Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at *Big Data Cogn. Comput.* **2018**, *2*, 5; doi:10.3390/bdcc2010005

Article

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

Dixian Zhu [1]*, Changjie Cai [2], Tianbao Yang [3] and Xun Zhou [4]

[1] Department of Computer Science, University of Iowa; dixian-zhu@uiowa.edu
[2] Department of Occupational and Environmental Health, University of Iowa; changjie-cai@uiowa.edu
[3] Department of Computer Science, University of Iowa; tianbao-yang@uiowa.edu
[4] Department of Management Sciences, University of Iowa; xun-zhou@uiowa.edu
* Correspondence: dixian-zhu@uiowa.edu

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, $PM_{2.5}$ 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 exists some works applying machine learning to air quality prediction, most of the prior studies are restricted to small scale data and simply train standard regression models (linear or non-linear) to predict the hourly air pollution concentration. In this work, we propose refined models to predict the hourly air pollution concentration based on meteorological data of previous days by formulating the prediction of 24 hours as a multi-task learning problem. It 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 with several typical regularizations for multi-task learning including standard Frobenius norm regularization, nuclear norm regularization, $\ell_{2,1}$ norm regularization. Our experiments show the proposed formulations and regularization 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 composition and concentration [1]. Major outdoor air pollutants in cities include ozone (O_3), particle matters (PMs), 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 pollutant (such as O_3, PMs and SO_2) concentrations [3–5]. According to the report from the American Lung Association [6], 10 part per billion (ppb) increase in O_3 mixing ratio might cause over 3,700 premature deaths annually in the United States (U.S.). Chicago, like many other megacities in U.S, has struggled with air pollution due to the industrialization and urbanization. Although O_3 precursor (such as VOCs, NO_x, and CO) emissions have significantly decreased since the late 1970's, O_3 in Chicago has 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 human respiratory system [8]. $PM_{2.5}$, referring to particles diameter smaller than or equal to 2.5 micrometer (μm), has been increasingly concerned since they can deposit into the lung gas-exchange region-Alveoli [9]. The U.S. EPA revised the annual standard of $PM_{2.5}$ by lowering the concentration to 12 microgram per cubic meter (μg/m3) to provide improved protection against health effects associated with long- and short-term exposures [10]. SO_2, as an important precursor of new particle formation and particle growth, has also been found to be association with respiratory diseases in many countries [11–15]. Therefore, we selected O_3, $PM_{2.5}$ and SO_2 for testing in this study.

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 5; doi:10.3390/bdcc2010005

Meteorological conditions, including regional and synoptic meteorology, are critical in determining the air pollutant concentrations [16–21]. According to the study from Holloway et al. [22], the O_3 concentration over Chicago was found to be the most sensitive to air temperature, wind speed and direction, relative humidity, incoming solar radiation, and cloud cover. For example, the lower ambient temperature and incoming solar radiation slows down photochemical reactions and leads 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 (week dispersion/ventilation), the pollutants associated with traffic were found at highest concentrations [24,25]. However, the strong wind speed might form the 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 PMs, CO and SO_2), but with low concentrations of other air pollutants (such as NO_2 and O_3) due to various formation and removal mechanisms [25]. In addition, high humidity can be an indicator of precipitation events, which results in strong wet deposition leading to low concentrations of air pollutants [27]. Since various particle compositions and their interactions with light were found as the most important factors in attenuating visibility [28–30], low visibility could be an indicator of high PM concentrations. Cloud can scatter and absorb the solar radiation, which is significant for the formation of some air pollutants (e.g., O_3) [23,31]. Therefore, these important meteorological variables were selected to predict air pollutant concentrations in this study.

Statistical models have been applied for air pollution prediction based on meteorological data [33, 34,36]. However, existing studies on statistical modeling are mostly restricted to simply utilizing standard classification or regression models, which have neglected the nature of the problem itself or ignore the correlation between each hour's model. On the other hand, machine learning approaches have been developed for over 60 years and have achieved tremendous success in many areas [32, 42–46]. There exist various new tools and techniques invented in machine learning community, which allow for more refined modeling for 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 concentration based on historical metrological data and air pollution data. A striking difference between this work and the previous works is that we emphasize on 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 have collected 10 years of meteorological and air pollution data in the Chicago area. The meteorological data is from MesoWest [53] and the air pollutant data is from the EPA [51,52]. From their databases, we fetch 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 includes: meteorological variables like solar radiation, wind direction and speed, temperature, atmospheric pressure; and air pollutants include $PM_{2.5}$, O_3, SO_2. We use two methods for model regularization: (i) explicitly control the number of parameters in the model; (ii) explicitly enforce certain structure in the model parameters. For controlling the number of parameters in the model, we compare three different model formulations which can be considered in a unified multi-task learning framework with a diagonal or full matrix model. For enforcing the model matrix into a certain structure, we will consider the relationship between prediction models of different hours and compare three different regularizations with standard Frobenius norm regularization. The experimental results show that the model with the intermediate size and the proposed regularization that enforces the prediction models of two consecutive hours to be close achieve the best results and are much better than standard regression models. We also develop efficient optimization algorithms for solving different formulations and demonstrate their effectiveness through experiments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at *Big Data Cogn. Comput.* **2018**, *2*, 5; doi:10.3390/bdcc2010005

The rest of the paper is organized as follows. In section 2, we discuss related work. In section 3, we describe data collection and preprocessing. In section 4, we described 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

A large number of previous work has been done to apply machine learning algorithms onto air quality predictions. Some researchers aimed to predict targets into discretized levels. Kalapanidas et al. [34] 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, alarm) by using a lazy learning approach, Case Based Reasoning (CBR) system. Athanasiadis et al. [35] employed σ-fuzzy lattice neurocomputing classifier to predict and categorize O3 concentration into 3 levels (low, mid, and high) based on meteorological features and other pollutants like SO_2, NO, NO_2 and so on. Kunwar et al. [36] utilized principle component analysis (PCA) and ensemble learning models to predict categorized air quality index (AQI) and combined air quality index (CAQI). 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 worked on predicting concentrations of pollutants. Corani [37] worked on training neural network models to predict hourly O_3 and PM10 concentration based on data from the previous day. Performances of Feed Forward Neural Network (FFNN) and Pruned Neural Network (PNN) were mainly compared. More efforts have been made on FFNN, Fu et al. [38] applied a rolling mechanism and gray model to improve traditional FFNN models. Jiang et al. [39] explored multiple models (physical & chemical model, regression model, multiple layer perceptron) on the air pollutant prediction task and their result shows statistical models are competitive to the classical physical & chemical models. Ni, X. Y. et al. [40] compared multiple statistical models based on $PM_{2.5}$ data around Beijing, which implies linear regression models sometimes can be better than the other models.

Multi-task Learning (MTL) focuses on learning multiple tasks that have commonalities together [41], which can improve the efficiency and accuracy of the models. It has achieved tremendous successes in many fields such as: natural language processing [42], image recognition [43], bioinformatics [44,45], marketing prediction [46], etc. A variety of regularizations can be utilized to enhance the commonalities of the tasks including $\ell_{2,1}$-norm [47], nuclear-norm [48], spectral norm [49]], Frobenius norm [50], etc. However, most of former machine learning works on air pollutant prediction don't consider the similarities between the models and only focus on improving model performance for a single task.

Therefore, we decide to use meteorological and pollutant data to do prediction for hourly concentration based on linear models. In this work, we focus on three different prediction model formulations and using MTL framework with different regularizations. To the best of our knowledge, this is the first work that utilizes MTL on air pollutant prediction task. We exploit analytical approaches and optimization techniques to obtain the optimal solutions. The model evaluation metric is rooted 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 pollutants include the concentrations of O_3, $PM_{2.5}$ and SO_2 in this study. We downloaded the air pollutant data from the U.S. Environmental Protection Agency's (U.S. EPA) Air Quality System (AQS) database (https://www.epa.gov/outdoor-air-quality-data), which has been widely used for model evaluation [51,52]. We selected the meteorological variables which would affect the air pollutant

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 5; doi:10.3390/bdcc2010005

Table 1. Summary of measurement sites and observed variables

Measurement sites	Variables
Alsip Village (AV)	O_3 and $PM_{2.5}$
Lemont Village (LV)	O_3 and SO_2
Lansing Municipal Airport (LMA)	temperature, relative humidity, wind speed, wind direction, wind gust, precipitation accumulation, visibility, dew point, wind cardinal direction, pressure, and weather condition
Lewis University (LU)	The same as LMA site

concentrations including air temperature, relative humidity, wind speed, wind direction, wind gust, precipitation accumulation, visibility, dew point, wind cardinal direction, pressure, and weather condition. 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 [53].

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 south Cook County, Illinois (AQS ID: 17-031-0001; lat/long: 41.670992/-87.732457). The Lemont Village (LV) air quality monitoring site is located in a suburban residential area, which is in southwest Cook County, Illinois (AQS ID: 17-031-1601; lat/long: 41.66812/-87.99057). The weather state situated in Lansing Municipal Airport (LMA) is the closest meteorological site (MesoWest ID: KIGQ; lat/long: 41.54125/-87.52822) to the AV air quality monitoring site. The weather station positioned in Lewis University (LU) is the closest meteorological site (MesoWest ID: KLOT; lat/long: 41.60307/-88.10164) to the LV air quality monitoring site.

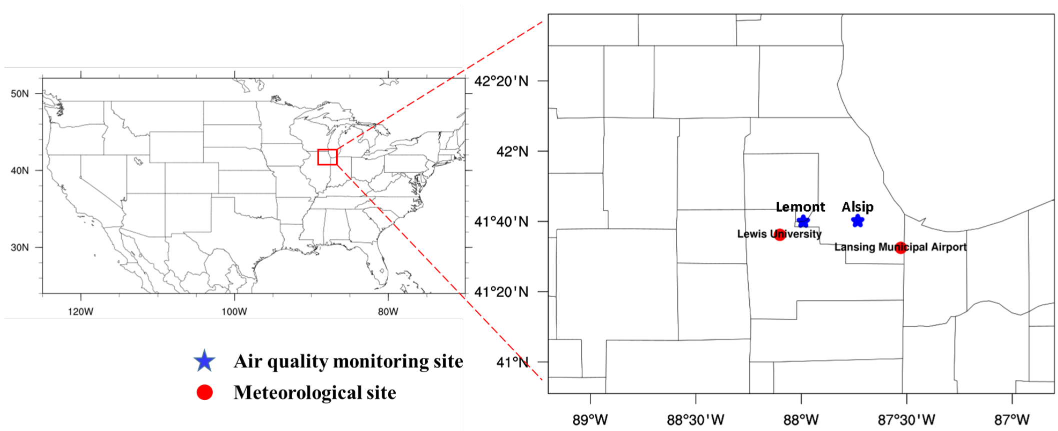


Figure 1. Locations of measurement sites. *Blue stars* denote the two air quality monitoring sites. *Red circles* denote the two meteorological sites.

3.2. Preprocessing

We paired the collected meteorological data and air pollutant data based on time to obtain the required data format of applying the machine learning methods. In particular, for each variable we form one value for each hour. However, the original data may contain multiple values or missing values at some hours. To preprocess the data, we calculated the hourly mean value of each numeric variable if there are multiple observed values within an hour, and chose the category with the highest frequency per hour for each categorical variable if there are multiple values. Missing values exist for some variables, which are not tolerable of applying the machine learning methods used in this study. Therefore, we imputed the missing value by using the closest neighbor value for four continuous

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 5; doi:10.3390/bdcc2010005

variables and one categorical variable, including wind gust, pressure, altimeter, precipitation, and weather condition. We deleted the days which still have missing values after imputing. We applied dummy coding for two categorical variables, the cardinal wind direction (16 values) and weather condition (31 values). Then, we added weekday and weekend as two boolean features. Finally, we obtained 60 features in total (9 numerical meteorological features + 16 dummy coding for wind direction + 31 dummy coding for weather condition + 2 boolean feature for weekday/weekend + 1 numerical feature for pollutant + 1 bias term). We apply normalization for all the features and pollutant targets to make their values fall in $[0, 1]$.

4. Machine Learning Approaches for Air Pollution Prediction

In this section, we will 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 next day based on the historical meteorological and air pollutant data. In this work we focus on using the former day's data to predict this day's hourly based pollutants. In particular, let $(\mathbf{x}_i; y_i)$ denote the i-th training data where $y_i \in \mathbb{R}^{24\times 1}$ denotes the concentration of a certain air pollutant at a day and $\mathbf{x}_i = (\mathbf{u}_i; \mathbf{v}_i)$ denotes the observed data in the previous day that include two components, where ';' represents column layout. The first component $\mathbf{u}_i = (\mathbf{u}_{i,1}; \ldots; \mathbf{u}_{i,D}) \in \mathbb{R}^{24\cdot D\times 1}$ include all meteorological data of 24 hours in the previous day, where $\mathbf{u}_{i,j} \in \mathbb{R}^{24\times 1}$ denotes the j-the meteorological feature of 24 hours and D is the number of meteorological features, and the second component $\mathbf{v}_i \in \mathbb{R}^{24\times 1}$ include hourly concentration of the same air pollutant in 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 denote the parameters of the model, $f(W, \mathbf{x}_i)$ denotes the prediction of 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 least 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 only one at the k-th position and zeros 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. Denote by $W = (w_0, w_1, \ldots, w_{D+1})^\top$ for this model. It is notable that this model predicts the hourly concentration based on the same hour historical data in the previous day and has $D + 2$ parameters. This simple model assumes that all 24 hours share the same model parameter.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at *Big Data Cogn. Comput.* **2018**, *2*, 5; doi:10.3390/bdcc2010005

- **Heavy Model**. The second model will take all the data in previous day into account when predicting the concentration of every hour in the second day. In particular, for the k-th 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 $w_{k,0} \in \mathbb{R}$. For this model, we define:

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

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

- **Light Model**. The third model is between the baseline model and the heavy model. It considers the 24 hours pattern of the air pollutant in the previous day and the same hour meteorological data in 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 $\mathbf{w}_{k,D+1} \in \mathbb{R}^{24\times 1}$. For this model, we define:

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

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

4.2. Regularization of Model Parameters

In this section, we describe different regularizations for the model parameter matrix W in the heavy and light model. We consider the problem as multi-task learning, where predicting the concentration of air pollutant at one hour is one task. In literature, a number of regularizations have been proposed by considering the relationship between different tasks. We first describe three baseline regularizations in literature and then present the proposed regularization that take the dimension of time into consideration for modeling the relationship between models at different time.

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

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

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

- **$L_{2,1}$-norm regularization**. The $L_{2,1}$ norm regularization has been used for feature selection in multi-task learning. It is formed by first computing the ℓ_2 norm of each row of the W matrix (across different tasks) and the 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 j-th row of W. We will consider a $L_{2,1}$-norm regularizer $\varphi(W) = \lambda\|W\|_{2,1}$.

- **Nuclear norm regularization**. 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 of using a low rank matrix is that models for consecutive hours are highly correlated, which could render the matrix W to be low rank. 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 concentration of air pollutants for two consecutive hours are close to each other. Denote 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 consecutive-close 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*

Except that Frobenius norm regularized model (with ℓ_2 norm consecutive-close regularization or not) has a closed-form solution, we solve the other models via advanced stochastic optimization techniques. Denote $F(W, \mathbf{x}_i) = [f_1(W, \mathbf{x}_i), ..., f_{24}(W, \mathbf{x}_i)]$, $Y_i = [y_{i,1}, ..., y_{i,24}]$, and the total number of feature is D. Although the standard stochastic (sub)gradient method [54] can be utilized to solve all the formulations considered in this work, it does not necessary yield the fastest convergence. To address this issue, we will consider advanced stochastic optimization techniques tailored for solving each formuation.

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

We utilize Accelerated Stochastic Subgradient (ASSG) Method [55] 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 use the proximal mapping [56]. 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 stage-wised stepsize, i is a sampled index, and $\mathbf{e}$ is a vector with all 1 as its elements. Then a proximal mapping is followed (denote by $\tilde{\lambda} = 2\eta_s\lambda$):

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 5; doi:10.3390/bdcc2010005

The above problem can solved exactly. Denote $\mathbf{w}_i$ as column vector for $W^\top$, $\mathbf{w}'_i$ as column vector for ${W'}_t^\top$. Then the solution to (4) can be computed by [47]:

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

The pseudocode of the algorithm is as following:

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

```
Input: X, Y, W_0, η_0, S, T
for s = 1,...,S do
    η_s = η_{s−1}/2
    for t = 1,...,T do
        Sample i ∈ {1,...,n}
        Update W'_t using equation (3)
        Update W_t using equation (4)
    end
    W_0 = Σ_{t=1}^T W_1/W_T
end
Output: W_0
```

4.3.2. Optimizing Nuclear norm Regularized Model

The challenge in solving a nuclear norm reguralized problem of most optimization algorithm lies at computing the full singular value decomposition (SVD) of the inovlved matrix W, which is an expensive operation. To avoid full SVD, SVD-freE CONvex-ConcavE Algorithm Extension to Stochastic Setting (SECONE-S) [57] is employed to solve the problem. The algorithm solves the following min-max 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:

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

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 following:

Algorithm 2: SECONE-S solving Nuclear norm Regularized Model

```
Input: X, Y, T, η_0, τ_0
for t = 1,...,T do
    Sample i ∈ {1,...,n}
    Update W_t and U_t using equation (6).
    η_t = η_0/√t, τ_t = τ_0/√t
end
Output: Ŵ_T = Σ_{t=1}^T W_t/T
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at *Big Data Cogn. Comput.* **2018**, *2*, 5; doi:10.3390/bdcc2010005

4.3.3. Optimizing Consecutive-Close Regularized Model

The challenge of tackling the proposed consecutive-close regularization lies that the standard proximal mapping cannot be computed efficiently. We address this challenge by using alternating directon method of multipliers. We utilize a recently proposed locally adaptive stochastic alternating direction method of multipliers (LA-SADMM) [58] to solve consecutive-close regularized model. Below, we discuss the updates for the choice of $p = 1$ (i.e., using the ℓ_1 norm) in (2). The updates for the choice of $p = 2$ can be derived similarly.

The objective function can be written as:

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

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

Then it can be solved by optimizing each variable alternatively. The update rules for stochatic ADMM (SADMM):

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

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

LA-SADMM solve the problem more efficiently by doing stage wised penalty increasing. The pseudocode for the algorithm is as following:

Algorithm 3: LA-SADMM solving Consecutive-Close Regularized problem with ℓ_1 norm

```
Input: X, Y, W_0, U_0, Λ_0, β_1, η_1, S, T
for s = 1, ..., S do
    for τ = 1, ..., T do
        Sample i ∈ {1, ..., n}
        Update W_τ, U_τ, Λ_τ using equation (8).
    end
    W_T = Σ_{τ=1}^{T} W_τ / T
    W_0 = W_T, U_0 = U_T, Λ_0 = Λ_T
    β_{s+1} = 2β_s, η_{s+1} = η_s / 2
end
Output: W_T
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at *Big Data Cogn. Comput.* **2018**, *2*, 5; doi:10.3390/bdcc2010005

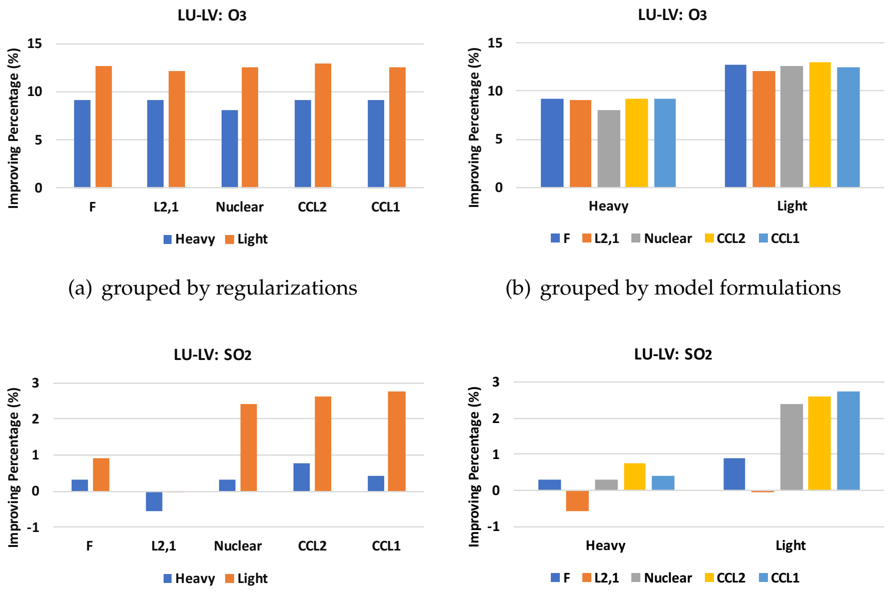
5. Experiments

We use the names of the paired air quality monitoring sites and two weather stations to denote the two datasets, i.e., LU-LV and LMA-AV. The LU-LV contains the data to predicting the concentration of two air pollutants, i.e., O_3 and SO_2. The LMA-AV contains the data to predicting the concentration of two air pollutants, i.e., O_3 and $PM_{2.5}$.

We compare 11 different models that are learned with different combinations of model formulations and regularizations. The 11 models are:

- Baseline: the baseline model with a standard Frobinus norm regularization.
- Heavy-F: the heavy model with a standard Frobinus norm regularization.
- Light-F: the heavy model with a standard Frobinus norm regularization.
- Heavy-$L_{2,1}$: the heavy model with a $L_{2,1}$-norm regularization.
- Heavy-Nuclear: the heavy model with a nuclear-norm regularization.
- Heavy-CCL2: the heavy model with the consecutive-close regularization using ℓ_2 norm.
- Heavy-CCL1: the heavy model with the consecutive-close regularization using ℓ_1 norm.
- Light-$L_{2,1}$: the light model with a $L_{2,1}$-norm regularization.
- Light-Nuclear: the light model with a nuclear-norm regularization.
- Light CCL2: the light model with the consecutive-close regularization using ℓ_2 norm.
- Light CCL1: the light model with the consecutive-close regularization using ℓ_1 norm.

We divide each dataset into two parts: training data and testing data. Each model is trained on the training data with a proper regularization parameter selected based cross-validation. Each trained model is evaluated on the testing data. The splitting of data is done by dividing all days into a number of chunks of 11 consecutive days, where the first 8 days are used for training and the next 3 days are used for testing. We use the root mean square error (RMSE) as the evaluation metric.



(a) grouped by regularizations (b) grouped by model formulations

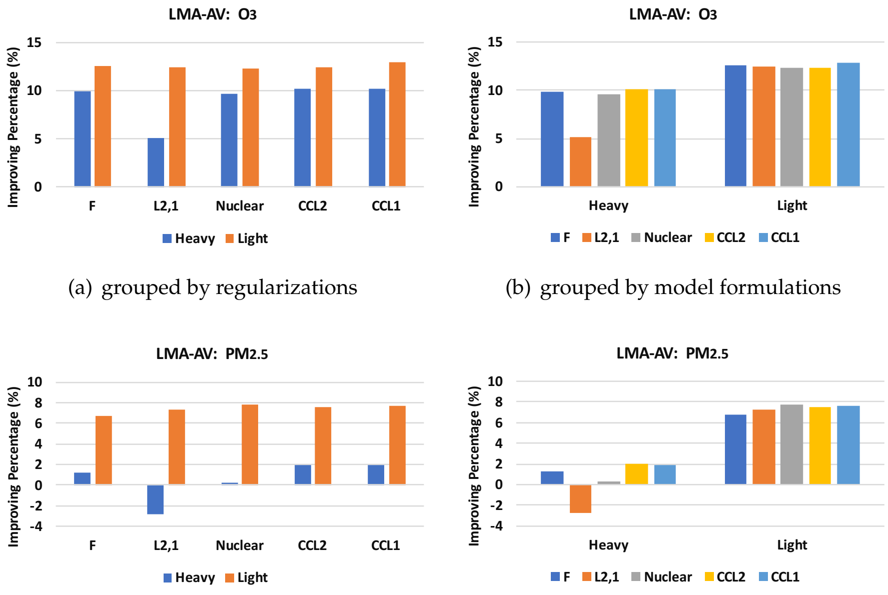
(c) grouped by regularizations (d) grouped by model formulations

Figure 2. Improvement of Different Methods over the Baseline Method on LU-LV dataset.

We first report the improvement of each method over the Baseline method. The improvement is measured by positive or negative percentage over the performance of the Baseline method, i.e., (RMSE of compared method - RMSE of the Baseline method)*100/RMSE of the Baseline Method. The results are shown in Figure 3 and 2. To facilitate the comparison between different methods, for each

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 5; doi:10.3390/bdcc2010005



(a) grouped by regularizations (b) grouped by model formulations

(c) grouped by regularizations (d) grouped by model formulations

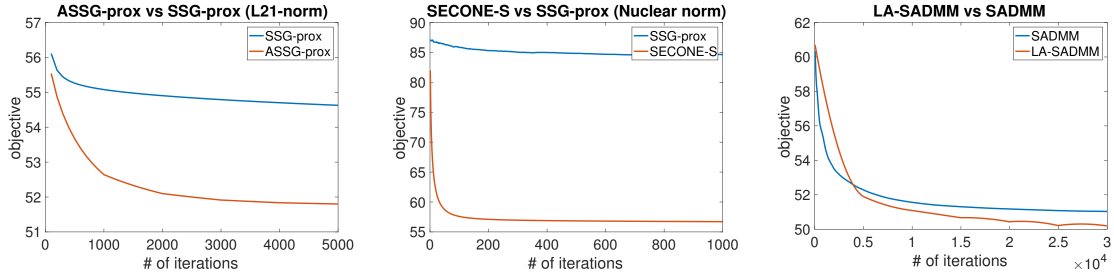
Figure 3. Improvement of Different Methods over the Baseline Method on LMA-AV dataset.

air pollutant of each dataset, we report two figures with one grouping the results by regularizations and another one grouping the results by the model formulations. From the results, we can see that (i) the light model formulation has clear advantage over the heavy model formulation and the baseline model formulation, which implies that controlling the number of parameters is important for improving generalization performance, and (ii) the proposed consecutive-close regularization yields better performance than other regularizations, which verifies that considering the similarities between models of consecutive hours are helpful. We also report the exact RMSE of each method in Table 2.

Table 2. Root Mean Squared Error (RMSE) for all approaches and datasets

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-$L2,1$	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-$L_{2,1}$	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 compare the convergence speed of the employed optimization algorithms with their standard counterparts. In particular, we compare ASSG vs SSG for optimizing $L_{2,1}$ regularized problem, compare vs SSG for solving nulcear norm regularized problem, and compare with SADMM for solving CC regularized problem. The results are plotted in Figure 4, which demonstrates that the employed advanced optimization techniques converge much faster than the classical techniques.



(a) Light-$L_{2,1}$ (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 formulated the problem as regularized multi-task learning and employed advanced optimization algorithms for solving different formulations. We have focused on alleviating model complexity by reducing the number of model parameters, and improving the performance by using structured regularizer. Our results shows that the proposed light formulation achieves much better performance than the other two model formulations, and the regularization by enforcing prediction models for two consecutive hours to be close can also boost the performance of prediction. We have also shown that advanced optimization techniques for important to improving the convergence of optimization and speed up the training process for big data. For future work, we can further consider the commonalities between nearby meteorology stations and combine them in a multi-task Learning framework, which may further provide a boosting for the prediction.

References

1. Curtis, Luke, et al. "Adverse health effects of outdoor air pollutants." *Environment international* 32.6 (**2006**): 815-830.
2. Mayer, Helmut. "Air pollution in cities." *Atmospheric environment* 33.24 (**1999**): 4029-4037.
3. Samet, Jonathan M., et al. "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* 94.pt 2 (**2000**): 5-79.
4. Dockery, Douglas W., Joel Schwartz, and John D. Spengler. "Air pollution and daily mortality: associations with particulates and acid aerosols." *Environmental research* 59.2 (**1992**): 362-373.
5. Schwartz, Joel, and Douglas W. Dockery. "Increased mortality in Philadelphia associated with daily air pollution concentrations." *American review of respiratory disease* 145.3 (**1992**): 600-604.
6. American Lung Association. *State of the air report*. New York: ALA, (**2007**): 19-27.
7. Environmental Protection Agency, EPA (**2009**), Region 5: State Designations, as of September 18, 2009, available at https://archive.epa.gov/ozonedesignations/web/html/region5desig.html, accessed 12/17/2017.
8. Hinds, William C. *Aerosol technology: properties, behavior, and measurement of airborne particles*. John Wiley & Sons, **2012**.
9. Soukup, Joleen M., and Susanne Becker. "Human alveolar macrophage responses to air pollution particulates are associated with insoluble components of coarse material, including particulate endotoxin." *Toxicology and applied pharmacology* 171.1 (**2001**): 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, Joel. "Short term fluctuations in air pollution and hospital admissions of the elderly for respiratory disease." *Thorax* 50.5 (**1995**): 531-538.
12. De Leon, A. Ponce, et al. "Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92." *Journal of Epidemiology & Community Health* 50.Suppl 1 (**1996**): s63-s70.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2017 doi:10.20944/preprints201712.0197.v1

Peer-reviewed version available at Big Data Cogn. Comput. 2018, 2, 5; doi:10.3390/bdcc2010005

13. Birmili, Wolfram, and Alfred Wiedensohler. "New particle formation in the continental boundary layer: Meteorological and gas phase parameter influence." *Geophysical Research Letters* 27.20 (**2000**): 3325-3328.
14. Lee, Jong-Tae, et al. "Air pollution and asthma among children in Seoul, Korea." *Epidemiology* 13.4 (**2002**): 481-484.
15. Cai, Changjie, et al. "Incorporation of new particle formation and early growth treatments into WRF/Chem: Model improvement, evaluation, and impacts of anthropogenic aerosols over East Asia." *Atmospheric Environment* 124 (**2016**): 262-284.
16. Kalkstein, Laurence S., and Peter Corrigan. "A synoptic climatological approach for geographical analysis: assessment of sulfur dioxide concentrations." *Annals of the Association of American Geographers* 76.3 (**1986**): 381-395.
17. Comrie, Andrew C. "A synoptic climatology of rural ozone pollution at three forest sites in Pennsylvania." *Atmospheric Environment* 28.9 (**1994**): 1601-1614.
18. Eder, Brian K., Jerry M. Davis, and Peter Bloomfield. "An automated classification scheme designed to better elucidate the dependence of ozone on meteorology." *Journal of Applied Meteorology* 33.10 (**1994**): 1182-1199.
19. Zelenka, Michael P. "An analysis of the meteorological parameters affecting ambient concentrations of acid aerosols in Uniontown, Pennsylvania." *Atmospheric environment* 31.6 (**1997**): 869-878.
20. Laakso, Lauri, et al. "Diurnal and annual characteristics of particle mass and number concentrations in urban, rural and Arctic environments in Finland." *Atmospheric Environment* 37.19 (**2003**): 2629-2641.
21. Jacob, Daniel J., and Darrell A. Winner. "Effect of climate change on air quality." *Atmospheric environment* 43.1 (**2009**): 51-63.
22. Holloway, Tracey, et al. "Change in ozone air pollution over Chicago associated with global climate change." *Journal of Geophysical Research: Atmospheres* 113.D22 (**2008**).
23. Akbari, Hashem. "Shade trees reduce building energy use and CO_2 emissions from power plants." *Environmental pollution* 116 (**2002**): S119-S126.
24. DeGaetano, Arthur T., and Owen M. Doherty. "Temporal, spatial and meteorological variations in hourly PM 2.5 concentration extremes in New York City." *Atmospheric Environment* 38.11 (**2004**): 1547-1558.
25. Elminir, Hamdy K. "Dependence of urban air pollutants on meteorology." *Science of the Total Environment* 350.1 (**2005**): 225-237.
26. Natsagdorj, L., D. Jugder, and Y. S. Chung. "Analysis of dust storms observed in Mongolia during 1937?1999." *Atmospheric Environment* 37.9 (**2003**): 1401-1411.
27. Seinfeld, John H., and Spyros N. Pandis. *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons, **2016**.
28. Koschmieder, Harald. "Theorie der horizontalen Sichtweite." *Beitrage zur Physik der freien Atmosphare* (**1924**): 33-53.
29. Appel, B. R., et al. "Visibility as related to atmospheric aerosol constituents." *Atmospheric Environment* (1967) 19.9 (**1985**): 1525-1534.
30. Deng, Xuejiao, et al. "Long-term trend of visibility and its characterizations in the Pearl River Delta (PRD) region, China." *Atmospheric Environment* 42.7 (**2008**): 1424-1435.
31. Twomey, Sean. "The influence of pollution on the shortwave albedo of clouds." *Journal of the atmospheric sciences* 34.7 (**1977**): 1149-1152.
32. Zhuoning Yuan, Xun Zhou, Tianbao Yang, James Tamerius, Ricardo Mantilla. Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study. *In 6th International Workshop on Urban Computing* (UrbComp 2017) in conjunction with ACM KDD **2017**
33. Zheng, Yu, Furui Liu, and Hsun-Ping Hsieh. "U-Air: When urban air quality inference meets big data." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, **2013**.
34. Kalapanidas, Elias, and Nikolaos Avouris. "Short-term air quality prediction using a case-based classifier." *Environmental Modelling & Software* 16.3 (**2001**): 263-272.
35. Athanasiadis, Ioannis N., et al. "Applying machine learning techniques on air quality data for real-time decision support." *First international NAISO symposium on information technologies in environmental engineering (ITEE'2003)*, Gdansk, Poland. **2003**.
36. Singh, Kunwar P., Shikha Gupta, and Premanjali Rai. "Identifying pollution sources and predicting urban air quality using ensemble learning methods." *Atmospheric Environment* 80 (**2013**): 426-437.

37. Corani, Giorgio. "Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning." *Ecological Modelling* 185.2 (**2005**): 513-529.
38. Fu, Minglei, et al. "Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model." *Neural Computing and Applications* 26.8 (**2015**): 1789-1797.
39. Jiang, Dahe, et al. "Progress in developing an ANN model for air pollution index forecast." *Atmospheric Environment* 38.40 (**2004**): 7055-7064.
40. Ni, X. Y., H. Huang, and W. P. Du. "Relevance analysis and short-term prediction of PM 2.5 concentrations in Beijing based on multi-source data." *Atmospheric Environment* 150 (**2017**): 146-161.
41. Caruana, Rich. "Multitask learning." *Learning to learn*. Springer US, **1998**. 95-133.
42. Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning." *Proceedings of the 25th international conference on Machine learning*. ACM, **2008**.
43. Fan, Jianping, Yuli Gao, and Hangzai Luo. "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation." *IEEE Transactions on Image Processing* 17.3 (**2008**): 407-426.
44. Widmer, Christian, et al. "Leveraging Sequence Classification by Taxonomy-Based Multitask Learning." *RECOMB*. **2010**.
45. Kshirsagar, Meghana, Jaime Carbonell, and Judith Klein-Seetharaman. "Multitask learning for host-pathogen protein interactions." *Bioinformatics* 29.13 (**2013**): i217-i226.
46. Lindbeck, Assar, and Dennis J. Snower. "Multitask learning and the reorganization of work: From tayloristic to holistic organization." *Journal of labor economics* 18.3 (**2000**): 353-376.
47. Liu, Jun, Shuiwang Ji, and Jieping Ye. "Multi-task feature learning via efficient l 2, 1-norm minimization." *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, **2009**.
48. Recht, Benjamin, Maryam Fazel, and Pablo A. Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." *SIAM review* 52.3 (**2010**): 471-501.
49. Argyriou, Andreas, Charles A. Micchelli, and Massimiliano Pontil. "On spectral learning." *Journal of Machine Learning Research* 11.Feb (**2010**): 935-953.
50. Maurer, Andreas. "Bounds for linear multi-task learning." *Journal of Machine Learning Research* 7.Jan (**2006**): 117-139.
51. Foley, K. M., et al. "Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7." *Geoscientific Model Development* 3.1 (**2010**): 205.
52. Yahya, Khairunnisa, et al. "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." *Atmospheric Environment* 152 (2017): 562-583.
53. Horel, John, et al. "Mesowest: Cooperative mesonets in the western United States." *Bulletin of the American Meteorological Society* 83.2 (**2002**): 211-225.
54. Zhang, Tong. "Solving large scale linear prediction problems using stochastic gradient descent algorithms." *Proceedings of the twenty-first international conference on Machine learning*. ACM, **2004**.
55. Xu, Yi, Qihang Lin, and Tianbao Yang. "Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence." *International Conference on Machine Learning*. **2017**.
56. Parikh, Neal, and Stephen Boyd. "Proximal algorithms." *Foundations and Trends in Optimization* 1.3 (**2014**): 127-239.
57. Xiao, Yichi, et al. "SVD-free Convex-Concave Approaches for Nuclear Norm Regularization." *IJCAI*. **2017**.
58. Xu, Yi, et al. "ADMM without a Fixed Penalty Parameter: Faster Convergence with New Adaptive Penalization." *Advances in Neural Information Processing Systems*. **2017**.