K-1} - \mathbf{w}_K)]$. The CC regularization is given by

$$\varphi(W) = \lambda \sum_{j=1}^{K-1} \|\mathbf{w}_j - \mathbf{w}_{j+1}\|_p^p \tag{2}$$

where $p = 1$ or $p = 2$.

4.3. Stochastic Optimization Algorithms for Different Formulations

With the exception that the Frobenius norm regularized model (with ℓ_2-norm CC regularization or not) has a closed-form solution, we solved the other models via advanced stochastic optimization techniques. We denote the following: $F(W, \mathbf{x}_i) = [f_1(W, \mathbf{x}_i), ..., f_{24}(W, \mathbf{x}_i)]$ and $Y_i = [y_{i,1}, ..., y_{i,24}]$; the total number of features is D. Although the standard stochastic (sub)gradient method [55] could be utilized to solve all the formulations considered in this work, it does not necessary yield the fastest convergence. To address this issue, we considered advanced stochastic optimization techniques tailored for solving each formulation.

4.3.1. Optimizing $\ell_{2,1}$-Norm Regularized Model

We utilized the accelerated stochastic subgradient (ASSG) method [56] with proximal mapping to optimize this model. The algorithm runs in mutliple stages, and each stage calls the standard stochastic gradient method with a constant step size. To handle the non-smooth $\ell_{2,1}$-norm, we used proximal mapping [57]. The stochastic gradient descent part is

$$W'_t = W_{t-1} - 2\eta_s \frac{\partial F(W_{t-1}, \mathbf{x}_i)}{\partial W_{t-1}} \mathbf{e}^\top (F(W_{t-1}, \mathbf{x}_i) - Y_i) \tag{3}$$

where η_s is the stage-wise step size, i is a sampled index, and $\mathbf{e}$ is a vector with 1 for all its elements. Then a proximal mapping is as follows (denoted by $\tilde{\lambda} = 2\eta_s\lambda$):

$$W_t = \arg\min_W \|W - {W'}_t\|_F^2 + \tilde{\lambda}\|W\|_{2,1} \tag{4}$$

The above problem has analytical solutions. We denote $\mathbf{w}_i$ as a column vector for $W^\top$ and $\mathbf{w}'_i$ as a column vector for ${W'}_t^\top$. Then the solution to Equation (4) can be computed by the following [51]:

$$\mathbf{w}_i = \begin{cases} (1 - \frac{\tilde{\lambda}}{\|\mathbf{w}'_i\|_2})\mathbf{w}'_i, & \tilde{\lambda} > 0, \|\mathbf{w}'_i\|_2 > \tilde{\lambda} \\ \mathbf{0}, & \tilde{\lambda} > 0, \|\mathbf{w}'_i\|_2 \leq \tilde{\lambda} \\ \mathbf{w}'_i, & \tilde{\lambda} = 0 \end{cases} \tag{5}$$

The pseudocode of the algorithm is as follows:

Algorithm 1: ASSG method with proximal mapping solving $\ell_{2,1}$-norm regularized model.

Input: X, Y, W_0, η_0, S, and T
for $s = 1, \ldots, S$ **do**
  $\eta_s = \eta_{s-1}/2$
  for $t = 1, \ldots, T$ **do**
    sample $i \in \{1, ..., n\}$
    update W'_t using Equation (3)
    update W_t using Equation (4)
  end
  $W_0 = \sum_{t=1}^{T} W_1/W_T$
end
Output: W_0

4.3.2. Optimizing Nuclear Norm Regularized Model

The challenge in solving the nuclear norm reguralized problem of most optimization algorithms lies with computing the full singular value decomposition (SVD) of the involved matrix W, which is an expensive operation. To avoid full SVD, the SVD-free convex–concave algorithm extension to a stochastic setting (SECONE-S) [58] was employed to solve the problem. The algorithm solves the following minimum–maximum problem:

$$\min_{W \in \mathbb{R}^{D \times K}} \max_{U \in \mathbb{R}^{D \times K}} \frac{1}{n} \sum_{i=1}^{n} \|F(W, \mathbf{x}_i) - Y_i\|_2^2 + \lambda \mathrm{tr}(U^\top W) - \rho[\|U\|_2 - 1]_+$$

Then stochastic gradient descent and ascent are used to update W and U at each iteration:

$$\begin{aligned} W_t &= W_{t-1} - \eta_{t-1}(2\frac{\partial F(W_{t-1}, \mathbf{x}_i)}{\partial W_{t-1}} \mathbf{e}^\top (F(W_{t-1}, \mathbf{x}_i) - Y_i) + \lambda U_{t-1}) \\ U_t &= U_{t-1} + \tau_{t-1}(\lambda W_{t-1} - \rho \partial[\|U_{t-1}\|_2 - 1]_+) \end{aligned} \tag{6}$$

where $\rho \geq \|Y\|_F^2$ and $\partial[\|U_t\|_2 - 1]_+$ can be computed by $\mathbf{u}_1 \mathbf{v}_1^\top \mathbf{1}[\sigma_1 > 1]$, with $(\mathbf{u}_1, \mathbf{v}_1)$ being the top-left and -right singular vectors of U_t and σ_1 being the top singular value. The pseudocode for the algorithm is as follows:

Algorithm 2: SECONE-S solving nuclear norm regularized model.

Input: X, Y, T, η_0, and τ_0
for $t = 1, \ldots, T$ **do**
  sample $i \in \{1, ..., n\}$
  update W_t and U_t using Equation (6)
  $\eta_t = \eta_0/\sqrt{t}$, and $\tau_t = \tau_0/\sqrt{t}$
end
Output: $\hat{W}_T = \sum_{t=1}^{T} W_t/T$

4.3.3. Optimizing Consecutive Close Regularized Model

The challenge of tackling the proposed CC regularization lies in that the standard proximal mapping cannot be computed efficiently. We addressed this challenge by using the alternating-direction method of multipliers. We utilized a recently proposed locally adaptive stochastic alternating-direction method of multipliers (LA-SADMM) [59] to solve the CC regularized model. Below, we discuss the updates for the choice of $p = 1$ (i.e., using the ℓ_1-norm) in Equation (2). The updates for the choice of $p = 2$ can be derived similarly.

The objective function can be written as

$$\min_{W\in\mathbb{R}^{D\times K}} \frac{1}{n}\sum_{i=1}^{n}\|F(W,\mathbf{x}_i)-Y_i\|_2^2+\lambda\|WE\|_{1,1}$$

Here, $E=(\hat{\mathbf{e}}_1,...,\hat{\mathbf{e}}_{k-1})$, where $\hat{\mathbf{e}}_i=(0,...,1,-1,...,0)^T$, $i=1,...,k-1$, the ith element is 1 and the $(i+1)$th element is -1. Therefore, $Cons(W)=WE$. A dummy variable $U=WE$ was introduced to decouple the last term from the first term, and a Lagrangian function was formed as follows:

$$L(W,U,\Lambda)=\frac{1}{n}\sum_{i=1}^{n}\|F(W,\mathbf{x}_i)-Y_i\|_2^2+\lambda\|U\|_{1,1}-\mathrm{tr}(\Lambda^\top(WE-U))+\frac{\beta}{2}\|WE-U\|_F^2 \tag{7}$$

where Λ is the Lagrangian multiplier and β is the penalty parameter.

This could then be solved by optimizing each variable alternatively. The update rules for SADMM are as follows:

$$\begin{aligned}
W_\tau &= \arg\min_{W\in\mathbb{R}^{D\times K}} L(W,U_{\tau-1},\Lambda_{\tau-1}) = \arg\min_{W\in\mathbb{R}^{D\times K}} \tilde{F}(W_{\tau-1},\mathbf{x}_i)+\mathrm{tr}\{\frac{\partial\tilde{F}(W_{\tau-1},\mathbf{x}_i)}{\partial W}^\top(W-W_{\tau-1})\}\\
&\quad+\frac{\beta}{2}\|WE-U_{\tau-1}-\frac{1}{\beta}\Lambda_{\tau-1}^T\|_F^2+\frac{\|W-W_{\tau-1}\|_F^2}{\eta_{\tau-1}}\\
U_\tau &= \arg\min_{U\in\mathbb{R}^{D\times K}} L(W_\tau,U,\Lambda_{\tau-1}) = \arg\min_{U\in\mathbb{R}^{D\times K}} \gamma\|U\|_{1,1}+\frac{\beta}{2}\|W_\tau E-U-\frac{1}{\beta}\Lambda_{\tau-1}^T\|_F^2\\
\Lambda_\tau &= \Lambda_{\tau-1}-\beta(W_\tau E-U_\tau)^T
\end{aligned} \tag{8}$$

where $\tilde{F}(W_{\tau-1},\mathbf{x}_i)=\|F(W_{\tau-1},\mathbf{x}_i)-Y_i\|_2^2$.

LA-SADMM solves the problem more efficiently by doing stage-wise penalty increasing. The pseudocode for the algorithm is as follows:

Algorithm 3: LA-SADMM solving consecutive close (CC) regularized problem with ℓ_1-norm.

Input: X, Y, W_0, U_0, Λ_0, β_1, η_1, S, and T
for $s=1,\ldots,S$ **do**
  for $\tau=1,\ldots,T$ **do**
    sample $i\in\{1,...,n\}$
    update W_τ, U_τ, and Λ_τ using Equation (8)
  end
  $W_T=\sum_{\tau=1}^{T}W_\tau/T$
  $W_0=W_T$, $U_0=U_T$, and $\Lambda_0=\Lambda_T$
  $\beta_{s+1}=2\beta_s$, and $\eta_{s+1}=\eta_s/2$
end
Output: W_T

4.4. Extensive Discussion

It is noteworthy that the main contribution of this work is the incorporation of model parameter reduction and MTL with regularization into air pollutant prediction. As the previous content has illustrated, for the parameter reduction part, our light formulation reduces model parameters by removing heavy meteorological parameters of the other hours for one hour's submodel. For the MTL part, we considered that there could be some similarities for consecutive hours' models; therefore, we could add appropriate regularizers for this purpose.

The high-level idea of MTL lies in transfer learning, which generally aims to transfer knowledge from a related source task to a target task and consequently improve the performance for the target task. There are multiple variants for transfer learning, such as inductive transfer learning, transductive transfer learning and unsupervised transfer learning, and the approaches for transfer learning mainly

include instance transfer, feature-representation transfer, parameter transfer and relational-knowledge transfer [60]. One of the most common examples is feature-representation transfer for deep neural networks. After either supervised or unsupervised learning from other related datasets, the pretrained model can be appropriately reused for learning the target task with a better performance. The MTL technique in this work is an example of parameter transfer in an inductive-transfer-learning setting.

A similar idea can be applied to other kinds of work. First, if the submodels are not built for each hour but for each day (or even for each location from a spatial perspective), we can still apply the parameter reduction idea that only keeps more important information and removes the information with low priority. Second, for the MTL part, we can still add regularizations for the similarities of the submodels. Furthermore, in this work, the submodel $\mathbf{w}_i$ was a linear regression model; it is also practical to replace it with support vector regression (SVR), nonlinear regression, neural networks, and so on. Finally, the techniques used in this work can be further combined with many other transfer learning techniques, such as feature-representation transfer for deep neural networks.

5. Experiments

We used the names of the paired air quality monitoring sites and two weather stations to denote the two datasets, that is, LU–LV and LMA–AV. LU–LV contained the data to predict the concentration of the two air pollutants O_3 and SO_2. LMA–AV contained the data to predict the concentration of the two air pollutants O_3 and $PM_{2.5}$.

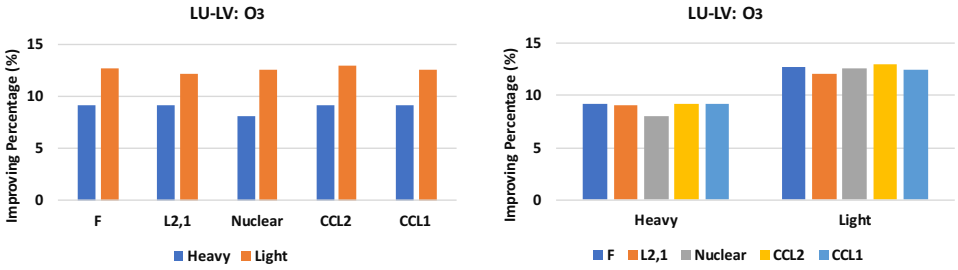
We compared 11 different models that were learned with different combinations of model formulations and regularizations. The 11 models were the following:

- Baseline: the baseline model with standard Frobenius norm regularization.
- Heavy–F: the heavy model with standard Frobenius norm regularization.
- Light–F: the heavy model with standard Frobenius norm regularization.
- Heavy–$\ell_{2,1}$: the heavy model with $\ell_{2,1}$-norm regularization.
- Heavy–nuclear: the heavy model with nuclear-norm regularization.
- Heavy–CCL2: the heavy model with CC regularization using the ℓ_2-norm.
- Heavy–CCL1: the heavy model with CC regularization using the ℓ_1-norm.
- Light–$\ell_{2,1}$: the light model with $\ell_{2,1}$-norm regularization.
- Light–nuclear: the light model with nuclear-norm regularization.
- Light–CCL2: the light model with CC regularization using the ℓ_2-norm.
- Light–CCL1: the light model with CC regularization using the ℓ_1-norm.

It is noteworthy that we also added the standard Frobenius norm regularizer for the heavy/light–nuclear, –CCL2, and –CCL1 models, because their regularizers were mainly considered for controlling the similarities of submodels and may not have been enough for preventing overfitting. We divided each dataset into two parts: training data and testing data. Each model was trained on the training data with proper regularization parameters and the learning rate selected on the basis of 5-fold cross-validation. Each trained model was evaluated on the testing data. The splitting of the data was done by dividing all days into a number of chunks of 11 consecutive days, for which the first 8 days were used for training and the next 3 days were used for testing. We have used the RMSE as the evaluation metric.

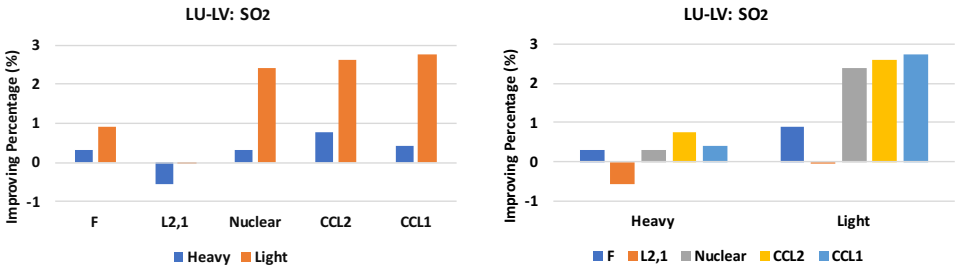
We first report the improvement of each method over the baseline method. The improvement was measured by a positive or negative percentage over the performance of the baseline method, that is, (RMSE of compared method - RMSE of the baseline method)$\times$100/RMSE of the baseline method. The results are shown in Figures 2 and 3. To facilitate the comparison between different methods, for each air pollutant of each dataset, we report two figures, with one grouping the results by regularizations and the other grouping the results by the model formulations. From the results, we can see that (i) the light model formulation had a clear advantage over the heavy model formulation and the baseline model formulation, which implied that controlling the number of parameters is important

for improving generalization performance; and (ii) the proposed CC regularization yielded a better performance than other regularizations, which verified that considering the similarities between models of consecutive hours is helpful. We also report the exact RMSE of each method in Table 2.



(**a**) Grouped by regularizations.

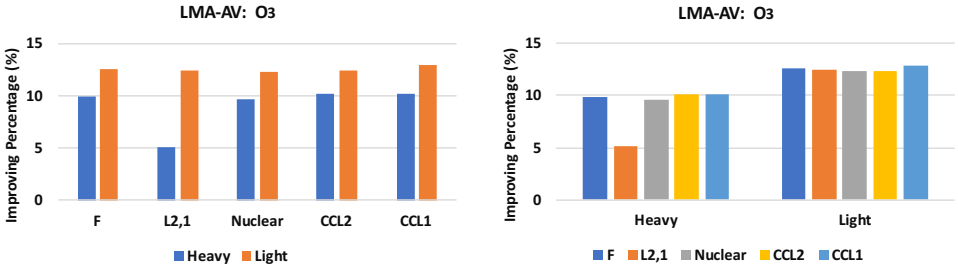
(**b**) Grouped by model formulations.



(**c**) Grouped by regularizations.

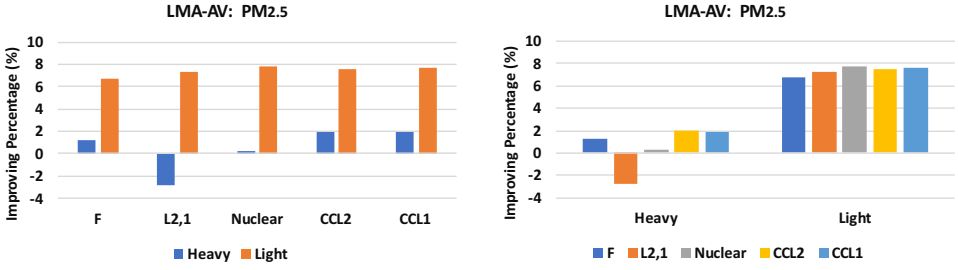
(**d**) Grouped by model formulations.

Figure 2. Improvement of different methods over the baseline method for Lewis University–Lemont Village (LU–LV) dataset.



(**a**) Grouped by regularizations.

(**b**) Grouped by model formulations.



(**c**) Grouped by regularizations.

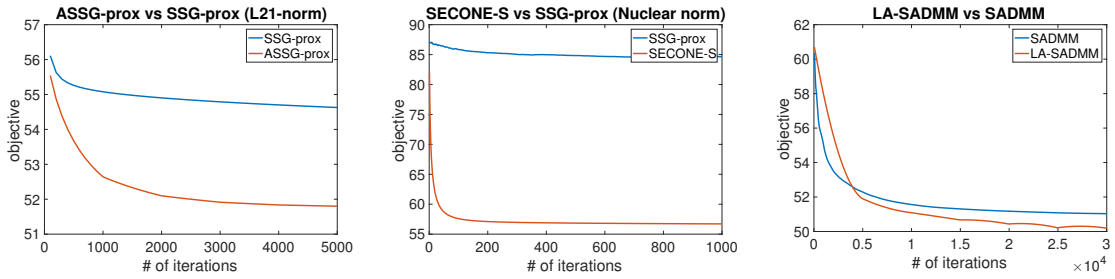
(**d**) Grouped by model formulations.

Figure 3. Improvement of different methods over the baseline method for Lansing Municipal Airport–Alsip Village (LMA–AV) dataset.

Table 2. Root-mean-squared error (RMSE) for all approaches and datasets. The best approaches are marked as bold.

Approaches	LMA-AV: O_3	LMA-AV: $PM_{2.5}$	LU-LV: O_3	LU-LV: SO_2
Baseline	0.1324	0.0399	0.0971	0.0334
Heavy–F	0.1193	0.0394	0.0882	0.0333
Heavy–$\ell_{2,1}$	0.12569	0.041	0.0883	0.033591
Heavy–nuclear	0.1197	0.0398	0.0893	0.0333
Heavy–CCL2	0.11896	0.0391	0.0882	0.033148
Heavy–CCL1	0.11897	0.039134	0.0882	0.033261
Light–F	0.1158	0.0372	0.0848	0.0331
Light–$\ell_{2,1}$	0.11591	0.037	0.085376	0.033411
Light–nuclear	0.1161	**0.0368**	0.0849	0.0326
Light–CCL2	0.116	0.0369	**0.0845**	0.03253
Light–CCL1	**0.11535**	0.03684	0.085	**0.03248**

Finally, we compared the convergence speed of the employed optimization algorithms with their standard counterparts. In particular, we compared the ASSG and SSG methods for optimizing the $\ell_{2,1}$-norm regularized problem, and SSG for solving the nuclear norm regularized problem, and and SADMM for solving the CC regularized problem. The results are plotted in Figure 4 and demonstrate that the employed advanced optimization techniques converged much faster than the classical techniques.



(**a**) Light–$\ell_{2,1}$. (**b**) Light–nuclear. (**c**) Light–CCL1.

Figure 4. Optimization techniques.

6. Conclusions

In this paper, we have developed efficient machine learning methods for air pollutant prediction. We have formulated the problem as regularized MTL and employed advanced optimization algorithms for solving different formulations. We have focused on alleviating model complexity by reducing the number of model parameters and on improving the performance by using a structured regularizer. Our results show that the proposed light formulation achieves much better performance than the other two model formulations and that the regularization by enforcing prediction models for two consecutive hours to be close can also boost the performance of predictions. We have also shown that advanced optimization techniques are important for improving the convergence of optimization and that they speed up the training process for big data. For future work, we will further consider the commonalities between nearby meteorology stations and combine them in a MTL framework, which may provide a further boosting for the prediction.

Acknowledgments: Authors would like to thank the support from Environmental Health Sciences Research Center at University of Iowa, and National Science Foundation Grant No. IIS-1566386 for funding and facilitating this research.

Author Contributions: Dixian Zhu, Tianbao Yang, and Xun Zhou conceived and designed the experiments; Changjie Cai collected the data; Dixian Zhu and Changjie Cai analyzed the data; Dixian Zhu performed the experiments; Xun Zhou and Tianbao Yang contributed to the progress of research idea; Tianbao Yang, Changjie Cai and Dixian Zhu wrote the paper. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Curtis, L.; Rea, W.; Smith-Willis, P.; Fenyves, E.; Pan, Y. Adverse health effects of outdoor air pollutants. *Environ. Int.* **2006**, *32*, 815–830.
2. Mayer, H. Air pollution in cities. *Atmos. Environ.* **1999**, *33*, 4029–4037.
3. Samet, J.M.; Zeger, S.L.; Dominici, F.; Curriero, F.; Coursac, I.; Dockery, D.W.; Schwartz, J.; Zanobetti, A. The national morbidity, mortality, and air pollution study. Part II: Morbidity and mortality from air pollution in the United States. *Res. Rep. Health Eff. Inst.* **2000**, *94*, 5–79.
4. Dockery, D.W.; Schwartz, J.; Spengler, J.D. Air pollution and daily mortality: Associations with particulates and acid aerosols. *Environ. Res.* **1992**, *59*, 362–373.
5. Schwartz, J.; Dockery, D.W. Increased mortality in Philadelphia associated with daily air pollution concentrations. *Am. Rev. Respir. Dis.* **1992**, *145*, 600–604.
6. American Lung Association. *State of the Air Report*; ALA: New York, NY, USA, 2007; pp. 19–27.
7. Environmental Protection Agency (EPA). Region 5: State Designations, as of September 18, 2009. Available online: https://archive.epa.gov/ozonedesignations/web/html/region5desig.html (accessed on 17 December 2017).
8. Hinds, W.C. *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
9. Soukup, J.M.; Becker, S. Human alveolar macrophage responses to air pollution particulates are associated with insoluble components of coarse material, including particulate endotoxin. *Toxicol. Appl. Pharmacol.* **2001**, *171*, 20–26.
10. Environmental Protection Agency (EPA). CFR Parts 50, 51, 52, 53, and 58-National Ambient Air Quality Standards for Particulate Matter: Final Rule. *Fed. Regist.* **2013**, *78*, 3086–3286.
11. Schwartz, J. Short term fluctuations in air pollution and hospital admissions of the elderly for respiratory disease. *Thorax* **1995**, *50*, 531–538.
12. De Leon, A.P.; Anderson, H.R.; Bland, J.M.; Strachan, D.P.; Bower, J. Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92. *J. Epidemiol. Community Health* **1996**, *50* (Suppl. 1), s63–s70.
13. Birmili, W.; Wiedensohler, A. New particle formation in the continental boundary layer: Meteorological and gas phase parameter influence. *Geophys. Res. Lett.* **2000**, *27*, 3325–3328.
14. Lee, J.-T.; Kim, H.; Song, H.; Hong, Y.C.; Cho, Y.S.; Shin, S.Y.; Hyun, Y.J.; Kim, Y.S. Air pollution and asthma among children in Seoul, Korea. *Epidemiology* **2002**, *13*, 481–484.
15. Cai, C.; Zhang, X.; Wang, K.; Zhang, Y.; Wang, L.; Zhang, Q.; Duan, F.; He, K.; Yu, S.-C. Incorporation of new particle formation and early growth treatments into WRF/Chem: Model improvement, evaluation, and impacts of anthropogenic aerosols over East Asia. *Atmos. Environ.* **2016**, *124*, 262–284.
16. Kalkstein, L.S.; Corrigan, P. A synoptic climatological approach for geographical analysis: Assessment of sulfur dioxide concentrations. *Ann. Assoc. Am. Geogr.* **1986**, *76*, 381–395.
17. Comrie, A.C. A synoptic climatology of rural ozone pollution at three forest sites in Pennsylvania. *Atmos. Environ.* **1994**, *28*, 1601–1614.
18. Eder, B.K.; Davis, J.M.; Bloomfield, P. An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. *J. Appl. Meteorol.* **1994**, *33*, 1182–1199.
19. Zelenka, M.P. An analysis of the meteorological parameters affecting ambient concentrations of acid aerosols in Uniontown, Pennsylvania. *Atmos. Environ.* **1997**, *31*, 869–878.
20. Laakso, L.; Hussein, T.; Aarnio, P.; Komppula, M.; Hiltunen, V.; Viisanen, Y.; Kulmala, M. Diurnal and annual characteristics of particle mass and number concentrations in urban, rural and Arctic environments in Finland. *Atmos. Environ.* **2003**, *37*, 2629–2641.
21. Jacob, D.J.; Winner, D.A. Effect of climate change on air quality. *Atmos. Environ.* **2009**, *43*, 51–63.

22. Holloway, T.; Spak, S.N.; Barker, D.; Bretl, M.; Moberg, C.; Hayhoe, K.; Van Dorn, J.; Wuebbles, D. Change in ozone air pollution over Chicago associated with global climate change. *J. Geophys. Res. Atmos.* **2008**, *113*, doi:10.1029/2007JD009775.
23. Akbari, H. Shade trees reduce building energy use and CO_2 emissions from power plants. *Environ. Pollut.* **2002**, *116*, S119–S126.
24. DeGaetano, A.T.; Doherty, O.M. Temporal, spatial and meteorological variations in hourly PM 2.5 concentration extremes in New York City. *Atmos. Environ.* **2004**, *38*, 1547–1558.
25. Elminir, H.K. Dependence of urban air pollutants on meteorology. *Sci. Total Environ.* **2005**, *350*, 225–237.
26. Natsagdorj, L.; Jugder, D.; Chung, Y.S. Analysis of dust storms observed in Mongolia during 1937–1999. *Atmos. Environ.* **2003**, *37*, 1401–1411.
27. Seinfeld, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
28. Appel, B.R.; Tokiwa, Y.; Hsu, J.; Kothny, E.L.; Hahn, E. Visibility as related to atmospheric aerosol constituents. *Atmos. Environ. (1967)* **1985**, *19*, 1525–1534.
29. Deng, X.; Tie, X.; Wu, D.; Zhou, X.; Bi, X.; Tan, H.; Li, F.; Jiang, C. Long-term trend of visibility and its characterizations in the Pearl River Delta (PRD) region, China. *Atmos. Environ.* **2008**, *42*, 1424–1435.
30. Twomey, S. The influence of pollution on the shortwave albedo of clouds. *J. Atmos. Sci.* **1977**, *34*, 1149–1152.
31. Zheng, Y.; Liu, F.; Hsieh, H.-P. U-Air: When urban air quality inference meets big data. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013.
32. Kalapanidas, E.; Avouris, N. Short-term air quality prediction using a case-based classifier. *Environ. Model. Softw.* **2001**, *16*, 263–272.
33. Kurt, A.; Oktay, A.B. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Syst. Appl.***2010**, *37*, 7986–7992.
34. Kleine Deters, J.; Zalakeviciute, R.; Gonzalez, M.; Rybarczyk, Y. Modeling PM2.5 urban pollution using machine learning and selected meteorological parameters. *J. Electr. Comput. Eng.* **2017**, *2017*, 5106045.
35. Bougoudis, I.; Demertzis, K.; Iliadis, L.; Anezakis, V.-D.; Papaleonidas, A. FuSSFFra, a fuzzy semi-supervised forecasting framework: The case of the air pollution in Athens. In *Neural Computing and Applications*; Springer: Berlin, Germany, 2017; pp. 1–14.
36. Yuan, Z.; Zhou, X.; Yang, T.; Tamerius, J.; Mantilla, R. Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study. In Proceedings of the 6th International Workshop on Urban Computing (UrbComp 2017), Halifax, NS, Canada, 14 August 2017.
37. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008.
38. Fan, J.; Gao, Y.; Luo, H. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE Trans. Image Process.* **2008**, *17*, 407–426.
39. Widmer, C.; Leiva, J.; Altun, Y.; Rätsch, G. Leveraging sequence classification by taxonomy-based multitask learning. In *Annual International Conference on Research in Computational Molecular Biology*; Springer: Berlin/Heidelberg, Germany, 2010.
40. Kshirsagar, M.; Carbonell, J.; Klein-Seetharaman, J. Multitask learning for host-pathogen protein interactions. *Bioinformatics* **2013**, *29*, i217–i226.
41. Lindbeck, A.; Snower, D.J. Multitask learning and the reorganization of work: From tayloristic to holistic organization. *J. Labor Econ.* **2000**, *18*, 353–376.
42. Foley, K.M.; Roselle, S.J.; Appel, K.W.; Bhave, P.V.; Pleim, J.E.; Otte, T.L.; Mathur, R.; Sarwar, G.; Young, J.O.; Gilliam, R.C.; et al. Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7. *Geosci. Model Dev.* **2010**, *3*, 205–226.
43. Yahya, K.; Wang, K.; Campbell, P.; Chen, Y.; Glotfelty, T.; He, J.; Pirhalla, M.; Zhang, Y. Decadal application of WRF/Chem for regional air quality and climate modeling over the US under the representative concentration pathways scenarios. Part 1: Model evaluation and impact of downscaling. *Atmos. Environ.* **2017**, *152*, 562–583.
44. Horel, J.; Splitt, M.; Dunn, L.; Pechmann, J.; White, B.; Ciliberti, C.; Lazarus, S.; Slemmer, J.; Zaff, D.; Burks, J.; et al. Mesowest: Cooperative mesonets in the western United States. *Bull. Am. Meteorol. Soc.* **2002**, *83*, 211–225.

45. Athanasiadis, I.N.; Kaburlasos, V.G.; Mitkas, P.A.; Petridis, V. Applying machine learning techniques on air quality data for real-time decision support. In Proceedings of the First international NAISO Symposium on Information Technologies in Environmental Engineering (ITEE'2003), Gdansk, Poland, 24–27 June 2003.
46. Corani, G. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. *Ecol. Model.* **2005**, *185*, 513–529.
47. Fu, M.; Wang, W.; Le, Z.; Khorram, M.S. Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. *Neural Comput. Appl.* **2015**, *26*, 1789–1797.
48. Jiang, D.; Zhang, Y.; Hu, X.; Zeng, Y.; Tan, J.; Shao, D. Progress in developing an ANN model for air pollution index forecast. *Atmos. Environ.* **2004**, *38*, 7055–7064.
49. Ni, X.Y.; Huang, H.; Du, W.P. Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data. *Atmos. Environ.* **2017**, *150*, 146–161.
50. Caruana, R. Multitask learning. In *Learning to Learn*; Springer: Boston, MA, USA, 1998; pp. 95–133.
51. Liu, J.; Ji, S.; Ye, J. Multi-task feature learning via efficient l 2, 1-norm minimization. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–21 June 2009.
52. Recht, B.; Fazel, M.; Parrilo, P.A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **2010**, *52*, 471–501.
53. Argyriou, A.; Micchelli, C.A.; Pontil, M. On spectral learning. *J. Mach. Learn. Res.* **2010**, *11*, 935–953.
54. Maurer, A. Bounds for linear multi-task learning. *J. Mach. Learn. Res.* **2006**, *7*, 117–139.
55. Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004.
56. Xu, Y.; Lin, Q.; Yang, T. Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
57. Parikh, N.; Boyd, S. Proximal algorithms. *Found. Trends Optim.* **2014**, *1*, 127–239.
58. Xiao,Y.; Li, Z.; Yang, T.; Zhang, L. SVD-free convex-concave approaches for nuclear norm regularization. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australian, 19–25 August 2017.
59. Xu, Y.; Liu, M.; Lin, Q.; Yang, T. ADMM without a Fixed Penalty Parameter: Faster Convergence with New Adaptive Penalization. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
60. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.