FIELD CALIBRATION OF LOW COST AIR QUALITY MONITORS IN TWO CHINESE CITIES WITH NON-LINEAR REGRESSION AND ARTIFICIAL NEURAL NETWORK TECHNIQUES

A Thesis

by

ZHENGLU WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Qi Ying
Committee Members,	Xingmao Ma
	Renyi Zhang
Head of Department,	Robin Autenrieth

May 2018

Major Subject: Civil Engineering

ABSTRACT

Exposure to fine particular matter ($PM_{2.5}$) is associated with various adverse health outcomes, including cardiovascular disease, cancer and respiratory related diseases. Reducing exposure assessment errors for epidemiologic studies, and facilitating new models of community-engaged research may benefit from improved understanding of the spatiotemporal distribution of PM with respect to mobile and stationary sources of particular emissions. In this study, the performance of a low-cost PM monitor based on the Shinyei PPD42NS PM sensor was tested and calibrated with a beta-attenuation based reference $PM_{2.5}$ monitor (BAM-1020) in two Chinese cities (Nanjing and Chengdu) using linear, power-law and artificial neural network (ANN) approaches. The first eight months' data in Nanjing showed that the low-cost monitor can provide reasonably accurate estimation of hourly $PM_{2.5}$ under non-condensing conditions (RH<95%). Among all three calibration methods, the ANN approach shows the highest correlation between the estimated and BAM-1020 measured hourly PM2.5 (R^2=0.76). PM2.5 estimated from the power-law equation demonstrates a slightly better agreement (R^2=0.70) with BAM-1020 hourly PM2.5 than linear fit method (R^2=0.68). Approximately 73% of the hourly PM2.5 estimated by the low-cost monitor with the ANN calibration approach is within the low-cost monitor performance guideline of the Ministry of Environment of China, which is better than linear and power-law approaches (approximately 64% and 67%, respectively). The better performance of ANN is mainly due to including temperature and relative humidity (RH) as input data in addition to the raw sensor output of Low-pulse Occupancy

Ratio (LOR). The performance of the low-cost monitor deteriorates after 5-6 months on continuous operation in polluted environments. R^2 for the first, second and third four-month periods based on the ANN approach are 0.80, 0.64 and 0.24, respectively. This suggests that regular replacement or cleaning of the PM optical sensing unit is needed to use the low-cost monitor for long-term community monitoring. However, in terms of monthly average concentrations, the low-cost monitor has a small error of 10%, even after long operation periods. The consistency of the low-cost sensors is also tested in this study. Poor correlation in sensor raw readings was found among three collocated low-cost monitors. This suggests that a screening of the sensors of consistency is needed to ensure consistent results. Also, the calibration parameters developed using the low-cost monitor data collected in Nanjing do not lead to good estimations of PM2.5 when sensor data in Chengdu are used. Variation of the sensor-to-sensor responses or different weather conditions are possible causes, but the root cause of the problem is still unclear and requires more investigation.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor, Professor Ying Qi for his continuous support of my study and research. During the two years working with Dr. Ying, I learned something that can benefit the rest of my life. His guidance helped me in all the time of research and writing this thesis. I could not have a better advisor and mentor for my Master's degree study.

I would also like to thank the rest of my committee: Professor Xingmao Ma, Professor Renyi Zhang, for their encouragement and insightful comments.

My sincere thanks also goes to Texas A&M University and Civil Engineering College for providing all the amazing resources.

Last but not the least. I would like to thank my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. I would not achieve the accomplishment I have right now with them. Thank you!

CONTRIBUTORS AND FUNDING SOURCES

Many analyses from this study were conducted through the advanced computing resources provided by Texas A&M High Performance Research Computing (HPRC) group.

This work was supervised by a thesis committee consisting of Dr. Ying qi advisor and committee chair. Other committee members include Dr. Xingmao Ma of Department of Environmental Engineering and Dr. Renyi Zhang of Department.

NOMENCLATURE

ANN	Artificial Neural Network
BAM	Beta Attenuation Monitor
CMAQ	Community Multi-Scale Air Quality System
COPD	Chronic Obstructive Pulmonary Disease
FEM	Federal Equivalent Method
GDP	Gross Domestic Product
IRED	Infrared Emitting Diode
LOD	Limit of Detection
LOR	Low Pulse Occupancy Channel Ratio
MLR	Multi Linear Regression
MNB	Mean Normalized Bias
MNE	Mean Normalized Error
PANDA	Portable and Affordable Nephelometric Data Acquisition System
PM	Particulate Matter
PVWP	Portable University of Washington Particle
TEOM	Tapered Element Oscillating Microbalance
VGP	Village Green Project
WHO	World Health Organization

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1 INTRODUCTION

1.1 Statement of Problems

Air pollution has caused a significant amount of health related problems. In 2014, 92% of the world population lived in the places that did not meet WHO air quality guideline requirements (WHO, 2016). There are over 3.2 million premature deaths and more than 76 million people lost disability-adjusted life time each year (Hankey, Marshall, & Brauer, 2012; Lim et al., 2012). The rapid Chinese economic growth, in which GDP grew on average 10% each year for over a decade, has come at the expense of its environment and public health (Xu, 2016). The average annual exposure to $PM_{2.5}$ in 272 Chinese cities was 56 $\mu g/m^3$, significantly above the WHO air quality guideline of 10 $\mu g/m^3$. For each 10 $\mu g/m^3$ increase in air pollution, there will be a 0.29 percent increase in all respiratory mortality and a 0.38 percent increase in chronic obstructive pulmonary disease (COPD) mortality. (Chen et al., 2017). Among all the air pollutants, $PM_{2.5}$ and ozone are ranking high on the health burden factor list around the world (Holstius, Pillarisetti, Smith, & Seto, 2014).

Therefore, it is crucial to monitor the real time air quality. In the urban areas of many developed countries, there might be one air quality monitoring station that covers about 100,000 people, whereas in the cities of some severely polluted developing countries, one air quality monitoring station might cover millions of people. For instance, there are about 300 monitoring stations in the United Kingdom, but only around 600 monitoring sites in India (DEFRA, 2011; CPCB, 2017). Same

problem was in China with only about 1500 air quality monitoring sites covering over 190 cities in 2015 (Rohde & Muller, 2015). The existing air quality monitoring infrastructure cannot fulfill the need to characterize urban $PM_{2.5}$ concentrations spatial and temporal variability (Holstius et al., 2014).

At the same time, most of the existing air quality measurement devices are made for professionals. They are expensive and are normally not accessible to the general public. Because of this need, small, low-cost and portable air quality detection devices have a great market potential. Also they can greatly improve the ability to generate higher resolution of spatial and temporal aerosol concentrations with a fairly low system cost. The data collected from this low-cost monitoring system can be utilized for supplementing the official air pollution monitors and provide a better understanding of exposure estimates to raise air pollution awareness in communities. (Abraham & Li, 2016)

However, data quality is a major concern of these low-cost monitors. Several studies have begun to investigate the utility of low-cost monitors in charactering outdoor and indoor air pollutant levels (Klepeis et al., 2013; Snyder et al., 2013). Unreliable data may mislead the users to make inappropriate decisions, such as outdoor activities, school cancellation, etc. For data quality assurance, sensor calibration process is necessary. It includes routine laboratory calibration under certain conditions by the end-user. For large sensor systems, it is more practical to utilize statistical techniques for data quality assurance. Before using low cost sensors, their characteristics need to be first understood. For instance, detection limit, range of

concentration, influence of temperature and relative humidity, consistency between individual sensors, temporal variation sensitivity, etc.

1.2 Background

There has been a considerable amount of previous academic researches on low cost particular matter sensors. Major themes of this study are outlined as following:

- Build of the low-cost air quality monitoring platforms.
- The performance of low cost sensors & data quality analysis with consideration of potential factors, such as temperature and relative humidity.
- Nonlinear regression and artificial neural network techniques used to correlate sensor reading with particulate matter reading from reference instruments.

1.2.1 Build of the Low-Cost Air Quality Monitoring System

There are many different custom-build low cost air quality monitoring platforms. For instance, Portable and Affordable Nephelometric Data Acquisition System (PANDA), Village Green Project (VGP), Portable University of Washington Particle (PUWP) and a lot more (Gao, Cao, & Seto, 2015; Holstius et al., 2014; Jiao et al., 2015). The design considerations for most platforms included: (1) compact physical size and structural design supporting their portability, (2) enabling real-time data acquisition for air pollution monitors, (3) wireless communication features, (4)

components are commercially available to smooth technology transfer and increase system flexibility, (5) collected data quality assurance.

The platforms included in the above studies always operate alongside an officially approved reference instrument. Beta Attenuation Monitors (BAM-1020) is a very popular EPA federal equivalent method (FEM) reference instrument for $PM_{2.5}$ and PM_{10}. Tapered element oscillating microbalance (TEOM) is often used to evaluate the accuracy of low cost sensors' measurements as well. TEOM based instruments have been approved by US EPA for air quality monitoring as well (Taylor, 2016).

The most essential component of a platform is the low-cost sensor, which is the one of the central topics of this study. Shinyei PPD42NS and Sharp GP2Y are two of the preferred sensors used by varies researchers. The PPD42NS sensor has a partially enclosed chamber with a single light emitting diode, a plastic lens, and an optical receiver at a forward angle of approximately 45 degrees. It also comes with a removable cap, which makes it easier to swab residue off the lens. The details are showed in Figure 1-1. The Sharp GP2Y contains an infrared emitting diode (IRED) and a phototransistor. The IRED illuminate particles in the air flow with a 10ms pulse driven waveform with a duty ratio of 0.032. Scattered light intensity is converted to a 0-3.5 V analog signal by phototransistor (Li & Biswas, 2017). Both of the sensors cost less than 15 US dollars.

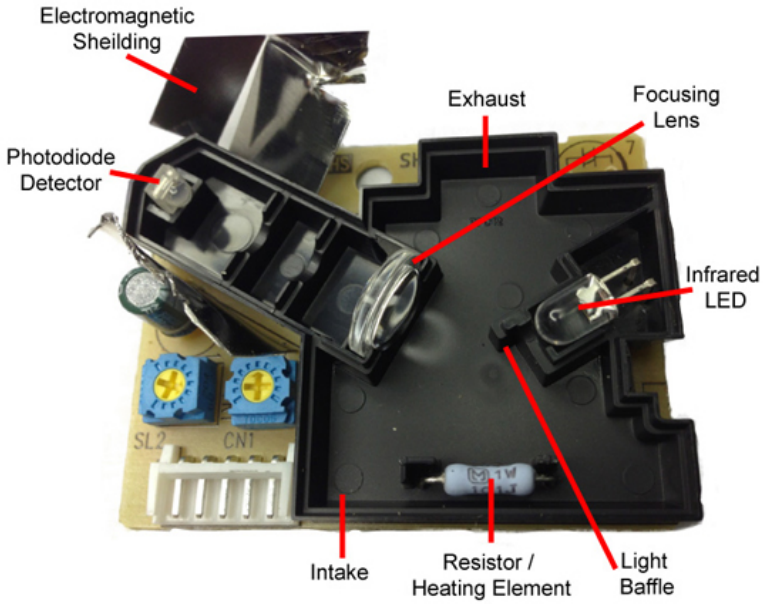

Figure 1-1. Inside the Shinyei PPD42NS (Michael, 2013). It demonstrates the structure of PPD42NS sensor and the function of each component.

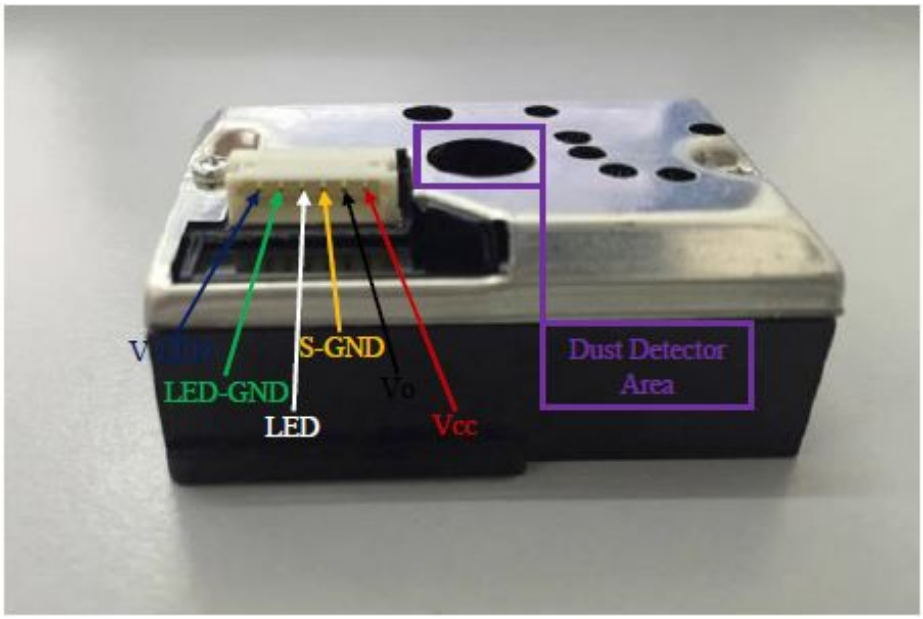

Figure 1-2. Inside the Sharp GP2Y Dust Sensor (mybotic, 2016)**.** It demonstrates the structure of Sharp sensor and the function of each component.

1.2.2 Sensor Performance and Data Quality

A good amount of studies has been reported on the performance of low cost optical PM sensors (Austin, Novosselov, Seto, & Yost, 2015; Holstius et al., 2014; Sousan et al., 2016; Wang et al., 2015). Most of the studies found a good linear relation between the readings of optical PM sensors and PM mass concentration from non-FEM instruments, such as SidePak or DustTrak (Amaral, de Carvalho, Costa, & Pinheiro, 2015), scanning mobility particle sizers (SMPSs) and aerodynamic particle sizers (APS) (Sioutas, Abt, Wolfson, & Koutrakis, 1999). However, using SidePak and DustTrak as reference instrument is problematic, because both are photometric sensors and all PM sensors used in previous studies are based on the same principle, which means if the principle itself has problem, then all the sensors could not sufficiently show the problem (Liu, Zhang, Jiang, & Chen, 2017). SMPS and APS estimated PM mass concentration from number-based particle size distributions under the assumption that particles are spherical and particle density is known. The measuring cycle of SMPS/APS with some low-cost PM sensors may different, which could bring another issue that is the time delay between the responses of reference and evaluated PM sensors. Time delay could result in potential error during the sensor performance calibration process. Therefore, calibrate PM sensors in laboratory by using proper reference instruments such as BAM-1020 and TEOM can provide high

accuracy measurements and they are independent of the size composition and distribution of particulate matter.

The specification of low cost and portable PM sensors, such as model size, sampling method, detection principle, detectable size range, concentration range and output signal, were tested in a previous study (Liu et al., 2017). Four low cost optical PM sensors, Shinyei PPD42NS, Samyoung DSM501A, Sharp GP2Y1010AU0F and Oneair CP-15-A4, were selected because of their compact sizes and popularity. All these optical PM sensors are based on light scattering technique. Thermal-driven upwind flows are generated in both Shinyei and Samyoung PM Sensors. Different from Shinyei and Samyoung, the Oneair sensor has a build-in fan to move particles through the sensing volume. However, the Sharp sensor only relies on diffusion to move particles through the sensor. Their results showed that the responses of all tested PM sensors have a linear relationship with particle mass concentration, but the composition of particles had a significant effect on the performance of optical PM sensors.

Study shows that a simple linear function is generally sufficient to provide moderate to high correlation coefficient values when the sensor response was calibrated with the reference measurements (Rai et al., 2017). However, a few other investigations (Austin et al., 2015; Kelly & Sukhatme, 2017; Manikonda, Zikova, Hopke, & Ferro, 2016; Wang et al., 2015) have reported that the sensor response started to saturate at high particle concentrations (higher than 50-100 $\mu g\ m^{-3}$) so that other data fitting methods, such as power-law or higher order polynomial functions

(Rai et al., 2017), will be needed to correlate the measurements. Therefore, it is essential to choose an appropriate method to calibrate particular sensors.

Low cost sensors typically perform better (with high R^2 values) under laboratory conditions (R^2 ranges from 0.7 to almost 1) compared with field condition (R^2 ranges from 0.4 to nearly 1). The deteriorated performance in the real-world condition is because of changing particle composition, sizes and the variation of environmental factors. Thus, on-site calibration is very important and laboratory calibrated sensors should not be used directly in field measurement (Rai et al., 2017). Furthermore, study shows that small variations in the sensor specifications may lead to significant differences in their responses (Hojaiji, Kalantarian, Bui, King, & Sarrafzadeh, 2017), the variation of the sensors needs to be better understood by co-located field measurements.

Before analyzing the data collected by low-cost sensors, all measurements were first time-matched to the nearest second (Austin et al., 2015). The sensor data were collected over 5-second intervals. Spline function was used to interpolate concentrations between the 5-second slots. The Shinyei PPD42NS sensors collect data for 1 second intervals. After being time-matched, the data were averaged to every 2-minute.

Coefficients of determination were used to quantify and compare the strengths of correlations for pairwise data. Empirical and simulated R^2 were computed for two collocated BAM-1020 to give the range of 1-hour integration R^2 values. Root mean squared errors (RMSE) were computed to provide the accuracy of linear calibrations

(Holstius et al., 2014). Shinyei PPD42NS sensor was used in the same study to test the performance of the sensor when compared with reference instruments. The results showed high correlations between individual low-cost monitors and reference instrument. The accuracies of linear models based on each monitor were essentially the same (RMSE for PANDAs, GRIMM, Dylos, and DustTrak = 3.4-3.6; 3.4 and 3.5; and 3.5 $\mu g\ m^{-3}$ respectively). Given the substantial cost differences, the agreement of PANDAs and commercially available devices was outstanding (Holstius et al., 2014). Similar study (Wang et al., 2015) has been done to evaluate the calibration methods of three low cost sensors, Shinyei PPD42NS, Samsung DSM501A and Sharp GP2Y. The response of the three sensors agreed well on particle mass concentration fall in the range of 0-1000 $\mu g\ m^{-3}$. The correlations of the sensors were higher than 0.78 after pairwise output readings.

The limit of detection (LOD) of the Shinyei sensor was calculated based on 348 observations for which the measurement was 1 $\mu g/m^3$ or less (Austin et al., 2015). The result proves that LOD of Shinyei is low enough to make it an appropriate sensor in most ambient and indoor conditions. Same study also shows that the precision of Shinyei sensors is quite high with extremely small standard deviation estimated for the slope and high R^2. This also indicates that the accuracy of Shinyei PPD42NS, when idiomatic sensor response is accounted for, is acceptable for wide deployment.

Relative humidity affected the performance of particle sensors in different ways. Water in the air may absorb infrared radiation and cause an overestimation of particle mass concentration because of the lessened light intensity. Also, high water

vapor content in the air could cause malfunction of the sensors. Last but not least, the reference instrument may also generate inaccurate outputs under high humidity conditions (Wang et al., 2015). Environmental factors, such as relative humidity, temperature and light, have been assessed by several studies (Austin et al., 2015; Hojaiji et al., 2017; Manikonda et al., 2016; Rai et al., 2017). For instance, an environmental chamber was used to study the effect of changing environmental factors on the performance of low-cost sensors (Wang et al., 2015). The study compared the outputs under different temperature and relative humidity conditions, while maintaining similar PM mass concentration. As they increased relative humidity from 20% to 67%, sensor output (Low Pulse Occupancy) first increased by 7-8% and then decreased 15% when relative humidity rose from 75% to 90%. Whereas the impact of temperature on sensor was ignorable compared to relative humidity.

Similar study showed that temperature and relative humidity had a large effect on output dust concentrations (Hojaiji et al., 2017). The results demonstrated that Shinyei sensor was less sensitive in high humidity conditions. The possible reasons for changing output with different relative humidity can be explained by (1) water absorbs radiation and results an overestimation of particle concentration, (2) inappropriate reference instrument used under high relative humidity. For example, scanning mobility particle sizers (SMPS). (3) possible failure of particle sensors at high relative humidity (Wang et al., 2015).

However, there are other investigations showing environmental factors such as temperature and relative humidity have negligible effect on sensor output (Bart et al.,

2014; Holstius et al., 2014; Jiao et al., 2015). (Holstius et al., 2014) averaged hourly data from the reference instrument (BAM-1020), using only Temperature, Relative Humidity and Light as predictors, served as negative controls: if there is no confound data, there should be no association. They found that variability in 1-hour BAM output cannot be explained by light or temperature. Though 1-hour relative humidity had some ability to predict 1-hour BAM responses.

1.2.3 Neural Network Account for Temperature and Relative Humidity

Artificial Neural Network (ANN) has been widely used to solve non-linear data fitting problems, which do not have explicit equations between dependent and independent variables.

$$True\ PM2.5 = f(Raw\ Readings, T, RH, \dots)$$

As demonstrated in Figure 1-2, ANN is based on many connected nodes or so called artificial neurons. The neurons between each connection can transmit signals to each other. The number of hidden neurons was determined through a series of neural network configurations with 1 to 100 neurons. Model performs better with 10-50 hidden neurons (Zu et al., 2017). Generally, the signals between connections are real numbers, and the output of neurons is calculated by using a non-linear function to sum up the corresponding inputs. Typically, neurons and connections have a weight that can be adjusted with the learning process. Data or signals are transferred from one

layer to another via several series processes. Input layer contains input parameters. Hidden layers evaluate input parameters by different algorithms and transfer them to output layer (Ozkan, 2010).

Recent studies applied statistical techniques to predict PM concentrations by using the readings of other pollutants or the past measurements. ANN is one of the leading statistical techniques to make predictions and its pattern recognition function can be well applied in environmental field. Studies also showed that ANN generally performed better than other prediction models. (Park et al., 2018)

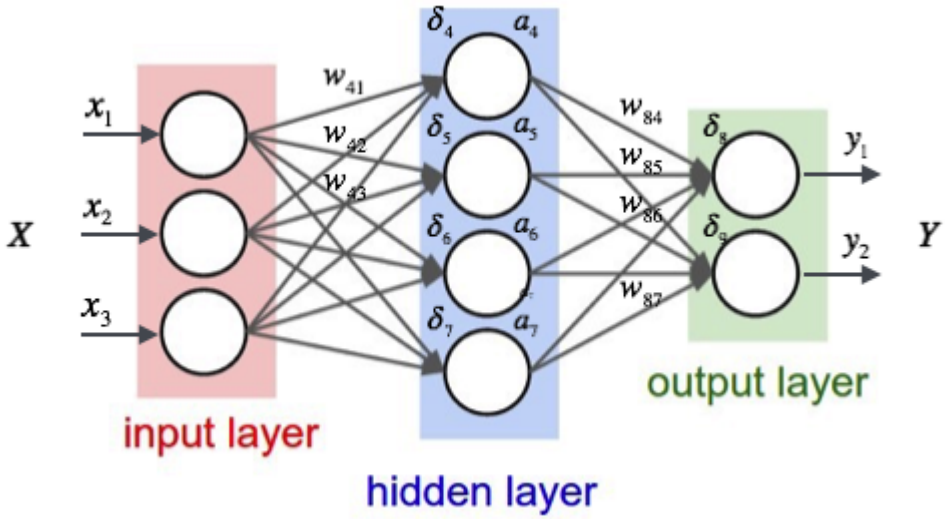

Figure 1-3. Artificial Neural Network System (Dormehl, 2017). It explains what is artificial neural network and the principle behind neural network.

In the US, EPA developed the Models-3 Community Multi-Scale Air Quality system (CMAQ) to provide 24-48 hours' air quality forecasts. Because of its ability to model different pollutants independently, CMAQ has been widely used for modeling air pollution for nearly two decades. The CMAQ forecast function was compared with Artificial Neural Network. The result showed that ANN generally results more

accurate prediction of future pollution levels for a single grid cell (Lightstone, Moshary, & Gross, 2017). However, it is limited in predicting extreme events.

To set up the Neural Network, the collected $PM_{2.5}$ data were used as the input. The neural network was created and tested using historical data. The input includes surface air temperature, surface pressure, planetary boundary layer height (PBLH), relative humidity and horizontal wind. Seasonal variation was taken into account as an input as well. In order to optimize the neural network performance, some preliminary tests were done to better use meteorological input variables. The most obvious result was the dramatic improvement of neural network during night and morning hours, where CMAQ model has the most trouble with. This is because of the machine learning approach in neural network, where the input, forecast periods, time have a significant effect on output performance (Lightstone et al., 2017).

Researchers (Zu et al., 2017) developed artificial neural network by using the *fitnet* and *train* functions in MATLAB. The levenberg-Marquardt network was deployed using 10 hidden nodes. As a means to provide better and more reliable predictions, 20 neural networks were trained and averaged to give final prediction. Compared to Multi Linear Regression (MLR) analysis, the ANN matches the measured concentration better, with R^2 of 0.74 and root mean square error (RMSE) of 32.05. However, the model still has difficulties to predict well under high pollutant levels.

1.3 Research Objectives

The objective of this study is to test the performance of custom-built low-cost monitor with Shinyei PPD42NS sensor by using different data fitting methods to compare the collected output with reference instrument values in two Chinese cities: Nanjing and Chengdu.

There are extensive researches on comparing low cost gas sensors and calibrating data with artificial neural network, but very few of them combined these two aspects together. Further, low-cost sensor (Shinyei PPD42NS) has been tested not performing well under laboratory high relative humidity conditions, but rarely tested under real world conditions.

In this study, we will be looking at two different sites (Chengdu, Sichuan Province and Nanjing, Jiangsu Province) in China. Chengdu is located in one of the most polluted and humid regions in China. The primary air pollutants were PM2.5 and PM10, with annual concentration 7 times higher than the WHO guideline values. Nanjing is located in the Yangtze River Delta – one of the most economic developed regions in China, and is suffering from poor air quality as well (Qiao, Jaffe, Tang, Bresnahan, & Song, 2015). We deployed one low-cost monitor in Chengdu and another three monitors side by side in Nanjing. More details of the data collection and analyses are documented in Chapter 2.

2 FIELD CALIBRATION OF LOW-COST AIR QUALITY MONITOR

2.1 Introduction

Particular matter is a mixture of particles which differ in size and composition. It is generated by a wide variety of nature and anthropogenic activities. According to WHO, particular matter is a mixture of solid and liquid particles of inorganic and organic substance suspended into the air. It could be dust, smoke, ash, soot, exhaust particles and other particles (WHO, 2016). As mentioned in the previous chapter, portable low-cost PM monitors could substantially improve the resolution of spatial and temporal aerosol concentrations. The data collected from these monitors could be utilized for supplementing the officially approved air quality monitors. However, many researchers expressed concerns over the quality of data from low-cost PM sensors and the consistency of the sensor performance. Extensive field calibration of these low-cost monitors with reference instruments to understand their performances under different environmental conditions and their operational lifespan are essential in wide applications of these potentially very useful instruments in environmental quality and epidemiology studies. Previously, a low-cost optical aerosol sensor Shinyei PPD42NS was calibrated with official approved air quality monitor conducted by a regulatory agency and with some other optical sensors in Xi'an China. The results

demonstrated some great benefits of using affordable sensors at some severe polluted areas. Also, studies showed the good ability of PPD42NS to capture spatiotemporal variabilities in various environmental conditions (Gao et al., 2015).

2.2 Methodology

2.2.1 Low-Cost Air Quality Monitor

A custom-built low-cost monitor (Figure 2-1) used in this study incorporates the Shinyei PPD42NS (Shinyei Corp, 2010) as the PM sensor. It also incorporates a temperature and relative humidity sensor (DHT22), a real time clock (RTC) module (DS3231), an SD card extension board, a mini-fan and a small 1-inch OLED display (128x64 pixels). All these components are controlled by an Arduino MEGA 2560 R3 microcontroller and placed in a 3D printed ABS case with a dimension 30x15x10 cm^3. The low-cost monitor can be powered through the 9V or 5V DC connects on the Arduino board. The estimated cost of the whole device is about $50-60.

The Shinyei PPD42NS sensor has been proved to be a reliable sensor in normal weather conditions (Austin et al., 2015). The PPD42NS sensor is designed to detect particles larger than 1 μm in diameter. The sensor operates on a 5V (±10%) DC power supply. When the infrared light from the LED light emitter scattered by the particles is detected, the voltage on the output pin drops from 4.0V to 0.7V, creating a low voltage pulse that has a width of 10-90ms. The standard operation of the sensor requires counting the low pulse occupancy time for at least 30 sec and the fraction of

low pulse occupancy time during that accumulation period demonstrates a distinct dependency with the number concentration of particles, which is clear from the calibration curve shown in the factory specification sheet. Nominally, it has a detection limit of 0-2.8×10^7 particles m^{-3}. Since the factory calibration curve is for cigarette smoke particles, to apply the sensor for ambient air quality monitoring and to report particle mass concentrations, it is necessary perform field calibrations to correlate the raw low occupancy ratio (LOR) output with particle mass concentrations reported by a reference monitor. PPD42NS has two LOR output channels, one for particles greater than ~1 μm (LOR1) and the other for particles greater than ~2.5 μm (LOR2). The operation environmental conditions are temperature within 0-45°C, and RH less than 95% (without dew condensation). Study (Gao et al., 2015) shows that there is an approximate error of 25% in detecting particles under most of the operational range. At very low concentrations, the error becomes enormously larger. For instance, when the particle density is less than 100,000 particles per cubic meter, the error could be over 50%. The DHT22 is a relatively low cost temperature and humidity sensor. It operates on 5V DC (accepts 3.3-6V). Temperature measurement has a resolution of 0.1°C and error of < ±0.5°C. RH measurement has a resolution of 0.1% and error of ±2%. The operation range of the temperature and RH sensors are -40°C – 80°C and 5% RH - 99%RH, respectively. The DS3231 is a low-cost, extremely accurate RTC with an integrated temperature-compensated crystal oscillator (TCXO) and crystal. The DS3231 has an error of 2 ppm per day (0.17 seconds per day) under typical ambient environment temperature. It communicates with the Arduino broad using the

industrial standard I²C interface. This ensures that no additional adjustment of the clock is necessary during the expected lifetime of the monitor. The DS3231 used in this project has an external battery to ensure continuous operation when power supply to the low-cost monitor is disconnected. Programming of the instrument was done under the Arduino IDE development environment.

The low-cost monitors were configured to save two-minute average of LOR1 and LOR2, and relative humidity and temperature. In addition, instantaneous reading of fan speed (rpm) and internal DC voltage (mV) were also recorded. All recorded data are timestamped with time readings from the RTC.

Figure 2-1. Custom Build Low-Cost PM2.5 Monitor

2.2.2 Study Location

We chose two cities in China as our study locations – Chengdu in Sichuan Province and Nanjing in Jiangsu Province. As we mentioned, both of the cities are

suffering severe air pollution problems. Chengdu, the provincial capital of Sichuan Province, is a mega city located in the west side of the heavily polluted Sichuan Basin and the year-round climate is very humid (around 79-84%). Nanjing is sitting next to Yangtze River with four-season climate and the yearly average humidity is much lower than Chengdu. To calibrate the raw sensor readings and to assess inter-monitor variations, three low-cost PM monitors were co-located alongside with a beta-attenuation monitor (BAM-1020) on the campus of Nanjing University of Information Science and Technology (NUIST) in an open field. These three monitors are named A001-A003 in the following discussions. Another low-cost monitor was placed on campus of Sichuan University in Chengdu, which is also co-located with a BAM-1020 monitor. The A001 monitor in Nanjing is the first prototype and collected data from December 2015 to June 2017. A002 and A003 were made later and collected data from July 2016 to June 2017. The low-cost monitor in Chengdu was in operation from July 2016 to January 2017

2.2.3 Data Analysis

Linear regression, non-linear regression with a power law function and the artificial neural network (ANN) technique are used to relate the LOR readings with the BAM-1020 readings. The linear regression and power law regression only use hourly average LOR1 as the independent variable and the hourly BAM-1020 as the dependent variable, as shown in the equation (1) and (2),

$$C_{PM2.5}^{BAM} = aLOR1 + b \quad (1)$$

$$C_{PM2.5}^{BAM} = a(LOR1)^b \quad (2)$$

where $C_{PM2.5}^{BAM}$ represents hourly concentrations of PM2.5 measured by BAM-1020, and a and b are parameters to be determined from the regression analysis.

The ANN technique has been developed and applied in previous studies (Lightstone et al., 2017; Park et al., 2018). It is modified in this study to correlate the low-cost monitor raw data (LOR1, LOR2, temperature and RH, and hour of the day) with the BAM-1020. A summary of the technique is provided below. The ANN used in this study is a two-layer feed-forward neural network (NN) with one hidden layer. The optimal number of neurons in the hidden layer will be determine in this study. The fitnet and train functions in MATLAB's Neural Network Toolbox were used to create and train the NNs, respectively. The Levenberg-Marquardt (Hagan & Menhaj, 1994) was selected to use in the training function to determine the weight and bias parameters for the NN. An ensemble approach (Hansen & Salamon, 1990) that weight-average twenty separately-trained NNs was used to provide a final estimation of PM2.5 from the inputs. This ensemble approach could provide more stable estimations than a single NN (Rai et al., 2017). The weight of each NN (w) in the ensemble was determined by minimizing an objective function Q:

$$Q = \sum_{i=1}^{N} \left[C_{PM2.5,i}^{BAM} - \sum_{m=1}^{20} w_m C_{PM2.5,i}^{ANN,m} \right]^2 \quad (3)$$

Where N is the number of hourly PM2.5 data (including both testing data and validation data); $C_{PM2.5,i}^{BAM}$ is the BAM-1020 reported PM2.5 concentration for the ith

data point; $C_{PM2.5,i}^{ANN,m}$ is the estimated $PM_{2.5}$ for the i^{th} data point by the m^{th} NN in the ensemble. Typically, less than 10 members of the ensemble have a weight factor greater than 0.01. Thus, a 20-member ensemble seems sufficient to get optimized result and additional members are not likely lead to better performance. In each ensemble ANN run, the entire set of ANN model training data were partitioned randomly into 70% training data and 30% cross-validation data. The cross-validation is an effective way of reducing over-fitting during ANN training. The parameters from the trained ANN ensemble were saved in order to make prediction for the $PM_{2.5}$ in other sites.

The data collected by A001 at NUIST from the first year of operation (From December 2015 to December 2016; No data were collected on May, June, August, September and October 2016 and the instrument was powered off, leaving only 8 months of useful data) were used as input data for the correlation analysis. For all three correlation analyses, input data were divided into training data (odd number data points) and evaluation data (even number data points). Data quality control was performed to remove 1) data with hourly RH greater than 95% (as suggested by the manufacture), 2) hourly BAM-1020 PM2.5 greater than 800 μg m^{-3}, 3) hours with less than 20 valid 2-minute low-cost instrument data. In addition, data points with LOR1/LOR2 > 1000 or < 1, or with LOR1/BAM-1020 > 0.5 or LOR1/BAM-1020 < 0.005 were considered as outliers and were excluded from analysis. Overall, a total number of 3610 valid hourly data were used in the analysis.

Environmental factors were also investigated to check if low-cost sensor

readings were affected by different relative humidity and temperature combinations. Last but not least, by considering all datasets in Nanjing, sensor deterioration was plotted as well.

2.3 Results and Discussion

2.3.1 Training Data in Nanjing using linear and non-linear regression

Figure 2-2(a) shows the correlation of hourly averaged raw low-cost sensor signal – Low Pulse Occupancy Channel 1 ratio (LOR1) with the reference BAM-1020 PM2.5 using the odd data points (training data) from the first eight months. Higher PM2.5 concentrations are generally higher in the first few months, which is also clear from the time series plot of BAM-1020 data in Figure 2-2(b). The analysis shows a moderate correlation between LOR1 and the BAM-1020 PM2.5 with a coefficient determination R^2=0.6847. The maximum LOR1 reading goes to ~20% with a corresponding BAM-1020 PM2.5 of ~350 μg m^{-3}.

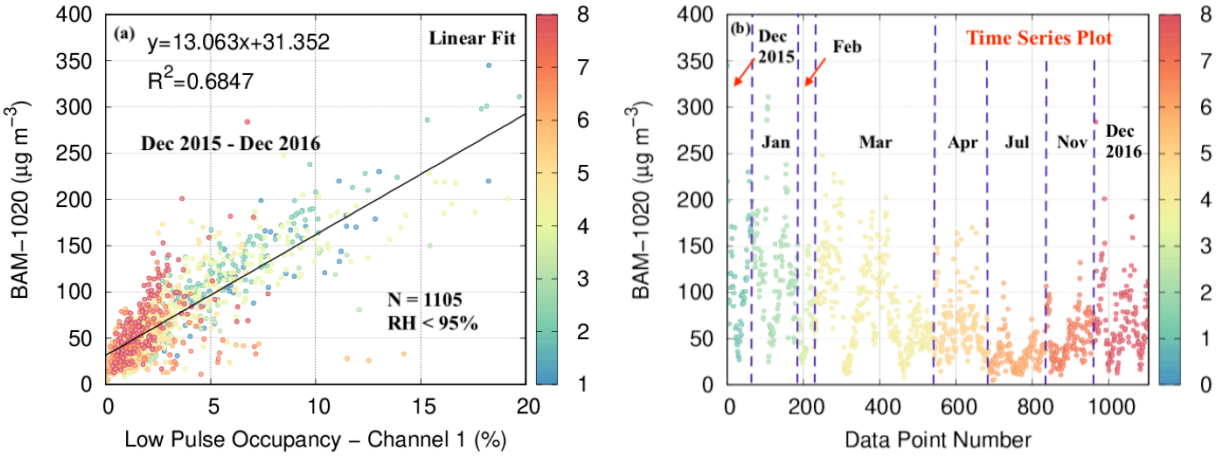

Figure 2-2. (a) Linear regression of LOR1 and BAM-1020 for low-cost monitor A001 on NUIST campus in Nanjing, China. Half of the valid hourly data between December 2015 and December 2016 (odd number data points; N=1105) were used in this regression analysis. The data points were colored to show the number of months the monitor has been in operation. Note that no data were collected in May, June, August, September and October 2016. Since the monitor was not running during these months, these missing months were not counted towards the total operation time. (b) Time series of the first year BAM-1020 PM2.5 concentrations for the odd number data used in panel (a).

Figure 2-3 shows the power law fit of LOR1 with BAM-1020 PM2.5 using equation (2). Simply looking at the coefficient determination R^2=0.6290 leaves an impression that this non-linear regression does not provide a better fit of the data. However, to evaluate if the power law fitting provides a better estimation of PM2.5 concentrations, it is necessary to compare the estimated PM2.5 with BAM-1020 PM2.5. The non-linear transformation of the LOR1 to estimated PM2.5 is expected to generate a different correlation coefficient. Using the linear equation (1) to estimate PM2.5 from LOR1, however, is going to yield a correlation coefficient with BAM-1020 PM2.5 the same as shown in Figure 2-2.

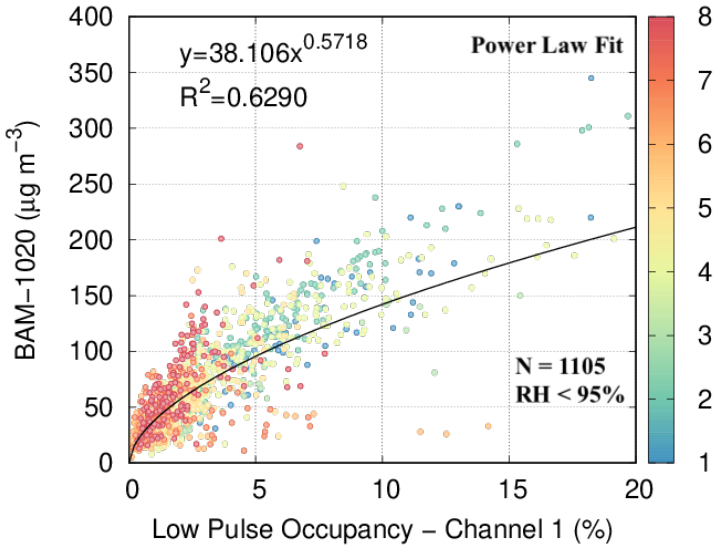

Figure 2-3. Same as Figure 2-2, using non-linear regression for power law equation (2).

As it shows in Figure 2-4 (b), the estimated PM2.5 from LOR1 and the power-law equation yields the coefficient determinant R^2=0.7016, which represents a slight improvement from the linear estimation (R^2=0.6847). The slope of the regression is very close to unity, suggesting that there is little bias in estimating ambient PM2.5 using the low-cost monitor LOR1 reading and both of the linear and power law equation. The number of data points that fall into the ±20 µg m^{-3} and ±20% area was also computed, as it indicated in Table 2-1. It is clear that higher percentage of data points with less than 100 µg m^{-3} will fall into the area than the data points with higher PM concentration (66.3% versus 57.2% for linear regression and 69.4% versus 60.9% for power law regression). More data points for power law estimations fall into the confined area than linear estimations (67.3% over 64.4%), which demonstrated a better data fitting ability than linear fit method.

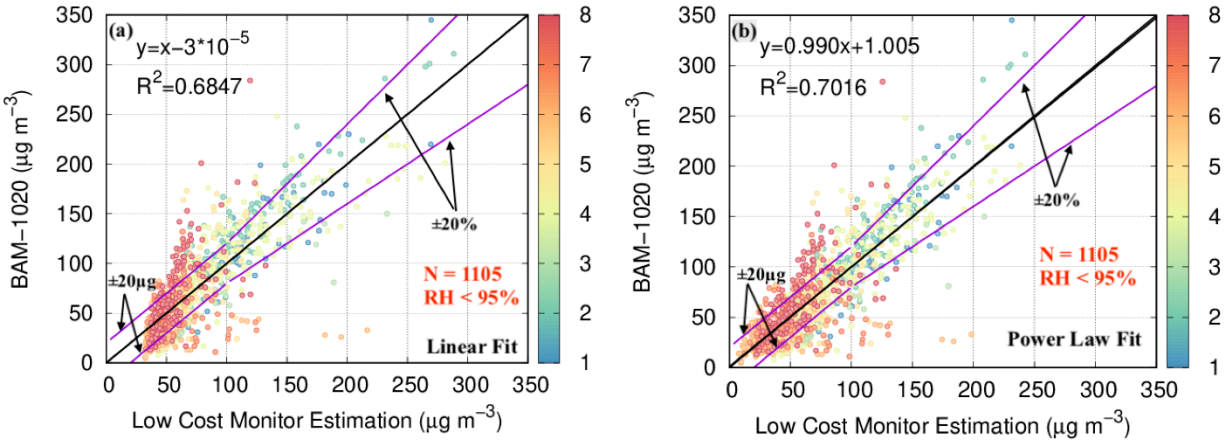

Figure 2-4. Linear correlation between low-cost monitor estimated PM2.5 in Nanjing using the linear & power law equation (see Figure 2-3) and BAM-1020 PM2.5 for training data set. The two purple lines show the required low-cost monitor performance for gridded community PM2.5 monitoring by the Ministry of Environment of China (Agency, 2007): an absolute error within ±20 μg/m^3 when the reference instrument reading is below 100 μg/m^3 and a relative error within ±20% when the reference instrument reading is above 100 μg/m^3. The data points were colored to show the number of months the monitor has been in operation.

	Number of Data Points With Less Than 100 μg m-3		Number of Data Points with Higher Than 100μg m-3		Total Percentage
Linear Estimation					
Within	581	66.3%	131	57.2%	64.4%
Not Within	295	33.7%	98	42.8%	35.6%
Total Number	876		229		1105
Power Law Estimation					
Within	579	69.4%	165	60.9%	67.3%
Not Within	255	30.6%	106	39.1%	32.7%
Total Number	834		271		1105

Table 2-1. Comparison of the number and percentage of data points with linear or power law estimated PM concentrations that either less or larger than 100μg m^{-3}. Same method was applied to calculate the number and percentage of data points that fall into the enclosed area of ±20μg m^{-3} and ±20% of the linear and power law estimations.

Figure 2-5 shows the normalized bias, MNB and MNE for both linear and power law fitting method in order to estimate the overall deviation between the predicted and measured values. It is clear that the normalized bias for low PM concentrations are significantly higher than the higher PM concentrations. Power law estimations also demonstrate better results for both MNE and MNB than linear estimations, as it indicates in Figure 2-5 (a) and (b). For linear regression method, there are 509 out of 1105 data points within ±20% of the normalized bias area, whereas 521 out of 1105 data points for power law regression method.

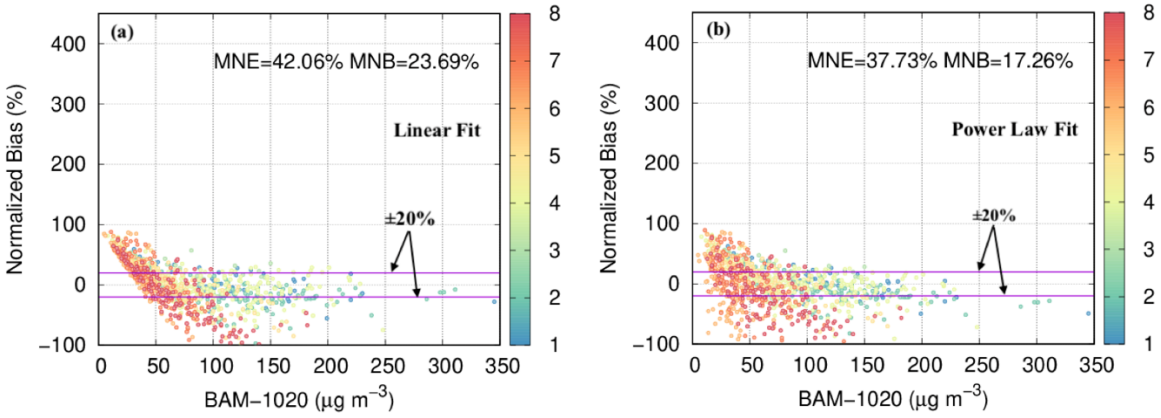

Figure 2-5 Normalized bias between the low-cost monitor PM2.5 and the BAM-1020 PM2.5 (PM2.5 low-cost monitor – PM2.5 BAM-1020) as a function of the BAM-1020 PM2.5. The data points were colored to show the number of months the monitor has been in operation. MNE: mean normalized error; MNB: mean normalized bias.

Figure 2-6 shows the correlation between ANN estimated PM2.5 and the BAM-1020 PM2.5 using different number of neurons in the hidden layer. As the number of neurons increases from 10 to 20, the coefficient determinant R^2 increases from 0.833 to

0.865. Adding more neurons in the hidden layer leads to decreases in R2. Thus, the subsequent analyses are based on the results from the 20-neuron setup.

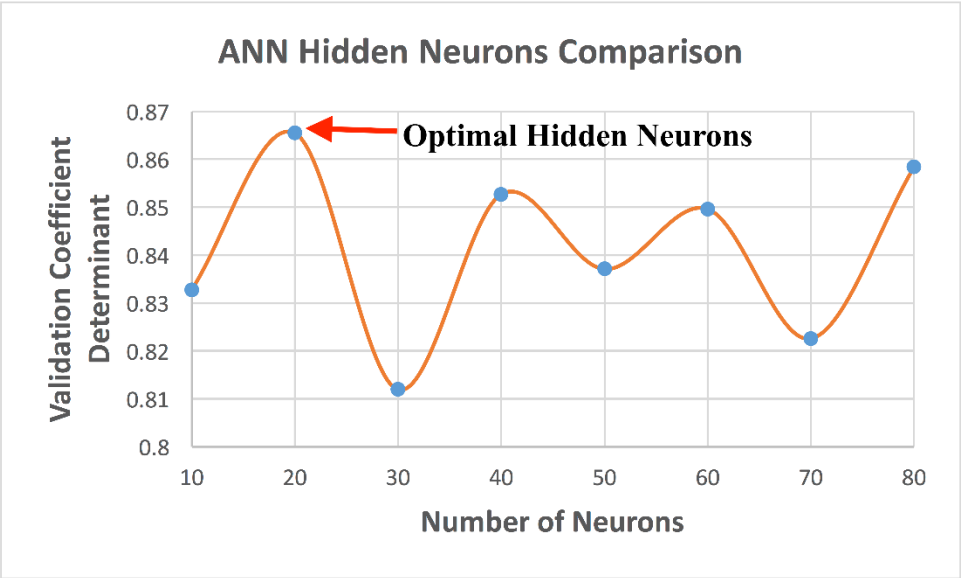

Figure 2-6. Optimal nodes in hidden layer to avoid both over fitting and under fitting.

Figure 2-7 shows the correlation of neural network estimation with BAM-1020 reference readings. Neural network shows the highest agreement (R^2=0.7597) compared with linear fit and power law fit. Most of the data can be observed within the area enclosed by the two proposed low-cost monitor performance guidelines. For PM concentration less than 100 µg m^{-3}, 600 out of 823 (72.9%) data points are within the purple line enclosed area. 210 out of 282 (74.5%) data points fall into the higher concentration area. The overall percentage of data points falling into the enclosed area is 73.3% for neural network, which is significantly higher than linear and power law regression (64.4% and 67.3%, respectively). Also, neural network shows better prediction results for higher PM concentrations rather than lower concentrations

(74.5% compares to 72.9%). However, linear and power law regression show a reverse trend, as it mentioned in Table 2-1.

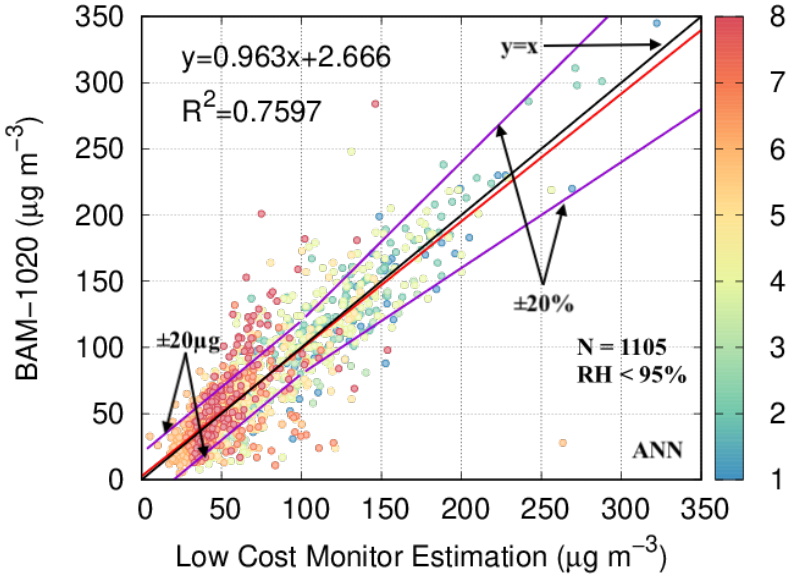

Figure 2-7. Same as Figure 2-4. Linear correlation between low-cost monitor estimated PM2.5 in Nanjing using artificial neural network and BAM-1020 PM2.5 for training data set.

Normalized bias and normalized error for artificial neural network method are calculated and shown in Figure 2-8. The normalized error decreases with the increase of PM2.5 concentrations. Compared to linear and power law fit, as it demonstrated in Figure 2-5, the neural network approach leads to a better estimation of PM2.5 based on the statistical measures of MNE (34.82% to 42.06% and 37.73%) and MNB (16.18% to 23.69% and 17.26%).

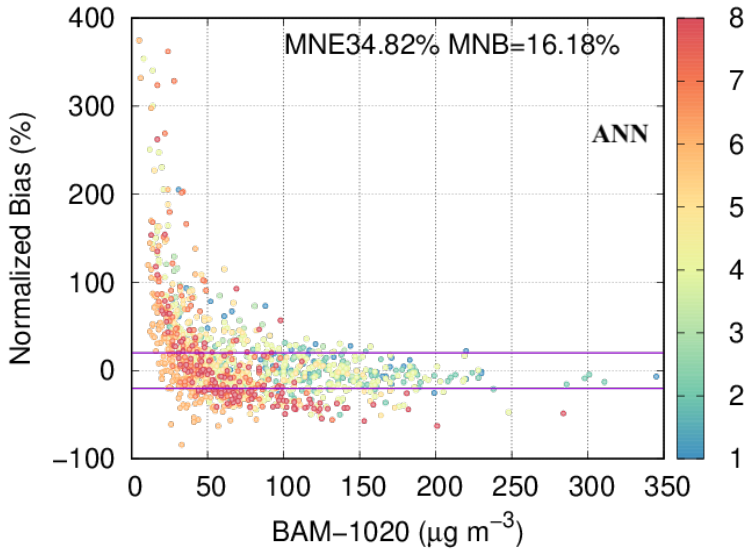

Figure 2-8. Same as Figure 2-5. Normalized bias between the low-cost monitor PM2.5 and the BAM-1020 PM2.5 (PM2.5 low-cost monitor – PM2.5 BAM-1020) as a function of the BAM-1020 PM2.5.

2.3.2 Data Evaluation

To evaluate if the parameters for linear, power-law and ANN approaches using the training data can be applied to reliably estimate PM2.5 concentrations for other data, the odd number of data points from December 2015 to December 2016 generated from low-cost monitor A001 were used as inputs. The estimated hourly PM2.5 concentrations were compared with BAM-1020 and shown in Figure 2-9. The training process is successful as the performance of the three approaches with the evaluation data set is similar to what they have achieved with the training data set. The ANN, as it is able to consider the influence of temperature and RH on sensor readings, still shows the best performance with R^2=0.7565 and 71.5% of the estimated PM2.5 concentrations are within the low-cost monitor operation guidelines. For linear and

power law methods, the percentage of data points fall into the confined area are 65.2% and 67.9%. Further, all three approaches demonstrate that lower estimated PM2.5 concentrations have higher chance to locate in the guideline area than higher concentrations. The percentages of data points that are within the low and high PM2.5 concentrations confined area are (66.25%, 61.16%), (70.64%, 59.7%), (74.2%, 64.4%) for linear, power law and ANN approaches. It also performs best in estimating high PM2.5 concentrations (> 200 $\mu g\ m^{-3}$) than the other two approaches. The MNB and MNE for high PM2.5 concentrations are (23.28%, 41.71%), (16.93%, 37.51%) and (16.23%, 36.45%) for linear, power-law and ANN approaches respectively. While the ANN approach shows best performance among the three, the power-law and linear equations are relatively easy to implement and could provide acceptable estimations as well after calibrating with a reference instrument.

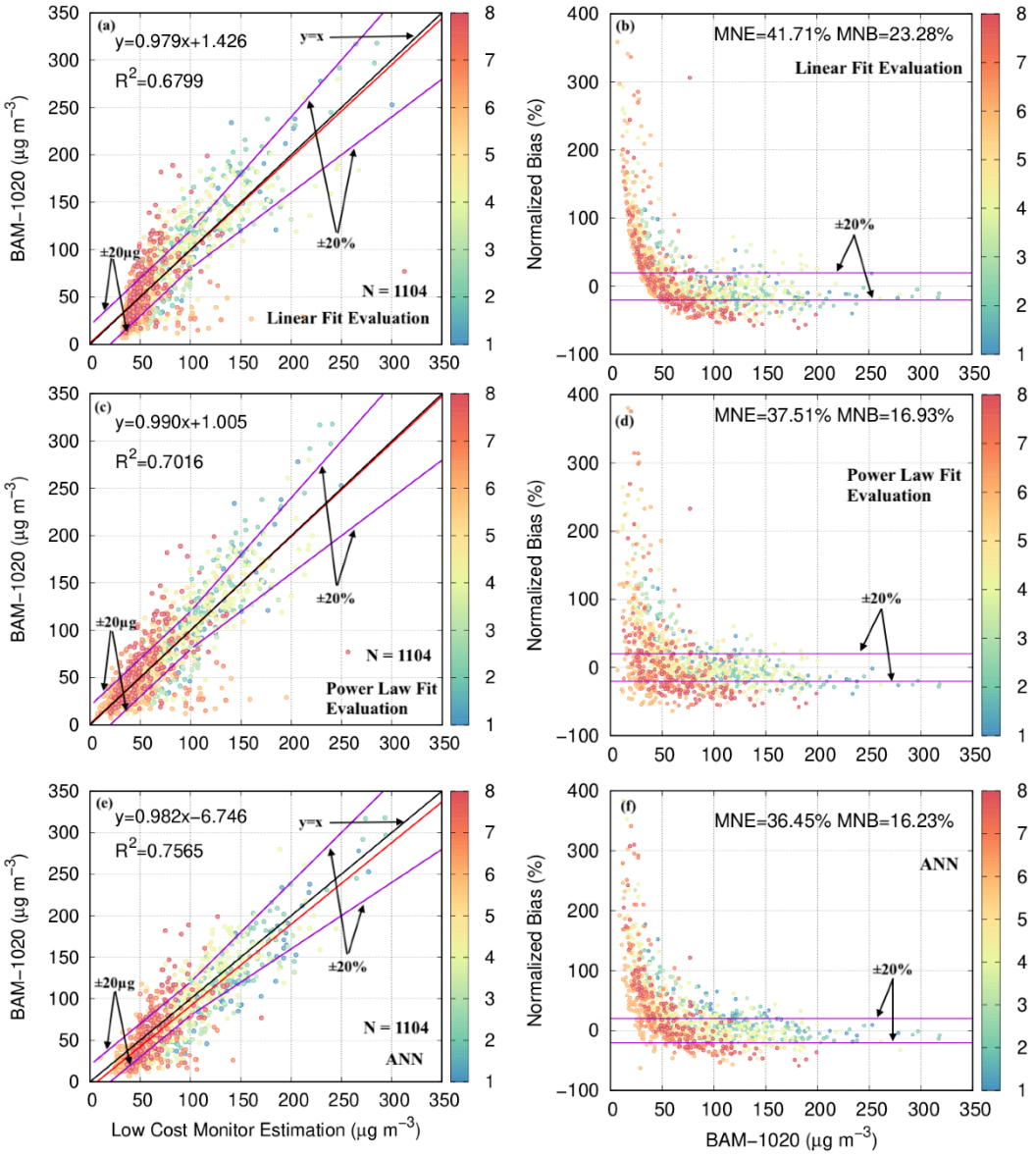

Figure 2-9. Evaluate the linear (a,b), power law (c,d) and neural network (e,f) approaches in estimating PM2.5 concentrations. The even number data points for the first eight months between December 2015 and December 2016 were used as input data. The parameters were derived using odd number data points. Panels (a, c, e) show the correlation between estimated and BAM-1020 PM2.5, and panels (b, d, f) show normalized bias, mean fractional bias and fractional error for the three approaches.

2.3.3 Sensor Deterioration Test

From the previous analyses, it is noticed that the agreement of PM2.5 between the low-cost monitor estimations and the BAM-1020 measurements in the first few months of operation are noticeably better than later months. The accuracy of the low-cost monitor appears to deteriorate in later months. This deterioration might happen due to aging of the electric components and/or accumulation of dust on the surface of the optical components as no maintenance was done during the operation of the low-cost monitor. In order to further confirm this, all input data collected by A001 in Nanjing from December 2015 to July 2017 are used as inputs to the ANN to estimate PM2.5 and results are compared with BAM-1020 measurements. Figure 2-10 shows estimation of all data in Nanjing generated by artificial neural network plot against the reference instrument readings. Even after considering all data points from December 2015 to June 2017, the sensor still demonstrated a moderate performance. By comparing Figure 2-10 (a) through (d), we can clearly observe a more dispersal trend for the sensor performance and significant dropping of R^2 as time goes by. The same conclusion can also be drawn from Figure 2-11. The PPD42NS sensor can maintain a good performance within the first 4-5 months' usage. But after operating for several months, it started to get larger normalized bias. It is also demonstrated in Table 2-2 that during the first four month of operation, the sensor can still generate good results with low MNE, MNB and high R^2. The overall agreement of monthly average of ANN estimations and the BAM-1020 measurements are acceptable.

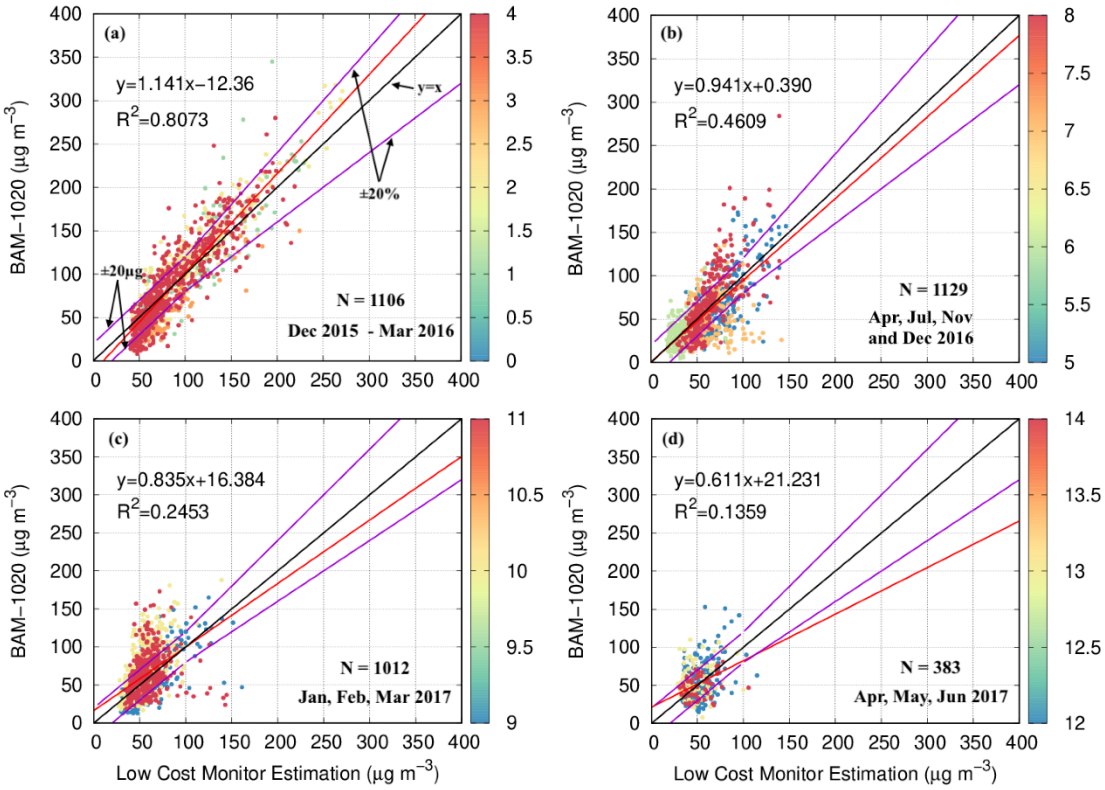

Figure 2-10. All 14 months of Nanjing data from December 2015 to June 2017 were broken down into four periods to check the sensor's overall performance and potential deterioration. Figure 2-10 (a) covers from December 2015 to March 2016, (b) covers April, July, November and December of 2016, (c) covers January through March of 2017, (d) covers the last three months, which are April, May and June of 2017. The total number of data points with relative humidity less than 95% was 3630.

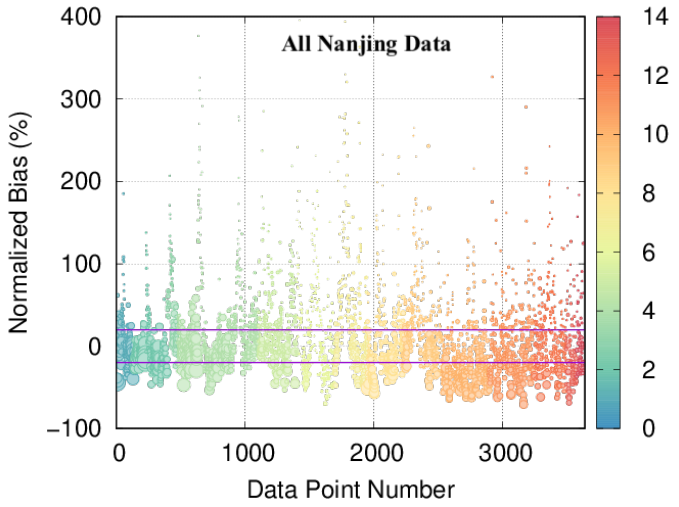

Figure 2-11. The normalized bias for all Nanjing data were calculated and plotted for the range of -100% to 400%. Different sizes of the circles represent the $PM_{2.5}$ concentrations from reference reading. The sensor had higher normalized bias after 6 months of usage.

Month	BAM Avg. ($\mu g/m^3$)	ANN Avg. ($\mu g/m^3$)	MNB (%)	MNE (%)	RMSE ($\mu g/m^3$)	R^2
Dec 2015	117	115	7.29	21.52	28.53	0.77
Jan 2016	127	114	-5.32	18.38	24.53	0.87
Feb 2016	69	94	55.11	56.92	30.98	0.78
Mar 2016	87	87	18.64	33.04	22.94	0.79
Apr 2016	65	71	26.64	38.70	23.96	0.52
Jul 2016	32	33	26.65	50.69	15.67	0.15
Nov 2016	51	63	55.76	68.62	29.69	0.05
Dec 2016	70	64	10.92	35.00	28.08	0.57
Jan 2017	52	58	28.89	38.16	19.54	0.60
Feb 2017	75	56	-12.81	32.72	31.94	0.32
Mar 2017	65	59	6.27	34.83	29.17	0.05
Apr 2017	52	51	14.68	36.40	21.78	0.21
May 2017	57	49	5.89	43.69	29.27	0.05
Jun 2017	48	52	16.27	34.79	17.90	0.07

Table 2-2. Comparison of BAM-1020 monthly average, ANN monthly average, MNB, MNE, RMSE and R^2 for each month of Nanjing data.

2.3.4 Chengdu Data

Figure 2-12 compares the estimated PM2.5 with BAM-1020 measurements in Chengdu by using different parameters from data training in Nanjing (linear, power law regression and artificial neural network.). From Figure 2-12 (a), (c) and (e), data points are more dispersed than Figure 2-9 (a), (c) and (e). The low R^2 values (R^2 = 0.382, 0.415 and 0.440 for linear, power law and ANN respectively) indicate that the data training parameters from Nanjing do not fit well in Chengdu. Figure 2-12 (b), (d), (f) show a much higher MNB, MNE in all three methodologies compared to Nanjing. This also proves that parameters from Nanjing cannot be directly used to predict PM2.5 concentration levels.

Therefore, the validity of Chengdu data is tested in this study. The analysis processes are essentially the same as the processes in Nanjing. By using the linear, power law fitting and ANN equations in section 2.2.3, new parameters are generated and compared to the parameters in Nanjing, as it shows in Table 2-3. For the same data training method, the parameters in these two cities are different. Based on the linear and power law regression parameters (slope, intercept and exponent), both the linear and power law equation for Nanjing data are closer to y=x compared to Chengdu, which means the data are better correlated in Nanjing. R^2 from the same table can further prove that the sensor in Nanjing performs better than that in Chengdu.

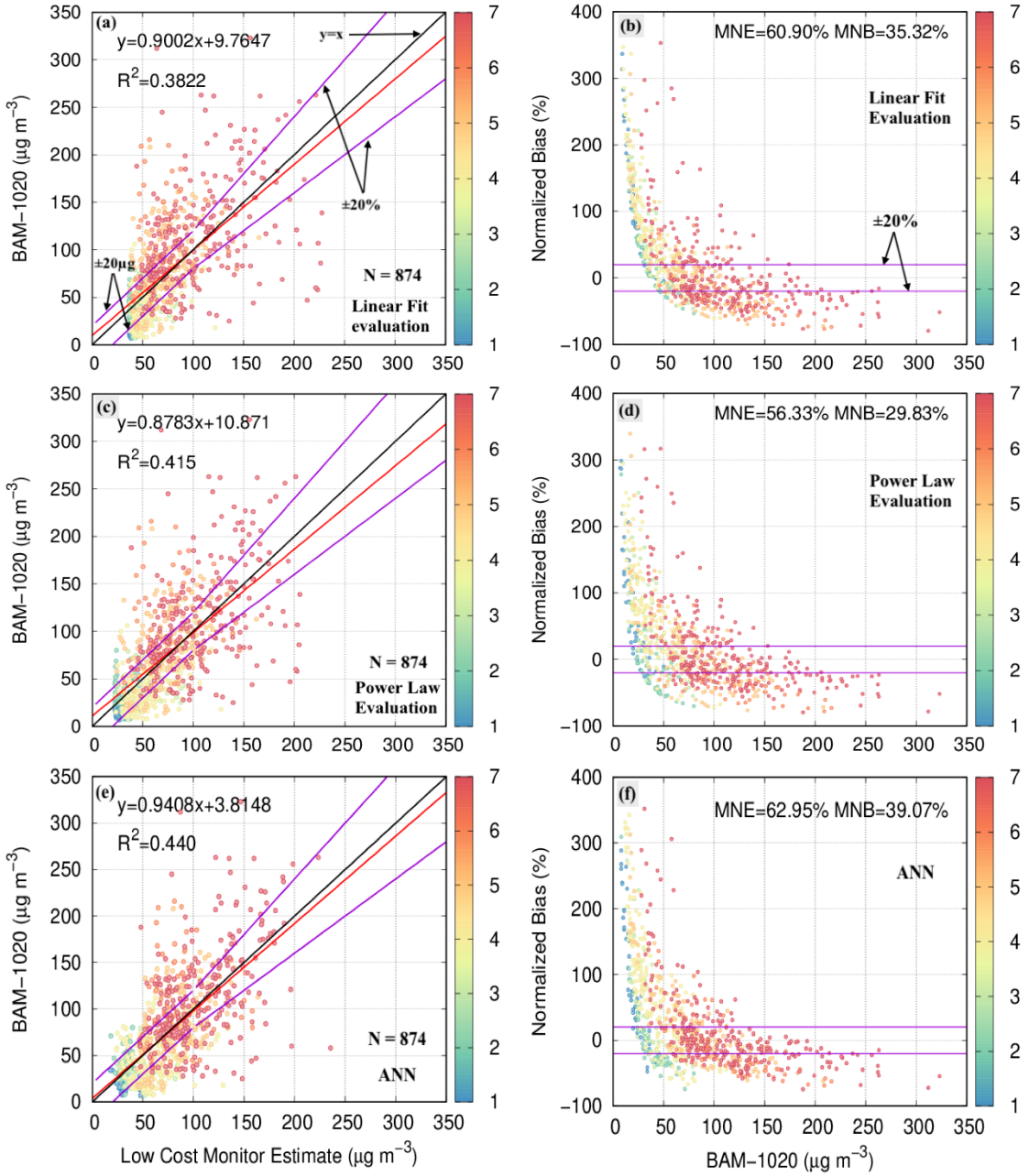

Figure 2-12. Evaluate the data training parameters from the linear (a,b), power law (c,d) and neural network (e,f) approaches in Nanjing on estimating PM2.5 concentrations in Chengdu. The all Chengdu data between July 2016 and January 2017 were used as input data. Panels (a, c, e) show the correlation between estimated and BAM-1020 PM2.5, and panels (b, d, f) show normalized bias, mean fractional bias and fractional error for the three approaches.

	Equations	Nanjing			Chengdu		
		a	b	R^2	a	b	R^2
Linear Fit	$C_{PM2.5}^{BAM} = aLOR1 + b$	13.0	31.3	0.68	11.7	37.9	0.38
Power Law Fit	$C_{PM2.5}^{BAM} = a(LOR1)^b$	38.9	0.61	0.70	45.1	0.52	0.42
ANN	$C_{PM2.5}^{BAM} = f(Raw\ Readings, T, RH, \dots)$			0.76			0.69

Table 2-3. Comparison of data training parameters from different methodologies in Nanjing and Chengdu. R^2 values are used to test which monitoring site gives better estimations.

Meteorology conditions such as different temperature and humidity patterns in these two cities may potentially affect the performance of the low-cost sensor. As it indicated in Figure 2-13, even though the general pattern of relative humidity and temperature are similar, the year-round lowest temperature in Nanjing is about 6-7 °C colder than Chengdu. Before analyzing meteorology inputs, data screening were first conducted following the same rule in Section 2.2.3. More than 60% of Chengdu data were excluded from this study, whereas only less than 50% of Nanjing data were excluded.

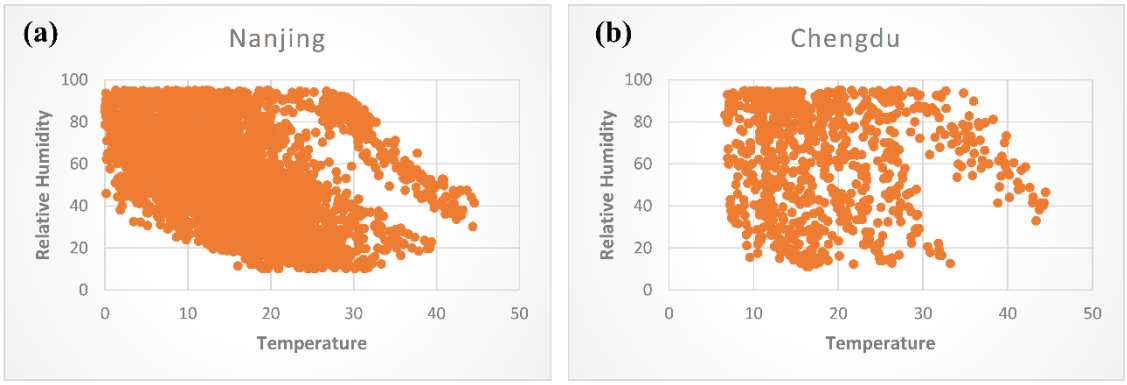

Figure 2-13. Temperature and relative humidity pattern comparison in Nanjing and Chengdu.

Table 2-4 and 2-5 showed the comparison of relative humidity and temperature in Nanjing and Chengdu. Higher humidity in Chengdu can be observed by comparing the median and mean values of relative humidity. Standard error mean was found lower in Nanjing because of larger sampling size. Higher temperature in Chengdu can also be observed.

		Nanjing	Chengdu
100%	Maximum	95	95
75.0%	Quartile	73	84
50.0%	Median	52	64
25.0%	Quartile	32	38
0.0%	Minimum	1	1
	Mean	52.3	59.6
	Std Dev	24.6	26.6
	Std Err Mean	0.41	0.90
	Number of Points	3630	874

Table 2-4. Summary statistics for relative humidity in Nanjing and Chengdu

		Nanjing	Chengdu
100%	Maximum	52	48
75.0%	Quartile	23	24
50.0%	Median	16	17
25.0%	Quartile	10	13
0.0%	Minimum	1	7
	Mean	16.8	19.2
	Std Dev	9.4	8.8
	Std Err Mean	0.15	0.30
	Number of Points	3630	874

Table 2-5. Summary statistics for temperature in both Nanjing and Chengdu

The different sensor performances in the two cities may also attributed by the variability the sensor itself. As it mentioned, three low-cost monitors (A001, A002 and A003) with the same sensors were deployed side by side on NUIST campus. Inter-comparison among all low-cost sensors were conducted in this study. Because the operation period for the sensors are different, time-match all the screened data are first approached. As it shows in Figure 2-14 (a), (b), (c), the LOR1 for each monitor is not correlated each other (the highest R^2 is only 0.037), which indicates that the reliability of the low-cost sensors is not sufficient enough to generate highly trustworthy results. When compares the temperature measurements with each of the low-cost monitor, all sensors show a great agreement (the lowest R^2 is 0.96), which demonstrates an excellent consistency of the temperature sensors, as it shows in panel (d), (e) and (f). Relative humidity measurements are compared as well (panel (g), (h) and (i)). Both A001 and A002 show a high to moderate correlation with A003 (R^2=0.82 and 0.55, respectively), whereas, A001 and A002 show a poor agreement with each other (R^2=0.20).

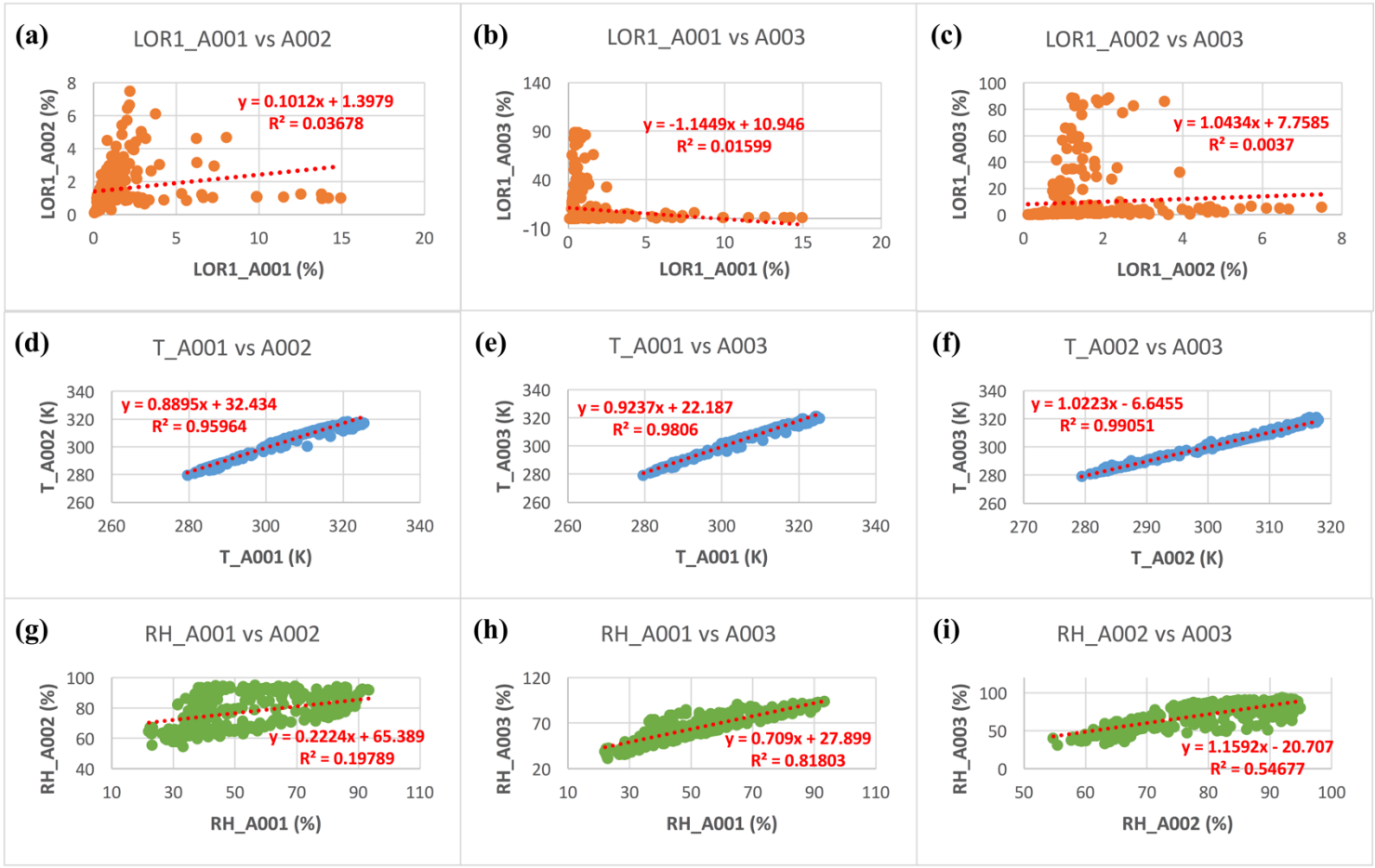

Figure 2-14. Inter-comparison of A001, A002, A003. Panel (a), (b) and (c) show the correlation of sensor raw results (LOR1) with each other (A001 vs. A002, A001 vs. A003 and A002 vs. A003). Panel (d), (e) and (f) compare each of the low-cost monitor's temperature readings (the comparison sequence is the same as (a), (b) and (c)). RH readings are compared as well, as it shows in panel (g), (h) and (i). The number of points for each monitor is 314.

2.3.5 Influence of input parameters on performance of ANN

As it indicated in table 2-6, the number of neural network input parameters were tested to find out with how many parameters neural network can generate the best results. All five parameters are Hour, Relative Humidity, Temperature, LOR1 and LOR2. The result shows that with three input parameters (Relative Humidity, Temperature and LOR1), the neural network generates the best estimation with the highest overall and training correlation coefficient. The Root Mean Standard Error (RMSE) and Mean Normalized Error were also the lowest. Furthermore, the neural network did not perform very well after eliminating any of the relative humidity and temperature data.

Number of Input Parameters	R_train	R_all	RMSE (μg/m3)	MNE (%)	MNB (%)
5	0.88	0.87	24.66	0.35	0.15
4 (w/o HR)	0.87	0.87	24.90	0.34	0.13
3 (w/o LOR2)	0.88	0.87	24.48	0.34	0.14
2 (w/o Temp.)	0.85	0.84	26.97	0.38	0.19
2 (w/o RH)	0.86	0.85	26.22	0.37	0.18
1 (only LOR1)	0.84	0.84	26.90	0.40	0.22

Table 2-6. Influence of input parameters on the ability of the neural network in estimating ambient PM2.5. The initial input data include five parameters, LOR1, LOR2, temperature, relative humidity (RH) and hour of the day. In each of the sensitivity simulations, one additional parameters were removed.

2.4 Conclusion

This study demonstrated that the low-cost PM sensor could be potentially used to improve existing PM2.5 sampling networks and as an affordable technology to enhance spatiotemporal resolution of PM2.5 datasets in both ambient monitoring networks and highly polluted areas. However, the sensor needs to be regularly calibrated with co-located reference instrument, such as BAM-1020.

Among all three different calibration methods used in this study (linear, power-law and ANN), ANN approach demonstrates the highest correlation between the sensor estimations and BAM-1020 hourly PM2.5 measurements (R^2=0.76). PM2.5 estimations from power-law equation shows a slightly better correlation than linear fit method (R^2=0.70 and 0.68 respectively). Further, approximately 73% of ANN estimations are located within the Ministry of Environment of China low-cost monitor performance guidelines, which is better than linear and power-law methods (~64% and ~67%, respectively). The main reason why ANN performs better than the other two approaches is that it not only takes the raw sensor output LOR1 as input training data, but also takes environmental and other factors into consideration as well. For instance, temperature, relative humidity and LOR2, etc.

The sensor deterioration was found after 5-6 months of continuous operation. All Nanjing data are separated into 4 periods with the first two periods have 4 months' data and the last two have 3-month data. The R^2 value for each of the four periods are 0.80, 0.64, 0.24 and 0.13, respectively. Therefore, for long-term use the low-cost sensor, regular replacement or cleaning of the PM optical sensor is needed. But on the other hand,

compare to the BAM measurements, the monthly average PM concentrations of the sensor bas a small error of 10%.

The inter-comparison of 3 co-located low-cost sensors were conducted in the study as well. However, poor correlations of ANN estimations for each sensor were found, which demonstrates that one sensor cannot be used to predict another, regards to provide consistent results.

Last but not the least, the calibration parameters developed by using low-cost monitor data in Nanjing do not fit well for the low-cost monitor in Chengdu. In addition to the sensor variation and environmental conditions, some further investigations need to be conducted to find the cause of this phenomenon.

There are some limitations in this study. More works are required to better understand the sensor's technological limitations and under what environmental conditions these sensors could be used. In addition, this study did not examine the potential effect of particles' chemical and optical compositions, seasonal variations on the low-cost monitor detection and calibration.

REFERENCES

Abraham, S., & Li, X. R. (2016). Design of A Low-Cost Wireless Indoor Air Quality Sensor Network System. *International Journal of Wireless Information Networks, 23*(1), 57-65. doi:10.1007/s10776-016-0299-y

Agency, U. E. P. (2007). Ambient Air Monitoring Network Assessment

Amaral, S. S., de Carvalho, J. A., Costa, M. A. M., & Pinheiro, C. (2015). An Overview of Particulate Matter Measurement Instruments. *Atmosphere, 6*(9), 1327-1345. doi:10.3390/atmos6091327

Austin, E., Novosselov, I., Seto, E., & Yost, M. G. (2015). Laboratory Evaluation of the Shinyei PPD42NS Low-Cost Particulate Matter Sensor (vol 10, e0137789, 2015). *Plos One, 10*(10). doi:ARTN e014192810.1371/journal.pone.0141928

Bart, M., Williams, D. E., Ainslie, B., McKendry, I., Salmond, J., Grange, S. K., . . . Henshaw, G. S. (2014). High Density Ozone Monitoring Using Gas Sensitive Semi-Conductor Sensors in the Lower Fraser Valley, British Columbia. *Environmental Science & Technology, 48*(7), 3970-3977. doi:10.1021/es404610t

Chen, R. J., Yin, P., Meng, X., Liu, C., Wang, L. J., Xu, X. H., . . . Zhou, M. G. (2017). Fine Particulate Air Pollution and Daily Mortality A Nationwide Analysis in 272 Chinese Cities. *American Journal of Respiratory and Critical Care Medicine, 196*(1), 73-81. doi:10.1164/rccm.201609-1862OC

Dormehl, L. (2017). What is an artificial neural network? Here's everything you need to know.

Gao, M. L., Cao, J. J., & Seto, E. (2015). A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. *Environmental Pollution, 199*, 56-65. doi:10.1016/j.envpol.2015.01.013

Grove. (2015). Grove - Dust Sensor User Manual. https://www.mouser.com/ds/2/744/Seeed_101020012-1217636.pdf

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks, 5*(6), 989-993. doi:10.1109/72.329697

Hankey, S., Marshall, J. D., & Brauer, M. (2012). Health impacts of the built environment: within-urban variability in physical inactivity, air pollution, and

ischemic heart disease mortality. *Environ Health Perspect, 120*(2), 247-253. doi:10.1289/ehp.1103806

Hansen, L. K., & Salamon, P. (1990). Neural Network Ensembles. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 12*(10), 993-1001. doi:Doi 10.1109/34.58871

Hojaiji, H., Kalantarian, H., Bui, A. A. T., King, C. E., & Sarrafzadeh, M. (2017). Temperature and Humidity Calibration of a Low-Cost Wireless Dust Sensor for Real-Time Monitoring. *2017 Ieee Sensors Applications Symposium (Sas)*.

Holstius, D. M., Pillarisetti, A., Smith, K. R., & Seto, E. (2014). Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California. *Atmospheric Measurement Techniques, 7*(4), 1121-1131. doi:10.5194/amt-7-1121-2014

Jiao, W., Hagler, G. S. W., Williams, R. W., Sharpe, R. N., Weinstock, L., & Rice, J. (2015). Field Assessment of the Village Green Project: An Autonomous Community Air Quality Monitoring System. *Environmental Science & Technology, 49*(10), 6085-6092. doi:10.1021/acs.est.5b01245

Kelly, J., & Sukhatme, G. S. (2017). Visual-Inertial Sensor Fusion: Localization, Mapping and Sensor-to-Sensor Self-calibration (vol 30, pg 56, 2011). *International Journal of Robotics Research, 36*(13-14), 1619-1619. doi:10.1177/0278364910382802

Klepeis, N., Hughes, S., Edwards, R., Allen, T., Johnson, M., Chowdhury, Z., . . . Hovell, M. (2013). Using Real-Time Monitor Feedback to Reduce Particle Levels in Smoking Homes. *Annals of Behavioral Medicine, 45*, S49-S49.

Li, J. Y., & Biswas, P. (2017). Optical Characterization Studies of a Low-Cost Particle Sensor. *Aerosol and Air Quality Research, 17*(7), 1691-1704. doi:10.4209/aaqr.2017.02.0085

Lightstone, S. D., Moshary, F., & Gross, B. (2017). Comparing CMAQ Forecasts with a Neural Network Forecast Model for PM2.5 in New York. *Atmosphere, 8*(9). doi:ARTN 16110.3390/atmos8090161

Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., . . . Memish, Z. A. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet, 380*(9859), 2224-2260. doi:10.1016/S0140-6736(12)61766-8

Liu, D., Zhang, Q., Jiang, J. K., & Chen, D. R. (2017). Performance calibration of low-cost and portable particular matter (PM) sensors. *Journal of Aerosol Science, 112*, 1-10. doi:10.1016/j.jaerosci.2017.05.011

Manikonda, A., Zikova, N., Hopke, P. K., & Ferro, A. R. (2016). Laboratory assessment of low-cost PM monitors. *Journal of Aerosol Science, 102*, 29-40. doi:10.1016/j.jaerosci.2016.08.010

Michael. (2013). Make Your Own AirCasting Particle Monitor. Retrieved from http://www.takingspace.org/make-your-own-aircasting-particle-monitor/

mybotic. (2016). HOW TO INTERFACE WITH OPTICAL DUST SENSOR.

Ozkan, O. (2010). Neural Network Forecast for Daily Average PM10 Concentrations. *Asian Journal of Chemistry, 22*(1), 582-588.

Park, S., Kim, M., Kim, M., Namgung, H. G., Kim, K. T., Cho, K. H., & Kwon, S. B. (2018). Predicting PK10 concentration in Seoul metropolitan subway stations using artificial neural network (ANN). *Journal of Hazardous Materials, 341*, 75-82. doi:10.1016/j.jhazmat.2017.07.050

Qiao, X., Jaffe, D., Tang, Y., Bresnahan, M., & Song, J. (2015). Evaluation of air quality in Chengdu, Sichuan Basin, China: are China's air quality standards sufficient yet? *Environmental Monitoring and Assessment, 187*(5). doi:ARTN 250 10.1007/s10661-015-4500-z

Rai, A. C., Kumar, P., Pilla, F., Skouloudis, A. N., Di Sabatino, S., Ratti, C., . . . Rickerby, D. (2017). End-user perspective of low-cost sensors for outdoor air pollution monitoring. *Science of the Total Environment, 607*, 691-705. doi:10.1016/j.scitotenv.2017.06.266

Rohde, R. A., & Muller, R. A. (2015). Air Pollution in China: Mapping of Concentrations and Sources. *Plos One, 10*(8). doi:ARTN e0135749 10.1371/journal.pone.0135749

Sioutas, C., Abt, E., Wolfson, J. M., & Koutrakis, P. (1999). Evaluation of the measurement performance of the scanning mobility particle sizer and aerodynamic particle sizer. *Aerosol Science and Technology, 30*(1), 84-92. doi:Doi 10.1080/027868299304903

Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W., . . . Preuss, P. W. (2013). The Changing Paradigm of Air Pollution Monitoring. *Environmental Science & Technology, 47*(20), 11369-11377. doi:10.1021/es4022602

Sousan, S., Koehler, K., Thomas, G., Park, J. H., Hillman, M., Halterman, A., & Peters, T. M. (2016). Inter-comparison of low-cost sensors for measuring the mass concentration of occupational aerosols. *Aerosol Science and Technology, 50*(5), 462-473. doi:10.1080/02786826.2016.1162901

Taylor, M. D. (2016). Low-cost air quality monitors Modeling and characterization of sensor drift in optical particle counters. *2016 Ieee Sensors*.

Wang, Y., Li, J. Y., Jing, H., Zhang, Q., Jiang, J. K., & Biswas, P. (2015). Laboratory Evaluation and Calibration of Three Low- Cost Particle Sensors for Particulate Matter Measurement. *Aerosol Science and Technology, 49*(11), 1063-1077. doi:10.1080/02786826.2015.1100710

WHO. (2016). Ambient (outdoor) air quality and health. Retrieved from http://www.who.int/mediacentre/factsheets/fs313/en/

Xu, E. A. a. B. (2016). China's Environmental Crisis. Retrieved from https://www.cfr.org/backgrounder/chinas-environmental-crisis

Zu, Y. Q., Huang, L., Hu, J. L., Zhao, Z., Liu, H., Zhang, H. L., . . . Chen, M. D. (2017). Investigation of relationships between meteorological conditions and high PM10 pollution in a megacity in the western Yangtze River Delta, China. *Air Quality Atmosphere and Health, 10*(6), 713-724. doi:10.1007/s11869-017-0472-1

APPENDIX A. SHINYEI PPD42NS SPECIFICATION

Items	Min	Norm	Max	Unit
VCC	4.75	-	5.25	V
Standby Current Supply	-	90	-	mA
Detectable Range of Concentration	-	0 ~ 28,000/ 0 ~ 8000	-	pcs/liter or pcs/0.01cf
Operating Temperature Range	0	-	45	℃
Output Method	Negative Logic, Digital output, Hi over 4.0V(Rev.2) Lo: under 0.7V			
Detecting the Particle Diameter	>1 um			
Dimensions	59(W) × 45(H) × 22(D) [mm]			
Humidity Range	95%rh or less			