Chapter 2

Particulate Matter Sampling Techniques and Data Modelling Methods

Jacqueline Whalley and Sara Zandi

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65054

Abstract

Particulate matter with 10 μm or less in diameter (PM_{10}) is known to have adverse effects on human health and the environment. For countries committed to reducing PM_{10} emissions, it is essential to have models that accurately estimate and predict PM_{10} concentrations for reporting and monitoring purposes. In this chapter, a broad overview of recent empirical statistical and machine learning techniques for modelling PM_{10} is presented. This includes the instrumentation used to measure particulate matter, data preprocessing, the selection of explanatory variables and modelling methods. Key features of some PM_{10} prediction models developed in the last 10 years are described, and current work modelling and predicting PM_{10} trends in New Zealand—a remote country of islands in the South Pacific Ocean—are examined. In conclusion, the issues and challenges faced when modelling PM_{10} are discussed and suggestions for future avenues of investigation, which could improve the precision of PM_{10} prediction and estimation models are presented.

Keywords: particulate matter, modelling, regression, artificial neural networks, instrumentation and measurement

1. Introduction

Particle pollution—also known as particulate matter or particulates—is a complex but stable gaseous suspension of liquid droplets and solid particles in the earth's atmosphere. Particle pollution is known to have many environmental effects from poor visibility to more serious consequences such as acid rain, which pollutes soil and water. The science of air quality is

complex, and many aspects of the problem are not understood fully. Particles are commonly classified according to their size as either coarse or fine. Fine particles have a diameter of 2.5 μm ($PM_{2.5}$) or less, and coarse particles are 10 μm or less (PM_{10}). Particulate matter that has a diameter over 100 μm tends not to stay airborne long enough to be measured. Fine particles are commonly generated through combustion or by secondary gas to particle reactions. These fine particles are typically rich in carbon, nitrates, sulphates and ammonium ions. Coarse particles are commonly the product of mechanical processes but also include naturally occurring wind-blown particles. A common example of coarse particulate matter is dust containing calcium, iron, silicon and other materials from the earth's crust.

Sources of particulate matter are often classified according to whether they originate from natural or anthropogenic sources. Natural sources include particles suspended in the atmosphere by volcanic eruptions, bush fires and pollen dispersal. Mechanistic processes cause natural particles such as dust and sea-salt particles to be suspended in the atmosphere. Biological sources of particulate matter are also natural sources; these consist largely of fungal spores (≤1 μm) and plant debris (normally < 2 μm) but also include microorganisms, viruses, pollen (≤10 μm) and fragments of living things (e.g. skin cells). Anthropogenic sources of biological particles include sources from farming, horticulture, waste disposal and sewage. Another anthropogenic source is emissions from combustion of fuels, for example, vehicle exhaust. In Europe, anthropogenic sources have been identified as the main contributor to PM_{10} due to urbanisation, high population density and areas of intensive industry. In New Zealand, the main contributors are also anthropogenic but are emissions from winter household heating (i.e. the wide use of wood-burning fires) and industry.

PM_{10} are so minute that they can be inhaled, penetrate the lungs and cause serious health problems. One event which illustrates the effect of particle pollution on human health is the 1952 'Great Smog' in London. Particle pollution from coal burning hung over the city for four days due to cold temperatures and lack of wind. Approximately 4000 deaths were linked to this single event [1]. As a result of events such as the Great Smog and obvious signs of climate change, many countries are now committed to international and national clean air legislation and air quality standards. These agreements require regular reporting of air quality including PM_{10} concentrations.

The economic costs of particulate pollution on a country can be significant. In the European Union in 2015, the cost of air pollution-related deaths was reported to be over US$1.4 trillion. In Israel, it is estimated that 2500 people a year die as a result of exposure to air pollutants [2]. In New Zealand (population ~ 4.4 million), it was reported that, despite relatively low air pollution when compared with other members of the Organisation for Economic Co-operation and Development, during 2012 a total of 1370 deaths, 830 hospital admissions and 2.55 million restricted activity days were linked to PM_{10} pollution [3]. Even low levels of PM_{10} have been found to significantly affect human health.

In order to make informed decisions, as individuals or as policymakers, it is critical that particulate matter is measured and modelled appropriately.

http://dx.doi.org/10.5772/65054

2. PM_{10} modelling

Models can be designed to estimate, predict or project. Discontinuities in data represent a real obstacle for time series analysis and prediction. Thus, estimating PM_{10} is important in situations where small periods of ground-truth data, acquired from sensors, are missing. Prediction models allow us to determine that something will happen in the future based on past data, generally with some level of probability, and are based on the assumption that future changes will not have a significant influence. In this sense, a prediction is most influenced by the initial conditions—the current situation from which we predict a change. Predicting short-range PM_{10} is important in order to identify days in which PM_{10} levels spike so that people with medical conditions which make them vulnerable to air pollution, such as asthmatics, can avoid exposure. It also allows for initiatives such as free public transport days to reduce commuter traffic volumes and thus reduce PM_{10} concentrations on a predicted high day. Models that allow for long-range projections are also important in order to assess the impact of different air quality management scenarios. A projection determines with a certain probability what could happen if certain assumed conditions prevailed in the future. Most PM_{10} models are designed to predict short range hourly, mean daily or maximum daily PM_{10} concentrations one day ahead.

A wide variety of techniques, ranging from simple to complex, have been used to predict PM_{10} concentrations. Mechanistic models are complex three-dimensional physiochemical models requiring theoretical information to simulate, using mathematical equations, the processes of particulate matter transportation and transformation (e.g. the air pollution model (TAPM) [4]). Such models are complex and time-consuming to implement and often prove inaccurate. Mechanistic models require a wide variety of input variables for which ground-truth data are not available. These missing data are either estimated or the model is simplified and all begin with meteorological forecasting, introducing both errors and uncertainties to a model.

Statistical models aim to discover relationships between PM_{10} concentrations and other explanatory variables. Statistical models work on a number of assumptions. Machine learning algorithms, on the other hand, are largely free of such assumptions and learn from the data they are presented with, finding patterns and relationships that are not necessarily obvious in the data. Machine learning approaches also tend to be good at modelling highly non-linear functions and can be trained to accurately generalise when presented with new, unseen data. As a result, machine learning methods have on the whole proven to be better at predicting PM_{10} concentrations than statistical models. This chapter focuses on statistical and machine learning approaches to PM_{10} modelling and prediction.

The vast majority of models in the last decade have been developed using a data-driven approach and have their origins in statistical modelling and machine learning. These models use ground-level sensor data and make no attempt to model the physical or chemical processes involved in PM_{10} generation, transportation and removal. They are reliant on measurements of pollutants and meteorological variables which are accurate only within a small area around

the monitoring stations. Thus, any model is limited by coverage, reliability and distribution of monitoring stations.

There are several steps in building an empirical PM_{10} model (**Figure 1**). The first is data acquisition from various types of particulate matter sensor. The next step is cleaning and preparing the raw data for analysis, including handling missing data, suspected errors and outliers. The next step, variable selection, is central to the performance of most models [5]. The aim of variable selection is to simplify the model by reducing the dimensions and removing any variables that do not significantly contribute to the model. The model is then built based on this subset of variables. Once a model is established, it is tested, after validation where required, by exposing the model to new data and measuring how well it predicts.

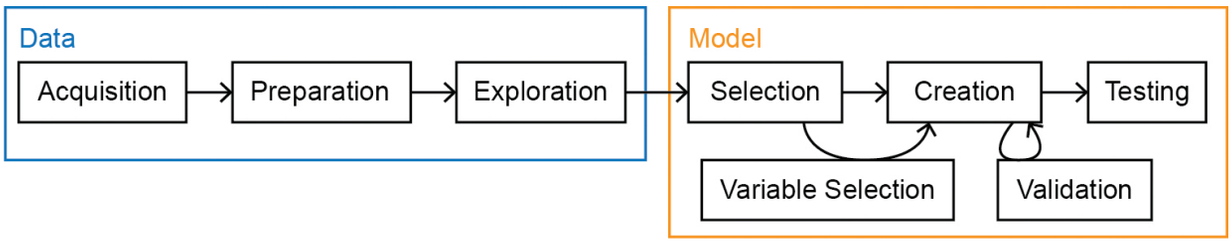

Figure 1. Key steps in the modelling process.

2.1. Particulate matter sampling techniques

The most common instruments for measuring particulate matter measure either its concentration or size distribution. The most accurate measurements are obtained from instruments that use a gravimetric (weighing) method. Air is drawn through a preweighed filter, and particles collect in the filter. The filter is then removed and reweighed. This approach has the added advantage that particles collected in the filter can be analysed chemically [6]. This method involves careful pre- and post-conditioning of the filter. Filter choice is also important as substrates are sensitive to environmental factors such as relative humidity. PTFE-bonded glass fibre has been found to be the most stable type of filter [7]. Accurate weighing is essential, and precise weighing protocols must be followed for results to be comparable [7]. This method is the most widely adopted by regulatory bodies including the EPA and the EU. However, it is not the most pragmatic method for PM_{10} modelling purposes because it is not real time and provides only average data for the period the filter was deployed. A manual process and consequently high operating costs limit the applications of this method. However, gravimetric measurements may be useful to provide a quick snapshot of PM_{10} at a site in order to determine locations for more intensive monitoring [8].

The TEOM™ sensor is the most commonly used instrument based on the microbalance method. TEOM™ uses a filter which is mounted on the end of a hollow tapered tube made of quartz. Particles collect on the filter and cause the oscillation frequency of the quartz tube to vary. PM_{10} measurements can be logged in near real time. A study which examined the

http://dx.doi.org/10.5772/65054

measurements on PM_{10} in New Zealand using microbalance measurement instruments found that the measurements were not equivalent to those from gravimetric methods [9].

Real-time monitoring of PM_{10} concentrations can be achieved using optical instruments. These instruments measure either light scattering, light absorption or light extinction caused by particulate matter. The most common instrument is an optical particle counter (OPC) which uses a light source, normally a laser diode, to illuminate particles and a photodetector to measure light scattered by those particles. Measurements may be periodically verified and calibrated using data from gravimetric instrumentation. OPC instruments have lower purchase and operating costs than gravimetric meters, but their lower precision and sensitivity mean that they are not considered appropriate for compliance monitoring [8]. However, the low cost of OPC instruments and real-time monitoring capability make OPCs suitable for particulate matter research.

Regardless of the data collection methods used, PM_{10} models are reliant on accurate and complete time series data from geographically localised monitoring stations.

2.2. Explanatory variables

Suspended PM_{10} regardless of location is dependent on many factors such as meteorological properties of the atmosphere, topo-geographical features, emission sources and the physical and chemical properties of the particles (size, shape and hygroscopicity). Many natural environmental factors influence PM_{10} concentrations from the time of year, to the weather, to extreme events such as volcanic eruptions and earthquakes. The effect of extreme events in nature on PM_{10} concentrations is well documented: high PM_{10} levels have been reported during heatwaves in Greece [10], as a result of forest fires [11], and in the aftermath of the Christchurch earthquakes in New Zealand [12]. Relatively low PM_{10} concentrations are observed during the monsoon season in India [13]. Of the myriad complex interrelated potential explanatory variables, only a small number have been used in the modelling of PM_{10} concentrations.

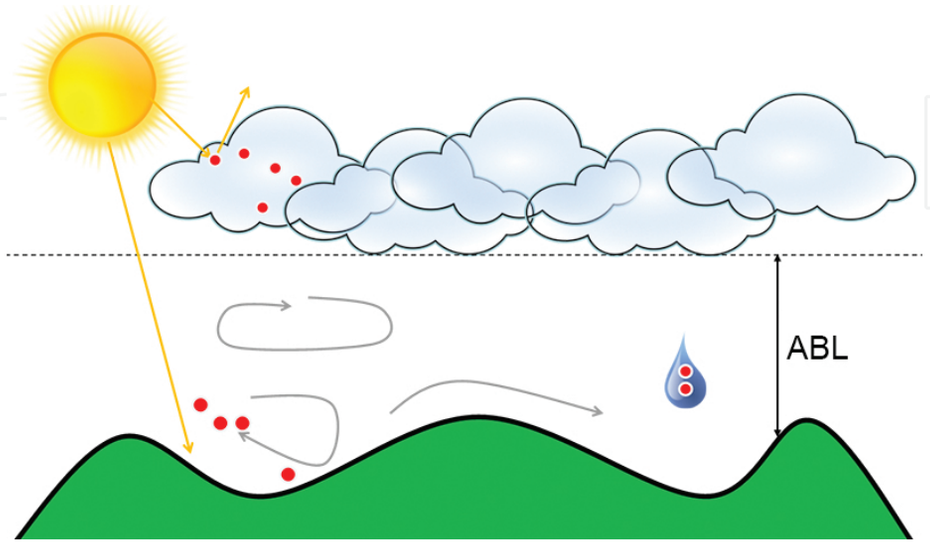

Figure 2. Particulate matter and the atmospheric boundary layer.

One key factor commonly used to explain and evaluate trends in PM_{10} data is the impact of meteorological conditions. The atmospheric boundary layer (ABL) is the lowest part of the earth's atmosphere (**Figure 2**). The thickness of the ABL can vary from 100 to 3000 m and extends from the ground to the point where cumulus clouds form. In the ABL wind, temperature and moisture fluctuate rapidly, and turbulence causes vertical and horizontal mixing. Suspended in the ABL, particles may undergo physical and chemical transformations triggered by factors such as the amount of water vapour, the air temperature, the intensity of solar radiation and the presence or absence of other atmospheric reactants. It is these physical processes, which help to explain why meteorological variables have such an influence on PM_{10} concentrations.

Having accurate and complete input data is critical to the success of any PM_{10} prediction model. As a result, most models make use of data that are readily recorded using weather station sensors. In cases where data are incomplete, the instance is often removed rather than imputed because of errors which may be introduced by estimation processes. The outputs of numerical weather forecast models can also be used as input variables in PM_{10} models. However, this is not common because of the uncertainties such variables introduce to PM_{10} predictions [14, 15].

Wind speed and temperature are the meteorological explanatory variables most frequently used in PM_{10} prediction models (**Table 1**). Wind variables have been found to be useful proxies for physical transportation factors; wind is critical to the horizontal dispersion of PM_{10} in the ABL. Wind direction controls the path that the PM_{10} will follow, while wind speed determines the distance it is carried and the degree to which PM_{10} is diluted due to plume stretching. The effect of wind speed and direction on PM_{10} varies with the geographical characteristics of a location. Low wind speed can be associated with high PM_{10} [16, 17]; this is common in hilly or mountainous regions. Conversely, in coastal or desert regions, high wind speeds result in high PM_{10} concentrations due to salt or dust suspension. In Europe, PM_{10} concentrations are significantly influenced by long-range transport contributions, which are independent of local emissions, so both wind direction and speed have a significant impact [18]. In Invercargill, New Zealand, where there are no close neighbours and thus little long-range transboundary PM_{10}, wind speed explains most of the variability in PM_{10} concentrations [19].

Cold temperatures increase the likelihood of an inversion layer forming in many locations. An inversion exists where a layer of cool air at the earth's surface is covered by a higher layer of warmer air. An inversion prevents the upward movement of air from the layers below and traps PM_{10} near the ground. As a result, cold temperatures tend to coincide with high concentrations of PM_{10}. However, in some locations days with high temperatures, no clouds and stable atmospheric conditions result in high PM_{10} [17]. In other locations when the difference between daily maximum and minimum temperatures is large and the height of the ABL mixing layer is low, high PM_{10} concentrations are observed [20].

PM_{10} levels can be reduced by rain, snow, fog and ice. Rain scavenging, a phenomenon in which below-cloud particles are captured and removed from the atmosphere by raindrops, is considered to be one of the major factors controlling the removal of PM_{10} from the air. The degree to which PM_{10} is removed is dependent on rainfall duration and intensity [21]. While rainfall is a primary factor in PM_{10} concentrations, it has not been used widely in models. This

http://dx.doi.org/10.5772/65054

is in part due to the fact that in some countries, there is no rain for long periods of time or little rainfall in summer. The lack of rain data means that it is not often included in PM_{10} models [14].

Study reference		**[16]**	**[26]**	**[25]**	**[34]**	**[33]**	**[35]**	**[35]**	**[39]**	**[14]**	**[23]**
Country of study (ISO 3166-1 alpha 3)		GRC	GRC	PRT	CHL	MYS	AUT	CZE	TUR	SAU	MYS
Predicted variable											
PM_{10}	Daily		Y	Y	Y	Y	Y	Y		Y	Y
	Hourly	Y							Y		
Explanatory variables											
	PM_{10} lag	Y		Y	Y	Y	Y	Y		Y	Y
Co-pollutants	CO_2			Y		Y				Y	Y
	SO_2			Y		Y				Y	Y
	NO			Y		Y					
	NO_2			Y		Y				Y	Y
	O_3					Y					Y
Meteorological data	Temperature	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Temperature lag						Y	Y			
	Wind direction	Y						Y	Y	Y	
	Wind direction lag										
	Wind speed	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Wind speed lag										
	Precipitation	Y			Y		Y				
	Solar radiation	Y									Y
	Sunshine hours										
	Air pressure				Y				Y		
	Dew point		Y								
	Humidity (%)	Y		Y	Y	Y			Y	Y	Y
	Cloud cover							Y			
	Date/time	Y	Y				Y	Y			
	Seasonal effects	Y						Y			
	Spatial variables										

Table 1. Explanatory variables used in recent MLR models for predicting PM_{10} concentrations.

Relative humidity has been used more frequently in models than rainfall. The relationship between PM_{10} concentration and relative humidity also depends on other meteorological conditions. For example, if humidity is high and there is also intense rainfall (such as during a monsoon season), then humidity has a negative correlation with PM_{10} due to rain scavenging. If high humidity is not accompanied by rainfall but is accompanied by high temperatures,

humidity has been found to contribute to higher PM_{10} concentrations. It has been suggested that when the relative humidity is over 55%, then PM_{10} concentrations are affected [22].

High solar radiation has also been shown to result in lower PM_{10}. When solar radiation is high, the surface of the earth is warmer; as a consequence the exchange of heat in the air results in turbulent eddies that disperse suspended particles [23].

Autocorrelation is a basic structural feature of the meteorological variables used in PM_{10} models. When a numeric time series correlates with its own past and future values, this is known as autocorrelation or *lagged correlation*. A positive autocorrelation indicates persistence and a tendency for a system to remain in the same state from one observation to the next. For example, if today is rainy, then tomorrow is more likely to be rainy. Most PM_{10} models rely solely on meteorological data from the same day (day t) and do not consider lagged ($t - n$) or lead ($t + n$) variables. However, the use of lagged variables has consistently increased the predictive power of such models. McKendry [24] found that one-day lagged ($t - 1$), two-day lagged ($t - 2$) and lead ($t + 1$) rainfall and lead ($t + 1$) wind direction contributed to models for estimating daily maximum PM_{10}. Lead ($t + 1$) daily mean temperature is also known to have good explanatory power for PM_{10}, whereas lagged daily mean temperature contributes to a lesser degree. PM_{10} is also autocorrelated and persistent, and therefore including lagged PM_{10} in the set of explanatory variables strengthens the predictive power of a model [24–26].

Co-pollutants—gases such as nitrogen monoxide (NO), nitrogen dioxide (NO_2), carbon monoxide (CO) and sulphur dioxide (SO_2)—have been found to be useful explanatory variables when used in conjunction with meteorological variables [24, 25, 27]. In many countries and especially in urban areas, road transportation is considered to be the largest contributor to PM_{10}. Road vehicles not only emit exhaust but also resuspend particulate matter [28]. Where data on traffic are not available, CO and NO_x can be used as a proxy for exhaust emissions [27].

Land usage can also influence PM_{10} concentrations, and therefore, land use type may be a useful explanatory variable. One study discovered spatial variations in PM_{10} with higher concentrations in commercial areas than in residential and industrial areas [29]. However, land use classifications are not common in PM_{10} models.

Another factor affecting PM_{10} concentrations is time. Various temporal variables have been used in models of PM_{10} concentration. Variables that reflect the seasonal cycle, such as sine and cosine of Julian day, are important for mean daily PM_{10} prediction because they reflect the dry, warm conditions typical in summer and therefore the role of photochemical production in increasing particulate matter concentrations [24]. Similarly, binary variables are sometimes used to indicate whether a period is cold or warm. For urban areas variables that reflect diurnal and weekly cycles are important due to high-density commuter and industrial traffic on weekdays contributing significantly to PM_{10} levels [16]. In urban areas of New Zealand PM_{10} has distinct diurnal cycles, with peaks between 10 pm and midnight and 8 am and 10 am, which have been found to be independent of population density [30]. Over the last 10 years, most PM_{10} models have included temporal variables.

A very recent approach to estimating PM_{10} concentrations is the use of satellite-based remote sensing in addition to ground-level meteorological variables. MODIS (Moderate Resolution Imaging Spectroradiometer) and MISR (Multi-angle Imaging Spectroradiometer) images are analysed using algorithms designed to calculate how much direct sunlight is prevented from reaching the ground by aerosol particles—the aerosol optical depth (AOD). Several studies have used AOD to estimate $PM_{2.5}$ and PM_{10} concentrations [31] and have shown that there is a high linear correlation between particulate matter concentrations and AOD [32]. One study published in 2014 used AOD along with meteorological variables to predict ground-level PM_{10} but did not evaluate the degree to which including AOD influenced the outcome of PM_{10} predictions [32].

2.3. Regression methods

Regression methods have been used as prediction and estimation tools in a wide range of disciplines including environmental pollution and climate studies. These methods are simple to implement and compute and provide models that are easily interpretable, hence their wide adoption. Among regression methods, multivariate linear regression (MLR) is probably the most commonly used statistical method for modelling air pollution and PM_{10}. **Table 1** summarises some recent MLR PM_{10} models reported in the literature, highlighting the explanatory variables which contributed to each model.

MLR is simply a process of finding a line that best fits a multidimensional cloud of data points. The line of best fit is computed to be the line in which the squared deviations of the observed points from that line are minimised. In other words, the constant term β_0 and the coefficients are calculated so that the average error ε is zero. This line of best fit provides a model (Eq. 1), which can be used to explain the relationship between one continuous response variable y, in this case PM_{10}, and two or more explanatory variables x_i:

$$y_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i \tag{1}$$

In MLR, it is assumed that a linear relationship exists between the response variable and the explanatory variables, all variables are normally distributed, there is little or no multicollinearity in the data (explanatory variables should not be highly correlated), and the residuals ($e_i = y_i - \hat{y}_i$) are homoscedastic (the variance around the predictor line is the same for all values of the response variable).

MLR models are considered to be limited models of PM_{10} concentration due to the inability to extend the response to non-central locations of the explanatory variables and to meet the other assumptions of the model [14, 33]. Despite possible nonconformity with one or more of the assumptions, MLR has been used extensively for predicting PM_{10} and is often used as a benchmark to which other methods are compared. Much of the PM_{10} modelling reported in the literature does not fully provide or explicitly address the data preparation and exploration

steps. Thus, it is often difficult to ascertain whether poor performance of MLR models is due to the fact that the data do not meet the assumptions of MLR, and therefore, MLR is a poor choice of model, or that the researchers chose not to refine the model when used only as a benchmark. In practice the assumption of linearity cannot be confirmed, and linear regression models are considered to be acceptable provided there are only minor deviations from this assumption. PM_{10} and its explanatory variables typically do not meet the assumption of linearity [22]. Often the variables do not have a normal distribution due to the presence of outliers.

Issues of linearity, non-normal distribution and homoscedasticity can be addressed by transforming the variable concerned. Such transformations are undertaken to linearise the relationship between the response and explanatory variables making model fitting simpler. Variable transformations should be handled carefully because in some cases, the transformation can introduce multicollinearity. Hourly PM_{10} concentrations in Athens are reported to have a logarithmic distribution so modellers performed a log transform of the PM_{10} response variable in order to improve the homoscedasticity of the residuals [16]. In Chile, maximum 24-hour moving average PM_{10} was also found to be logarithmically distributed. Again a log transform was used to normalise the data, and extreme outliers were removed [34]. After outlier analysis, all data were then normalised to ensure constant variance for each variable. In the case of Graz, Austria, the PM_{10} was gamma distributed, and a generalised linear modelling (GLM) approach using a log-link function was compared with MLR [35]. GLM is a generalisation of MLR that relates a linear model to its response variable by a link function and therefore allows for response variables that are not normally distributed. Little difference was observed between the two models, suggesting that the simpler MLR method was a better option than GLM in that case. Studies in other locales have reported that PM_{10} was normally distributed [26] or that PM_{10} concentrations were right skewed [14]. Studies have found that the use of curvilinear transformations of input variables (e.g. inverse transformation of wind speed) may result in improved regression models (e.g. see [36]).

Multicollinearity can be identified by examining the correlations among pairs of explanatory variables. However, looking at correlations only among pairs of predictors is not the best approach as even when pairwise correlations are small, it is possible that a linear dependence exists among three or more variables. Some PM_{10} modelling studies report on the linear correlation of PM_{10} with each of the possible explanatory variables but fail to examine correlations between those explanatory variables. A few recent studies have examined the correlations between pairs of explanatory variables but have not used this information to reduce multicollinearity [25]. Some researchers have used variance inflation factors (VIF) to ensure the assumption of no multicollinearity in the data before creating linear regression models [16, 22].

One approach for variable selection, in order to establish a parsimonious model, is stepwise regression: a systematic method for adding and removing terms from a multilinear model based on their statistical significance in a regression. The method begins with an initial model and then compares the explanatory power of incrementally larger (forward) or smaller (backward) models. In [16], a backward stepwise method was adopted to establish linear

regression models for daily average PM_{10} prediction one day ahead for four different sites in Athens. One problem with stepwise refinement is that a combination of variables to add or remove may be missed where their combined effect or non-effect is hidden by collinearity.

Pires et al. [25] investigated a number of alternative linear regression models including principal component regression (PCR), independent component regression (ICR), quantile regression (QR) and partial least squares regression (PLSR).

PCR uses principal component analysis (PCA) to create new variables, or principal components, that are orthogonal and uncorrelated linear combinations of the original explanatory variables. Linear regression is then used to determine a relationship between PM_{10} and selected principal components. PCR showed no improvement over MLR in terms of model performance [25]. In an earlier study [26], a PCR model for predicting one-day-ahead mean PM_{10} in Thessaloniki, Greece, was found to perform slightly better than MLR (for which no input selection was undertaken); PCR also better predicted high daily mean spikes in concentration.

ICR is another method that extends linear regression, but in this case, the input variables are independent components—linear combinations of latent variables—that are considered to be non-Gaussian and statistically independent.

QR models the relationship between a set of predictor variables and specific percentiles (or quantiles) of the response variable and thus gives a more complete picture of the effect of the predictors on the response variable. QR has been used very little for modelling pollutants. One study that compared models developed using QR with MLR found that QR was better at predicting hourly PM_{10} concentrations [14].

Like PCR, PLSR combines PCA and MLR, but in PLSR latent variables are selected such that they provide the maximum correlation with the response variable. If all the latent variables are used, then the results of PLSR will be very similar to those of PCR. PLSR is the most flexible of the extensions to MLR modelling and can be used in cases where MLR cannot. For example, PLSR is applicable where there are a large number of explanatory variables, and as a result, multicollinearity exists.

A comparison of regression models found the size of the data set is critical to performance [25]. Both ICR and QR were found to perform poorly on a large data set. MLR, PCR and PLSR all gave similar results and performed best on a large data set. On a smaller data set, the models that removed the correlation of the variables—PCA, ICR and PLSR—performed best.

Another recent regression-based modelling approach is cluster-wise linear regression which is founded on the 'mixtures of linear regression' statistical framework [37, 38]. When compared with generalised additive non-linear models, cluster-wise linear models performed extremely well and further exploration of such models has been recommended [37].

Some research suggests that, in general, non-linear regression methods outperform their linear counterparts. One study using AOD from satellite imagery along with meteorological variables—temperature, wind speed, wind direction, relative humidity and planetary boundary layer height—to compare a non-linear regression model with a linear regression model found that the non-linear models outperformed their counterparts [32]. In the non-linear

model, non-linear functions were used to reflect the non-linear influences of the explanatory variables. For example, an exponential function of relative humidity was used to account for the growth of the particle size with increasing humidity (Eq. 2):

$$PM_{10} = \left(e^{\beta_0 + \beta_T(T) + \beta_{WD}(WD)}\right) \times \left(e^{\beta_{RH}(RH)}\right) \times \left(AOD^{\beta_{AOD}}\right) \times \left(PBL^{\beta_{PBL}}\right) \times WS^{\beta_{WS}} \quad (2)$$

Some researchers have also explored using generalised additive models (GAMs) to model PM_{10} concentrations [14, 18, 27]. GAMs are a nonparametric extension of GLM that are flexible and able to handle non-linear relationships well. Like GLM, the response variable may have an exponential distribution (e.g. gamma, Poisson, exponential, etc.). GAMs assume that the mean of the response variable is dependent on additive predictors through a non-linear link function g and establish the nature of the relationship using smoothing functions which are determined by the data (Eq. 3):

$$g\left(E\left[y \mid x\right]\right) = \beta_0 + f_1\left(x_1\right) + \ldots + f_p\left(x_p\right) \quad (3)$$

In a daily mean PM_{10} GAM that employed co-pollutants, the previous day's mean PM_{10}, and meteorological explanatory variables, significant differences were seen between the mean observed and predicted PM_{10} concentrations [14]. The model was also found to underestimate occurrences of high PM_{10}. GAMs have not been used widely for PM_{10} modelling, possibly as they can be prone to overfitting, hard to interpret and computationally expensive.

2.4. Artificial neural networks

Artificial neural networks (ANNs) are the most common of the machine learning approaches used to model PM_{10} and have seen rapid growth in this field since the year 2000. ANNs are suited to modelling complex and dynamic non-linear systems and are particularly useful because they can be trained to accurately generalise when presented with new information. Air pollution models have been developed using ANNs for modelling and forecasting single air pollutant indicators such as ozone (O_3), NO_2, SO_2 and particulate matter. In almost all cases, ANNs, properly designed and trained, have been found to provide more accurate predictions than traditional linear statistical approaches.

ANNs are a family of computational machine learning algorithms inspired by the way biological nervous systems process and learn from information. ANNs consist of a large number of interconnected nodes, called neurons, that work together to learn and identify patterns. They have three or more layers: one input layer consisting of one or more neurons, one or more hidden layers where the learning occurs and one output layer. The neurons in this interconnected network send signals to each other along weighted connections. The most common variant of ANNs used in PM_{10} prediction is a feedforward ANN in which signals

http://dx.doi.org/10.5772/65054

travel in one direction from input neurons (explanatory variables) to output (response variable).

A multilayer perceptron (MLP) is a fully connected feedforward neural network and is in essence a logistic regression classifier that facilitates non-linear transformations between inputs and outputs. The weighted values of each input are passed to the hidden layer which generates output weightings for each neuron. Each neuron j receives incoming signals from every neuron i in the previous layer. The effective incoming signal N_j to the node j is the weighted sum of all the incoming signals (Eq. 4) [39]:

$$N_j = \sum_{i=1}^{m} x_i W_{ji} \tag{4}$$

where m is the number of input neurons converging into neuron j, x_i is incoming signal and W_{ji} is a synaptic weight associated with each x_i. An activation function is then applied to the N_j to produce the output signal of the neuron j. The most commonly used activation function in backpropagation (BP) networks is the logistic sigmoid function shown in Eq. 5:

$$\sigma\left(N_j\right) = \frac{1}{\left(1 + e^{-N_j}\right)} \tag{5}$$

In order to train a neural network, the weights of each input have to be adjusted so that the error between the expected output and the predicted output is reduced. Backpropagation is a widely used method for computing the error derivative of the weights. In BP, the input data are repeatedly presented to the neural network. If the desired output is not achieved, the error weighting is propagated backwards through the network and the synaptic weights adjusted. Once the ANN has been trained, the network can be used to perform a forecast using a testing data set.

When using ANNs to model PM_{10}, it is important to ensure that there is sufficient data to adequately train the network. PM_{10} concentration time series data are typically noisy and contain outliers. A suggested 'rule of the thumb' is that a very noisy target variable requires over 30 times as many training cases as weights. The necessity for copious training data is one of the main impediments to using ANNs [40]. A validation data set is used to tune the model by exposing it to new data prior to testing its ability to predict [41]. In the PM_{10} modelling literature, many researchers do not mention the way in which they dealt with the training and testing of their models beyond the fact that they measured the difference between the observed and predicted values. It is important when using machine learning approaches to consider the size of the data sets and the approach used to train, validate and test models.

Spatial and temporal variations of PM_{10} concentration are attributed to complex interactions between high numbers of input variables which mean an even higher number of weights are

needed to train a network with a fully connected topology. The more inputs, the more time that an ANN model takes to develop. As with regression models, eliminating irrelevant or interdependent variables can improve an ANN's performance—thus keeping the model complexity low while achieving the best fit possible. Often other bioinspired optimisation methods are used in order to select the inputs for an ANN. Genetic algorithms (GAs) use a natural selection process, inspired by Darwinian evolution, to try to find the fittest solution. Individual solutions evolve through mechanisms of reproduction and mutation to produce new solutions from which only the fittest or best survive, thus finding the optimal solution—in this case the best subset of explanatory variables for PM_{10} concentration. A GA can be used to set the initial weights of the input variables for an ANN; indeed, there are some instances where GAs have been used to select the best inputs for ANN modelling of PM_{10} concentrations [24, 42].

A common rhetoric throughout work using ANNs to model PM_{10} concentrations is that ANNs are more appropriate for modelling PM_{10} than conventional deterministic approaches such as regression because ANNs are better at dealing with multiple correlated factors, uncertainty and non-linearity [43–45]. It is also generally accepted that ANNs are useful in cases where a full theoretical approach is not available and a large number of different variables are involved [46].

The first reported use of an ANN as a prediction model for ground-level SO_2 pollution was in 1993 [47]. In 1999, an MLP was used to model hourly NO_x and NO_2 pollutant concentrations in central London using basic hourly meteorological data. Their results have shown that the ANN models outperformed their previous attempts to model the same pollutants using regression-based models [48]. Another early study which compared ANNs with MLR for predicting daily mean and maximum PM_{10} and $PM_{2.5}$ reported that neural network models showed little if any improvement over regression models [24]. Despite this finding, considerable interest and work continue in the area of ANNs for modelling air quality. One study reported that multilayer ANN models are able to predict the exceedances of PM_{10} concentration thresholds in about 75% of cases, suggesting that ANNs might be suitable for predicting episodic events in which air pollution is high [20].

Since then, the vast majority of air quality prediction models developed using ANN methods have used the MLP variant. MLP models have been used to successfully predict daily average PM_{10} concentrations one day in advance in urban areas of Belgium [49] and for Santiago in Chile [20]. A BP neural network was used to predict mean hourly air pollutant concentrations ten hours ahead (CO, NO_2, O_3 and PM_{10}) for Guangzhou in China [50]. A summary of some models developed in the last ten years using ANNs to model PM_{10} is presented in **Table 2**, contrasting the input variables. These studies have used various ANN architectures and different input parameters to obtain the most predictive models possible for their locations.

A number of studies have compared the performance of ANNs and MLR [17, 33, 43, 46] for the prediction of PM_{10} for various locations. In all cases ANNs were found to outperform MLR. ANNs were also found to give better predictions for PM_{10} concentrations than a deterministic dispersion model [43]. In [15] MLP, a radial basis function ANN and PCR models for prediction of mean hourly PM_{10} concentrations in Cyprus were compared. A variety of meteorological,

co-pollutant and temporal variables as well as lag PM_{10} were used as input to the models. Among these models, MLP was found to give the most accurate predictions for all the sites investigated.

Study reference		**[33]**	**[54]**	**[54]**	**[54]**	**[23]**	**[62]**	**[63]**	**[51]**	**[51]**	**[52]**	**[42]**	**[42]**
Country of study (ISO 3166-1 alpha 3)		MYS	CHN	CHN	CHN	MYS	NZL	NZL	MEX	MEX	TUR	IRN	IRN
Predicted variable													
PM_{10}	Daily	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Hourly												
Explanatory variables													
	PM_{10} lag	Y				Y					Y		
Co-pollutants	CO_2	Y				Y							
	SO_2	Y				Y							
	NO												
	NO_2	Y				Y							
	O_3	Y				Y							
Meteorological data	Temperature	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Temperature lag										Y		
	Wind direction					Y	Y	Y	Y	Y	Y		
	Wind direction lag										Y		
	Wind speed	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Wind speed lag										Y		
	Precipitation		Y	Y	Y								
	Solar radiation						Y						
	Sunshine hours		Y	Y	Y			Y					
	Air pressure		Y	Y	Y								
	Dew point												
	Humidity (%)	Y	Y	Y	Y		Y		Y	Y	Y	Y	Y
	Cloud cover												
	Date/time											Y	Y
	Seasonal effects												
	Spatial variables		Y	Y	Y			Y					

Table 2. Explanatory variables used in recent MLP models for predicting PM_{10} concentrations.

In order to improve on MLP model performance, one approach is to use clustering algorithms to find relationships between PM_{10} and meteorological variables using the clusters as input to the model [51]. When compared with MLP, it was found that using an ANN with k-means

clusters gave improved predictions for daily mean PM_{10} for Salamanca in Mexico. Another approach reported that improved daily predictions could be achieved by developing separate MLP models for winter and summer periods. PCA was used to create ANN input vectors which were components of the most significant lagged variables within a seven-day period [52].

A study forecasting PM_{10} using different variants of ANNs found that ANN models are effective tools for two-day-ahead prediction of incidences of high PM_{10} concentration. The models were based on four simple input variables—the daily mean wind intensity, wind speed, temperature and barometric pressure [53]. The data were preprocessed to eliminate errors introduced by instrumentation and the input variables normalised in the range of −1 to 1. The best-performing ANN was reported to be that with an Elman topology. The authors concluded that ANN models are an effective tool for early high-level air pollutant warning systems. In a recent study, three ANN models—MLP, Elman and support vector machines (SVM)—for six stations in Wuhan, China, were compared. Because of the rapid development and intensive construction occurring in the area, a Construction Index (CI) was included to represent the dust arising from building works. The time series data were decomposed into wavelet functions, and wavelet coefficients were predicted. All three models were found to give similar results, and when CI was included, the model gave improved predictions for peaks in PM_{10} concentration [54].

2.5. Alternative approaches

A few recent studies have explored other machine learning methods for modelling PM_{10}. One approach is Classification and Regression Trees (CART). CART models are obtained by recursively splitting the data and fitting a simple prediction model within each partition to create a decision tree. In one study CART was found to give better overall prediction than an ANN, but the ANN was found to better reflect temporal trends in PM_{10} [26]. Random forests (RFs) are an ensemble technique which consists of a number of learned decision trees that are used to determine the PM_{10} prediction. Predicting PM_{10} is usually considered to be a regression problem so the RF's tree responses are averaged to obtain an estimate of the dependent variable. One study showed that RFs gave better daily mean PM_{10} predictions than MLP or SVM approaches [5]. This finding is supported by another study which also found that an RF model outperformed an SVM model [55]. Tzima et al. explored a number of machine learning approaches including decision trees and RFs for predicting daily mean PM_{10} and found that logistic model trees—decision trees with logistic regression models at the leaves—gave the best predictions [56].

3. New Zealand PM_{10} trends and models

New Zealand is a country situated in the Pacific Ocean and comprises two main islands, the North Island and South Island, and numerous smaller islands. New Zealand has a mild and temperate maritime climate, but conditions can change rapidly and vary dramatically across

http://dx.doi.org/10.5772/65054

regions from very wet on the west coast of the South Island, to semiarid in Central Otago, to subtropical in Northland.

New Zealand has one of the best-reported air qualities in the OECD. Legislations and guidelines govern the level of PM_{10} that is acceptable. The World Health Organisation (WHO) has set guidelines for short- and long-term PM_{10} exposure levels. These guidelines state that there should be no more than three exceedances of the daily mean limit of 50 μg/m^3 in a year and annual concentrations should not exceed 20 μg/m^3 [57]. New Zealand also has its own local standards, the National Environmental Standards (NES), which specifies a daily mean PM_{10} threshold of 50 μg/m^3 measured between the hours of midnight and midnight. One exceedance per year is allowed. In 2012, 87% of New Zealand monitoring sites meet the WHO guidelines, while 50% had levels which exceeded the short-term standard [12]. The sites which exceeded WHO guidelines were mainly in the South Island where the winters are colder and a large proportion of home heating uses wood burners. In 2013, Christchurch, also in the South Island, is reported to have exceeded the NES limit on about 30 occasions [58].

During summer, anthropogenic sources are mainly traffic and industry. Thus, PM_{10} concentrations in New Zealand are linked strongly to season. **Figure 3** (left) shows a heat map of PM_{10} concentrations by month and time of day; the highest PM_{10} concentrations occur during the evening in the winter from May to August. As in other parts of the world, PM_{10} concentrations are autocorrelated (**Figure 3**, (right)) and dependent on meteorological conditions.

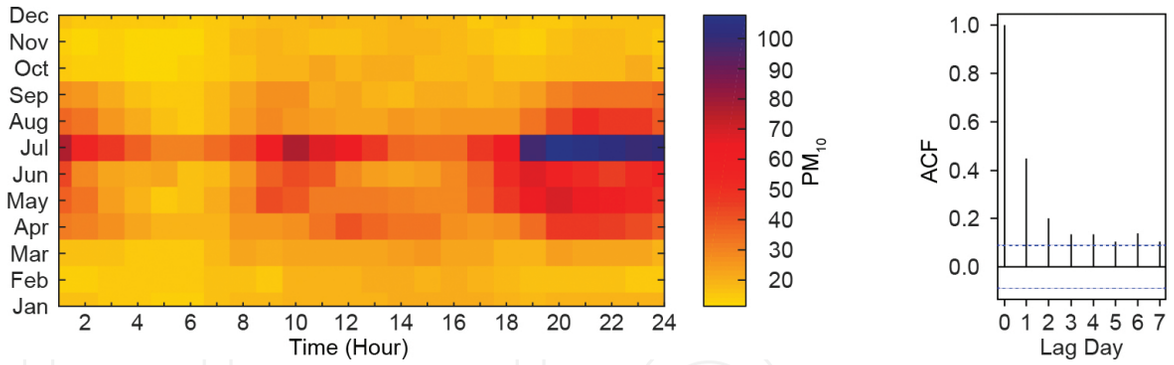

Figure 3. Heat map of PM_{10} concentrations in Timaru for 2012, data source [59] (left) and autocorrelation function (ACF) for a typical week of PM_{10} concentration observations in Auckland (right).

Christchurch is a coastal city with a flat topology surrounded by hills and is therefore subject to complex ABL winds. Wind speed and vertical temperature gradient influence PM_{10} concentrations by dispersing pollutants. A relatively recent study examined a method for studying PM_{10} emission trends by removing meteorological factors [58]. PM_{10} was calibrated and imputed using simple linear regression and transformed using the natural logarithm. Transformed PM_{10} values were then regressed against temperature, the calculated linear dependency was removed and summer observations were scaled so the PM_{10} time series was independent of seasonal fluctuations. This data set was input for multiple linear regression analysis with temperature and wind speed as explanatory variables. The resulting model could only explain 20% of the variation in PM_{10} concentrations. Residuals of the observed and

predicted values were calculated to investigate the variations in PM_{10} that were unexplained by the regression model. These residuals were added to the overall mean of the temperature-corrected PM_{10} data and a simple moving average filter applied to smooth the data. The modelled trend showed peak emissions in 2001 and 2002 with a subsequent steady decline. This trend did not match with those reported by local authorities in their three yearly emission inventories in which a steady decline was reported [58]. However, it is difficult to compare the two. In the inventory, constant emissions are assumed and then modified according to meteorological conditions and are only undertaken every three years. In [58], the method has been modified to allow for emissions that are not constant, and the model is based on hourly observations making it difficult to assess the success of the method.

In 2010, a study was undertaken to identify the influence of weather factors on occurrences of high PM_{10} concentrations—those in which the NES limits were breached—in Blenheim [60]. Blenheim is a small coastal town (population ~ 30,000) in the South Island of New Zealand. The town is on a flat area surrounded by hills on three sides. Blenheim has a dry climate with hot summers and cold winters. A boosted regression tree using a Gaussian link function was used to identify the meteorological variables which best explained the observed variance in PM_{10} concentrations. Mean daily wind speed and average temperature between 8 pm and midnight were found to best explain the variance; these variables were then used as input to a normal regression tree. It was discovered that low wind speed and low temperatures explained the majority of the NES exceedances. A similar result obtained for Invercargill using CART found that low wind speeds and low temperatures in the evening hours also accounted for most of the variation in PM_{10} levels [19]. In both studies, the model was used to account for trends in PM_{10} rather than to predict or estimate PM_{10}.

Much of the PM_{10} modelling undertaken in New Zealand until recently has been for areas in the South Island. This may be due to the fact that there is constant and historic time series data available from a well-maintained network of South Island PM_{10} monitoring stations or that frequent and higher exceedances of PM_{10} limits have been recorded for South Island regions than for regions in the North Island.

An in-depth study of Christchurch's daily mean PM_{10} employing statistical modelling approaches was undertaken using GLM, GAM, generalised additive mixed model with auto-correlated errors (GAMM + AR) and QR [61]. All of the models evaluated used a natural log transform of the PM_{10} response variable as a number of the explanatory meteorological variables impacted PM_{10} concentrations in a negative exponential form. It was concluded that simple linear regression modelling was not a suitable approach as the data violated all of the assumptions. A total of 41 meteorological variables were considered from which a subset of 20 in addition to lag PM_{10} were chosen by forward and backward stepwise selection. Models were built using the response PM_{10} data both without imputation and with missing values imputed by linear interpolation. The GAMM+AR model was found to be the best prediction model and able to explain around 70% of the variability in daily average PM_{10} concentrations [61].

There have been very few models developed using ANNs to estimate or predict PM_{10} concentrations in New Zealand. Gardner and Dorling [48] compared the performance of

different models such as linear regression, feedforward ANNs and CART approaches for modelling mean hourly PM_{10} in Christchurch, New Zealand. As with studies in other parts of the world, ANNs were found to be the best-performing modelling method. In another more recent study, ANNs were combined with a k-means clustering method to group and rank explanatory variables. The data used were from Auckland—New Zealand's most populated city with a population of over 1.4 million. It was found that the inclusion of cluster rankings, derived from k-means cluster analysis, as an input parameter to the ANN model showed a statistically significant improvement in the performance of the ANN model and that the model was also better at predicting high concentrations [62, 63].

Near-ground maximum PM_{10} concentrations for two sites in Timaru, a small rural town, were estimated using a feedforward backpropagation ANN with a hyperbolic tangent sigmoid function [41]. The response and explanatory variables were normalised. Additionally, due to the correlation between the seasonal changes and PM_{10} concentration, the PM_{10} data were divided into high season (winter/autumn) and low season (spring/summer) classes prior to creating the model. The inputs included one-day lagged meteorological variables and one-day lagged PM_{10}, in addition to meteorological variables for the day of estimation. Levenberg-Marquardt optimisation and Bayesian regularisation training were evaluated, and it was found that Bayesian regularisation was the best approach for tuning the weights and bias values for the network. This approach gave good estimations of daily mean PM_{10} concentration for both sites.

Some research has been conducted using TAPM, a deterministic global atmospheric pollutant model, [4] which includes fundamental fluid dynamics and scalar transport equations to predict meteorology and pollutant concentration [64–66]. Localised models of PM_{10} concentrations for two South Island towns, Alexandra (population ~ 5000) and Mosgiel (population ~ 10,000), were developed. Alexandra has a borderline oceanic semiarid climate—the country's coldest, driest and warmest—due to its geographic location as New Zealand's most inland town. Mosgiel is separated from Dunedin city by hills and is situated on a plain. It has a temperate climate with a significant annual average rainfall of 738 mm. TAPM was found to correctly predict daily PM_{10} concentration breaches and non-breaches of the NES 66% of the time in Alexandra and 71% of the time in Mosgiel [65]. Another study has looked at TAPM for simulating PM_{10} dispersion for a single winter in Masterton and also obtained good predictions of PM_{10} [65]. Yearlong PM_{10} was modelled using TAPM for Christchurch city. TAPM was reported to provide an acceptable simulation of ground-level weather and PM_{10} dispersion (with a 4 $\mu g/m^3$ difference in annually averaged concentration of modelled and measured PM_{10}), but the model tended to overestimate wind speed during still nights resulting in low PM_{10} estimates for those periods [66].

4. Summary

Although there are now several models available for predicting PM_{10}, it is difficult to compare them. The complex nature of ambient particulate matter composition and the physical and

chemical transformations that particulate matter can undergo between emission source and sampling location seems to mean that PM_{10} concentrations are largely explained by location-specific variables and events. Meteorological variables used in these localised models tend to be restricted to those which are routinely collected by local authorities.

It is also difficult to compare models because of the variation in PM_{10} instrumentation and measurement approaches used between different studies. In the future, improved sensor technology and lower costs associated with such monitoring could allow for more comprehensive coverage of areas—improving the inputs available for modelling. Ability to sense at different atmospheric levels should also enhance the data and in turn any empirical models. The use of geo-topological features such as elevation and land use could be considered as inputs for modelling as they reflect site-specific conditions and are readily available, but few models utilise these variables. Inclusion of air quality data, such as AOD measurements, from satellite-based remote sensing should also enhance models. Such data have the potential to provide a means of imputing missing values, to verify and enhance the accuracy of sensor-based ground-level observations and to provide additional inputs to models.

While general trends in PM_{10} concentrations can be explained and similarities can be seen between countries and factors contributing to PM_{10}, no empirical comparison can be made between models developed for specific locations. An attempt to develop a single general model for an area found that the general model performed poorly compared with site-specific models [32]. Some of these site-specific issues are removed when a deterministic physiochemical modelling approach is used, but accuracy of such models is currently limited as many of the actual mechanisms involved in pollution generation, dispersal, dilution and removal are not fully understood. However, it is possible that in the future, with better understanding, deterministic models could prove to be the way forwards.

Author details

Jacqueline Whalley* and Sara Zandi

*Address all correspondence to: jacqueline.whalley@aut.ac.nz

Auckland University of Technology, Auckland, New Zealand

References

[1] Stone R. Counting the Cost of London's Killer Smog. Science. 2002;298(5601):2107–7.

[2] European Environment Agency. Air quality in Europe-2015 report [Internet]. 2015 . Available from: http://www.eea.europa.eu/publications/air-quality-in-europe-2015 [Accessed: 2016-07-22]

[3] Statistics NZ. Health effects from exposure to PM10 [Internet]. 2012 . Available from: http://www.stats.govt.nz/browse_for_stats/environment/environmental-reporting-series/environmental-indicators/Home/Air/health-effects.aspx [Accessed: 2016-07-22]

[4] CSIRO. The Air Pollution Model (TAPM) [Internet]. 2015 . Available from: http://www.csiro.au/en/Research/OandA/Areas/Assessing-our-climate/Air-pollution/TAPM [Accessed: 2016-07-20]

[5] Siwek K, Osowski S. Data mining methods for prediction of air pollution. International Journal of Applied Mathematics and Computer Science. 2016;26(2):467–78.

[6] Nussbaumer T, Czasch C, Klippel N, Johansson L, Tullin C. Particulate emissions from biomass combustion in IEA countries. In: 16th European biomass conference and exhibition; 2–6 June; Valencia, Spain. 2008.

[7] Brown AS, Yardley RE, Quincey PG, Butterfield DM. Ambient air particulate matter: quantifying errors in gravimetric measurements. NPL Report DQL-AS. 2005 Jan:41.

[8] Ministry for the Environment. Good Practice Guide for Air Quality Monitoring and Data Management 2009 [Internet]. 2009. Available from: https://www.mfe.govt.nz/sites/default/files/good-practice-guide-for-air-quality.pdf [Accessed: 2016-07-22]

[9] Bluett J, Wilton E, Franklin P, Dey K, Aberkane T, Petersen J, et al. PM10 in New Zealand's urban air: a comparison of monitoring methods [Internet]. 2007. Available from: https://www.niwa.co.nz/sites/niwa.co.nz/files/import/attachments/CHC2007_059.pdf [Accessed: 2016-07-22]

[10] Pakalidou N, Katragkou E, Poupkou A, Zanis A, Bloutsos S, Karacostas, T. Analysis of Heat-Wave Events in Thessaloniki and Investigation of Impacts on PM10. In: Helmis G.C., Nastos T.P., editors. Advances in Meteorology, Climatology and Atmospheric Physics. Berlin, Heidelberg: Springer; 2013. p. 663–669. DOI: 10.1007/978-3-642-29172-2_94

[11] Junpen A, Garivait SB, Onnet P, Ongpullponsak A. Spatial and temporal distribution of forest fire PM10 emission estimation by using remote sensing information. International Journal of Environmental Science and Development. 2011;2(2):156–61. DOI: 10.7763/IJESD.2011.V2.115

[12] Ministry for the Environment and Statistics New Zealand. New Zealand's Environmental Reporting Series: 2014 Air domain report [Internet]. 2014 . Available from: www.mfe.govt.nz [Accessed: 2016-07-22]

[13] Tiwari S, Chate DM, Pragya P, Ali K, Bisht DS. Variations in mass of the PM 10, PM 2.5 and PM 1 during the monsoon and the winter at New Delhi. Aerosol and Air Quality Research. 2012;12(1):20–9. DOI: 10.4209/aaqr.2011.06.0075

[14] Sayegh AS, Munir S, Habeebullah TM. Comparing the performance of statistical models for predicting PM10 concentrations. Aerosol and Air Quality Research. 2014;14(3):653–65. DOI: 10.4209/aaqr.2013.07.0259

[15] Paschalidou AK, Karakitsios S, Kleanthous S, Kassomenos PA. Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models: implications to local environmental management. Environmental Science and Pollution Research. 2011;18(2):316–27.

[16] Grivas G, Chaloulakou A. Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. Atmospheric Environment. 2006;40(7):1216–29.

[17] Papanastasiou DK, Melas D, Kioutsioukis I. Development and assessment of neural network and multiple regression models in order to predict PM10 levels in a medium-sized Mediterranean city. Water, Air, & Soil Pollution. 2007;182(1–4):325–34.

[18] Barmpadimos I, Hueglin C, Keller J, Henne S, Prévôt ASH. Influence of meteorology on PM 10 trends and variability in Switzerland from 1991 to 2008. Atmospheric Chemistry and Physics. 2011;11(4):1813–35.

[19] Wilton E, Appelhans T, Baynes M, Zawar Reza P. Assessing long- term trends in PM 10 concentrations in Invercargill [Internet]. 2009 . Available from: http://www.environet.co.nz/environet/documents/TrendsPM10InvercargillFinal.pdf [Accessed: 2016-07-22]

[20] Perez P, Reyes J. Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile. Atmospheric Environment. 2002;36(28):4555–61.

[21] Olszowski T. Changes in PM10 concentration due to large-scale rainfall. Arabian Journal of Geosciences. 2016;9(2):1–11.

[22] Ul-Saufie AZ, Yahaya AS, Ramli NA, Rosaida N, Hamid HA. Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmospheric Environment. 2013;77:621–30.

[23] Afzali A, Rashid M, Sabariah B, Ramli M. PM10 pollution: its prediction and meteorological influence in PasirGudang, Johor. IOP Conference Series: Earth and Environmental Science. 2014;18: 012100. DOI: 10.1088/1755-1315/18/1/012100

[24] McKendry IG. Evaluation of artificial neural networks for fine particulate pollution (PM10 and PM2.5) forecasting. Journal of the Air & Waste Management Association. 2002;52(9):1096–101.

[25] Pires JCM, Martins FG, Sousa SI V, Alvim-Ferraz MCM, Pereira MC. Prediction of the daily mean PM 10 concentrations using linear models. American Journal of Environmental Sciences. 2008;4(5):445–53.

[26] Slini T, Kaprara A, Karatzas K, Moussiopoulos N. PM10 forecasting for Thessaloniki, Greece. Environmental Modelling & Software. 2006;21(4):559–65.

[27] Munir S, Habeebullah TM, Seroji AR, Morsy EA, Mohammed AMF, Saud WA, et al. Modeling particulate matter concentrations in Makkah, applying a statistical modeling approach. Aerosol and Air Quality Research. 2013;13(3):901–10.

[28] Chen L, Bai Z, Kong S, Han B, You Y, Ding X, et al.. A land use regression for predicting NO2 and PM10 concentrations in different seasons in Tianjin region, China. Journal of Environmental Sciences. 2010;22(9):1364–73.

[29] Yu CH, Fan Z-H, Meng Q, Zhu X, Korn L, Bonanno LJ . Spatial/temporal variations of elemental carbon, organic carbon, and trace elements in PM 10 and the impact of land-use patterns on community air pollution in Paterson, NJ. Journal of the Air & Waste Management Association. 2011;61(6):673–88.

[30] Trompetter WJ, Davy PK, Markwitz A. Influence of environmental conditions on carbonaceous particle concentrations within New Zealand. Journal of Aerosol Science. 2010;41(1):134–42.

[31] Péré JC, Pont V, Mallet M, Bessagnet B. Mapping of PM10 surface concentrations derived from satellite observations of aerosol optical thickness over South-Eastern France. Atmospheric Research. 2009;19(1):1–8.

[32] Sotoudeheian S, Arhami M. Estimating ground-level PM10 using satellite remote sensing and ground-based meteorological measurements over Tehran. Journal of Environmental Health Science & Engineering. 2014;12(1):122. DOI: 10.1186/s40201-014-0122-6

[33] Ul-saufie AZ, Yahaya AS, Ramli NA, Hamid HA. Comparison between Multiple linear regression and feedforward back propagation neural network models for predicting PM 10 concentration level based on gaseous and meteorological parameters. International Journal of Applied Science and Technology. 2011;1(4):42–29.

[34] Díaz-Robles LA, Ortega JC, Fu JS, Reed GD, Chow JC, Watson JG, et al. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmospheric Environment. 2008;42(35):8331–40.

[35] Stadlober E, Zuzana H. Forecasting of daily PM10 concentrations in Brno and Graz by different regression approaches. Austrian Journal of Statistics. 2012;41(4):287–310.

[36] Chaloulakou A, Kassomenos P, Spyrellis N, Demokritou P, Koutrakis P. Measurements of PM10 and PM2.5 particle concentrations in Athens, Greece. Atmospheric Environment. 2003;37(5):649–60.

[37] Poggi JM, Portier B. PM10 forecasting using clusterwise regression. Atmospheric Environment. 2011;45(38):7005–14.

[38] Misiti M, Misiti Y, Poggi J, Portier B. Mixture of linear regression models for short term PM10 forecasting in Haute Normandie (France). Case Studies in Business, Industry and Government Statistics. 2015;6(1):47–60.

[39] Ozdemir U, Taner S. Impacts of meteorological factors on PM10: Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) approaches. Environmental Forensics. 2014;15(4):329–26.

[40] Ordieres JB, Vergara EP, Capuz RS, Salazar RE. Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environmental Modelling & Software. 2005;20(5):547–59.

[41] Zandi S, Whalley J, Sallis P, Ghobakhlou A. Estimation of Near Ground PM10 Concentrations using Artificial Neural Networks. In: Weber T, McPhee MJ, Anderssen RS, editors. MODSIM2015, 21st International Congress on Modelling and Simulation; Queensland, Australia. Modelling and Simulation Society of Australia and New Zealand Inc.; 2015. p. 42.

[42] Asghari Esfandani M, Nematzadeh H. Predicting air pollution in Tehran: Genetic algorithm and back propagation neural network. Journal of AI and Data Mining. 2016;4(1):49–54.

[43] Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, et al. Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment. 2003;37(32):4539–50.

[44] Esplin GJ. Approximate explicit solution to the general line source problem. Atmospheric Environment. 1995;29(12):1459–63.

[45] Hewitson B, Crane RG. Neural Nets: Applications in Geography. 1st ed. Netherlands: Springer; 1994. 196 p.

[46] Chaloulakou A, Grivas G, Spyrellis N. Neural network and multiple regression models for PM10 prediction in Athens: a comparative assessment. Journal of the Air & Waste Management Association. 2003;53(10):1183–90.

[47] Boznar M, Lesjak M, Mlakar P. A neural network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain. Atmospheric Environment. Part B. Urban Atmosphere 1993;27(2):221–30.

[48] Gardner MW, Dorling SR. Regression modelling of hourly NO(x) and NO_2 concentrations in urban air in London. Atmospheric Environment. 1999;33:709–19.

[49] Hooyberghs J, Mensink C, Dumont G, Fierens F, Brasseur O. A neural network forecast for daily average PM10 concentrations in Belgium. Atmospheric Environment. 2005;39(18):3279–89.

[50] Cai M, Yin Y, Xie M. . Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. Transportation Research Part D: Transport and Environment. 2009;14(1):32–41. DOI: 10.1016/j.trd.2008.10.004

[51] Cortina-Januchs MG, Quintanilla-Dominguez J, Vega-Corona A, Andina D. Development of a model for forecasting of PM10 concentrations in Salamanca, Mexico. Atmospheric Pollution Research. 2015;6(4):626–34. DOI: 10.5094/APR.2015.071

[52] Taşpinar F. Improving artificial neural network model predictions of daily average PM10 concentrations by applying principle component analysis and implementing seasonal models. Journal of the Air & Waste Management Association. 2015;65:800–9. DOI: 10.1080/.

[53] Brunelli U, Piazza V, Pignato L, Sorbello F, Vitabile S. Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM10, NO_2, CO in the urban area of Palermo, Italy, Atmospheric Environment. 2007;41(14):2967–95.

[54] Feng Q, Wu S, Du Y, Xue H, Xiao F, Ban X, et al. Improving neural network prediction accuracy for PM10 individual air quality index pollution levels. Environmental Engineering Science. 2013;30(12):725–32. DOI: 10.1089/ees.

[55] Ishak AB, Moslah Z, Trabelsi A. Analysis and prediction of PM10 concentration levels in Tunisia using statistical learning approaches. Environmental and Ecological Statistics. 2016; 23:1-22. DOI: 10.1007/s10651-016-0349-8

[56] Tzima FA, Karatzas KD, Mitkas PA, Karathanasis S. Using data-mining techniques for PM10 forecasting in the metropolitan area of Thessaloniki, Greece. In: IEEE International Conference on Neural Networks; 2007. p. 2752–7.

[57] World Health Organization. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment [Internet]. 2006 . Available from: http://whqlibdoc.who.int/hq/2006/WHO_SDE_PHE_OEH_06.02_eng.pdf?ua=1 [Accessed: 2016-07-22]

[58] Appelhans T, Sturman A, Zawar-Reza, P. Modelling emission trends from non-constant time series of PM 10 concentrations in Christchurch, New Zealand . International Journal of Environment and Pollution 2010;43(4):354–63.

[59] Environment Canterbury. Data Catalogue [Internet]. 2016. Available from: http://data.ecan.govt.nz/

[60] Wilton E, Rijkenberg M, Bluett J. Assessing long-term trends in PM 10 concentrations in Blenheim [Internet]. 2010 . Available from: http://www.marlborough.govt.nz/Environment/Air-Quality/~/media/Files/MDC/Home/Environment/Air Quality/TrendsInAirQuality9.ashx [Accessed: 2016-07-21]

[61] Scarrott C, Reale M, Newell J. Statistical estimation and testing of trends in PM 10 concentrations: is Christchurch city likely to meet the NES target for PM 10 concentrations in 2013? [Internet]. 2013 . Available from: http://ecan.govt.nz/publications/Reports/PM10TrendsComplete.pdf [Accessed: 2016-07-22]

[62] Elangasinghe MA, Singhal N, Dirks KN, Salmond JA, Samarasinghe S. Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network

modelling and k-means clustering. Atmospheric Environment. 2014;94:106–16. DOI: 10.1016/j.atmosenv.2014.04.051

[63] Elangasinghe MA. Applications of semi-empirical and statistical techniques in urban air pollution modelling [thesis]. University of Auckland; 173 p. Available from: https://researchspace.auckland.ac.nz/handle/2292/23444

[64] Tate A. Wintertime PM10 measurements and modelling in Alexandra and Mosgiel, Otago, New Zealand [thesis]. University of Otago; 2012. 117 p. Available from: http://hdl.handle.net/10523/2255

[65] Xie S, Gimson N, Clarkson T. Modelling wintertime PM10 dispersion in Masterton, New Zealand: a tool for implementing national standards. WIT Transactions on Ecology and the Environment. 2006;86:65–74.

[66] Zawar-Reza P, Kingham S, Pearce J. Evaluation of a year-long dispersion modelling of PM10 using the mesoscale model TAPM for Christchurch, New Zealand. Science of the Total Environment. 2005;349(1–3):249–59.