STATISTICAL ANALYSIS OF PARTICULATE MATTER DATA IN DOHA, QATAR

CHARLES C. TAYLOR[1], ADIL E. YOUSIF[2] & KASSIM S. MWITONDI[3]
[1]University of Leeds, UK
[2]Qatar University, Qatar
[3]Sheffield Hallam University, UK

ABSTRACT

Pollution in Doha is measured using passive, active and automatic sampling. In this paper we consider data automatically sampled in which various pollutants were continually collected and analysed every hour. At each station the sample is analysed on-line and in real time and the data is stored within the analyser, or a separate logger so it can be downloaded remotely by a modem. The accuracy produced enables pollution episodes to be analysed in detail and related to traffic flows, meteorology and other variables. Data has been collected hourly over more than 6 years at 3 different locations, with measurements available for various pollutants – for example, ozone, nitrogen oxides, sulphur dioxide, carbon monoxide, THC, methane and particulate matter (PM1.0, PM2.5 and PM10), as well as meteorological data such as humidity, temperature, and wind speed and direction. Despite much care in the data collection process, the resultant data has long stretches of missing values, when the equipment has malfunctioned – often as a result of more extreme conditions. Our analysis is twofold. Firstly, we consider ways to "clean" the data, by imputing missing values, including identified outliers. The second aspect specifically considers prediction of each particulate (PM1.0, PM2.5 and PM10) 24 hours ahead, using current (and previous) pollution and meteorological data. In this case, we use vector autoregressive models, compare with decision trees and propose variable selection criteria which explicitly adapt to missing data. Our results show that the regression tree models, with no variable transformations, perform the best, and that attempts to impute missing values are hampered by non-random missingness.

Keywords: diurnal variation, missing data, multivariate linear regression, regression tree.

1 POLLUTION MONITORING IN DOHA

According to the World Health Organisation's Global Urban Ambient Air Pollution Database [1], which tabulates pollution levels in nearly 3,000 cities in over 100 countries, Doha's air is some of the most polluted in the world. Average levels (93 ug/m^3) of PM2.5 – small and fine particles – ranked 42nd on this list. Pollution in Doha is measured in three ways, from simple physical and chemical measurements, to sophisticated electronic methods, but they all involve measuring the ambient pollutant concentrations at a given place over a given period of time.

Passive sampling is the simplest and cheapest way to screen air quality, and gives a general indication of average pollution concentrations over a time period ranging from a week to several months. Active sampling methods (which are semi-automatic) use physical or chemical methods to collect pollutant samples. The common method involves pumping a known volume of air through a collector such as a filter or chemical solution for a known period of time. Using (continuous) automatic sampling methods, pollutants can be continually collected and analysed every hour (or more frequently). The sample is analysed on-line and in real time and the data is stored within the analyser, or a separate logger and may be downloaded remotely by a modem. The accuracy produced enables pollution episodes to be analysed in detail and related to traffic flows, meteorology and other variables. To ensure that the data produced is accurate and reliable a high standard of maintenance, calibration, operation and quality assurance/control procedures are required, making automatic sampling

WIT Transactions on Ecology and the Environment, Vol 230, © 2018 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR180101

methods the most expensive way routinely employed. Automatic techniques can be used to measure: ozone; oxides of nitrogen; sulphur dioxide; carbon monoxide; and PM10 particulates.

1.1 Objectives

Recent studies have established an explicit link between exposure to aerosols and increased rate of mortality and morbidity; see, for example, [2]. Our interest is to investigate the (relative) effects of meteorological and physicochemical factors on particulate matter in the city of Doha, and to consider also spatial and temporal variation. A methodologically related previous study, which used data from Kuala Lumpur, was able to decompose the meteorological effects from other pollution measurements through the use of principal components [3].

After describing our data and some issues connected to data cleaning in Section 2, we focus on the task of forecasting values of PM 24 hours ahead. This mimics a practical scenario, whereby if high levels of PM were predicted, then various types of action could be implemented in a timely manner. A simple time series model, which uses only time of day and day of the week, is considered in Section 3, and this serves as a benchmark. In Section 4 we give results for linear regression for models which are restricted to observations within location and/or only meteorological variables. The use of regression tree models is investigated in Section 5, with suggestions for fair comparisons to linear regression. In a final section, we discuss approaches to handling missing data, and compare results.

2 DATA USED IN THIS STUDY

The data used in this paper used automatic sampling methods. Measurements have been collected at three locations in Doha: “Aspire”, “Qatar University” (QU), and “Corniche”. These are shown on a map, with the longitude and latitude, in Fig. 1. The available data, recorded hourly, is from 1st January 2010 to midnight on 30th September 2016, i.e. 59,160 time points.

2.1 Summary of variables

Measurements are available for 13 pollutants and 5 types of meteorological data and, from the time points, we have the time of day and the day of the week. In the original data, there were many missing values; at each location there were only complete records for about 10% of the time points, and only 7% of the time points had complete data at all three locations. For the particulate matter (PM) observations – which is our main focus – nearly 40% of the PM1.0 and PM2.5 measurements were missing at each location, and PM10 measurements were missing at 8.5% (Corniche), 17% (Aspire) and 10% (QU). The full list of variables is:

Meteorological (5 variables)	air pressure (AP), temperature (TEMP), wind speed (WS), wind direction (WD), relative humidity (RH)
Pollutant (13 variables)	ozone (O3), nitric oxide (NO), nitrogen dioxide (NO2), nitrogen oxides (NOx), sulphur dioxide (SO2), carbon monoxide (CO), total hydrocarbon (THC), methane (CH4), hydrogen sulphide (H2S), non-methane hydrocarbon (NMHC) and particulate matter (PM1.0, PM2.5, PM10)

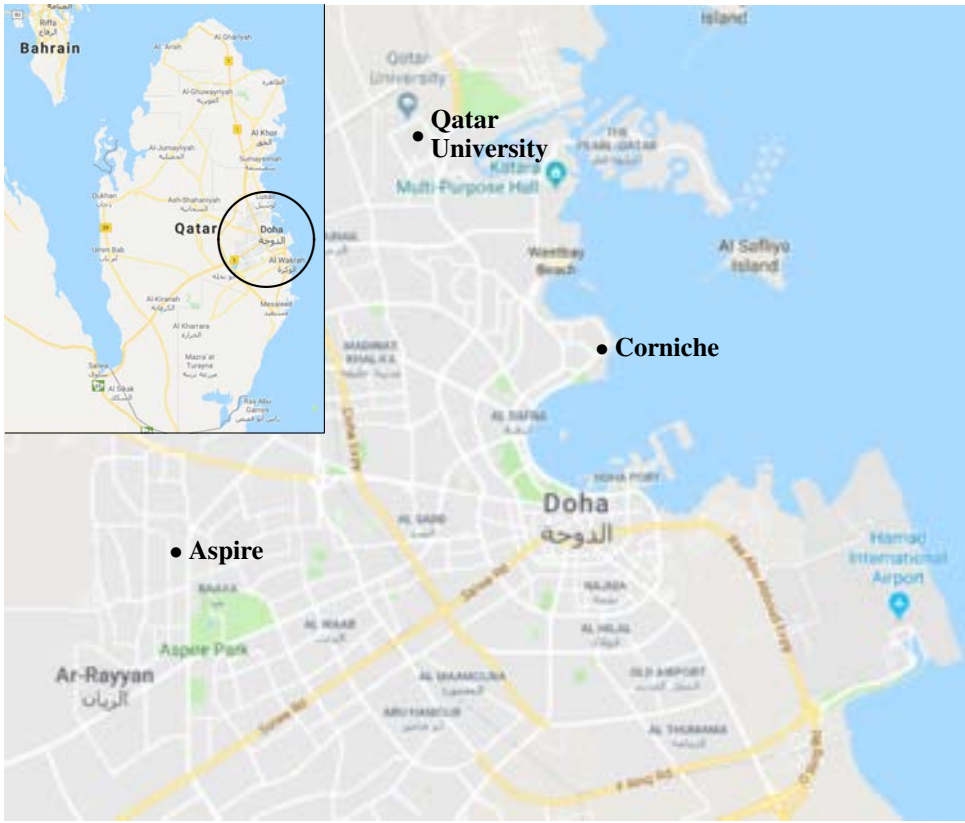

Figure 1: Location of monitoring stations (longitude, latitude): Aspire (25.28, 51.43), Qatar University (25.38, 51.49), and Corniche (25.33,51.54). *(Source: Map data ©2017 Google.)*

Several variables/locations were missing for more than 70% of the time points: Aspire measurements for all meteorological data, and all measurements for H2S. These were simply excluded from any further analysis, leaving 5 meteorological variables and 12 pollutants.

2.2 Data cleaning

For such data, only temperature can take negative values. However, given the lowest recorded temperature in Qatar is 1.5° all of the variables *should* take non-negative values, so we first replaced all negative values by "NA" (missing). This affected 18 of the $3 \times (12 + 5) = 51$ variables, and about 0.05% of the measured observations. Air pressure measurements had values outside the normal range, so any observations outside the range (870,1087) were replaced by NA. Pairwise scatter plots (using the three locations) were used to identify other outliers, and wind speeds above 40 m/s were identified and also replaced by NA.

Many variables had a excess of values which were equal to zero, suggesting that this may be an erroneous way of recording an invalid (or missing) observation. Clearly, zero is *valid* as a value, but it would not be expected to occur many more times than the next smallest observation. For example, a frequency table of the PM1.0 values (Corniche) starts with 0

(1590), 0.1 (71), 0.2 (64), 0.3 (88), ... In such a large dataset, a semi-automated rule for handling such instances is given by the following. If $x_1 < x_2 < \cdots$ are the ordered possible values taken by a variable, each occurring with corresponding frequency $f_1, f_2, \ldots$, then –in the case that $x_1 = 0$ and $f_1 > f_2$ (as in the above example) – we replace all such cases of x_1 by NA. Of course, some of these instances would have been valid, but we have no easy way to know which ones.

2.3 Transformations

Even after the cleaning described above, the distribution of several variables had very long tails. In such cases where the skewness of a variable exceeded 5 (at all 3 locations), we considered taking a log transform (with 0.0001 added to the data, to deal with observations equal to zero). This circumvented the issue of trying to decide a threshold for the high outliers, and made those observations less influential in the fitted models. The list of variables considered for log-transformation were: PM1.0, PM2.5, PM10, NO, SO2, CO, THC, NMHC.

For the directional variable (wind direction), we took a sin and cos transform. We also wanted to treat time of day as a discrete quantitative variable (rather than as an ordered categorical variable) so we replaced the time of day ($t \in \{0, 1, \ldots, 23\}$) by the pair (c, s) where $c = \cos(2\pi t/24)$ and $s = \sin(2\pi t/24)$. In addition, to allow for higher frequency diurnal variations we created new variables $c2 = \cos(4\pi t/24)$ and $s2 = \sin(4\pi t/24)$. The days of the week were simply treated as factors (categorical values).

3 VARIATION BY DAY OF WEEK AND TIME OF DAY

We focus on modelling PM (1.0, 2.5 and 10) at each of the three locations. In this section, we consider models which predict the log PMx values using only the time of day, and day of the week. Using multiple regression models, we find that day of the week is an important variable, and that there is a significant *interaction* with the time of day, c, and day of the week. These interaction terms have high significance in Corniche (p-values around 10^{-6}), but even more significant at the other two locations. Moreover, the second order ($s2$) terms were usually statistically significant. However, it should be noted that none of these simple models fits very well, with multiple correlation coefficients (R^2) only around 1%; this value becoming larger for the bigger particulate matter.

The fitted models for the log transformed data are shown in Fig. 2 in which the mean values at each time/day are compared to the fitted model values at each location, and for each PM variable. In the figure we can clearly see that the PM levels are lowest at Corniche, and highest at QU (for all particulates). We can also note that Friday has the lowest values, and that the PM Value usually peaks around 8–9 am with a secondary peak around 11 pm (which is least prominent for PM10) with the lowest values occurring around 2 pm. These findings are consistent with working patterns in Doha.

Alternative models, in which separate values were estimated for each time point did fit better than the trigonometric models with a day interaction. But with 168 degrees of freedom the adjusted R-squared values were not competitive, so it was decided to only use trigonometric terms and the day of the week to capture this component of variation.

4 FORECASTING PM VALUES USING LINEAR REGRESSION

Clearly, high correlations exist throughout the data. As expected PMx are correlated between locations, as are the various meteorological variables. Within a location, the largest correlations are to be found between PM1.0, PM2.5 and PM10, as well as between CH4 and THC, NO2 and NOx; correlations between the meteorological variables and the pollution

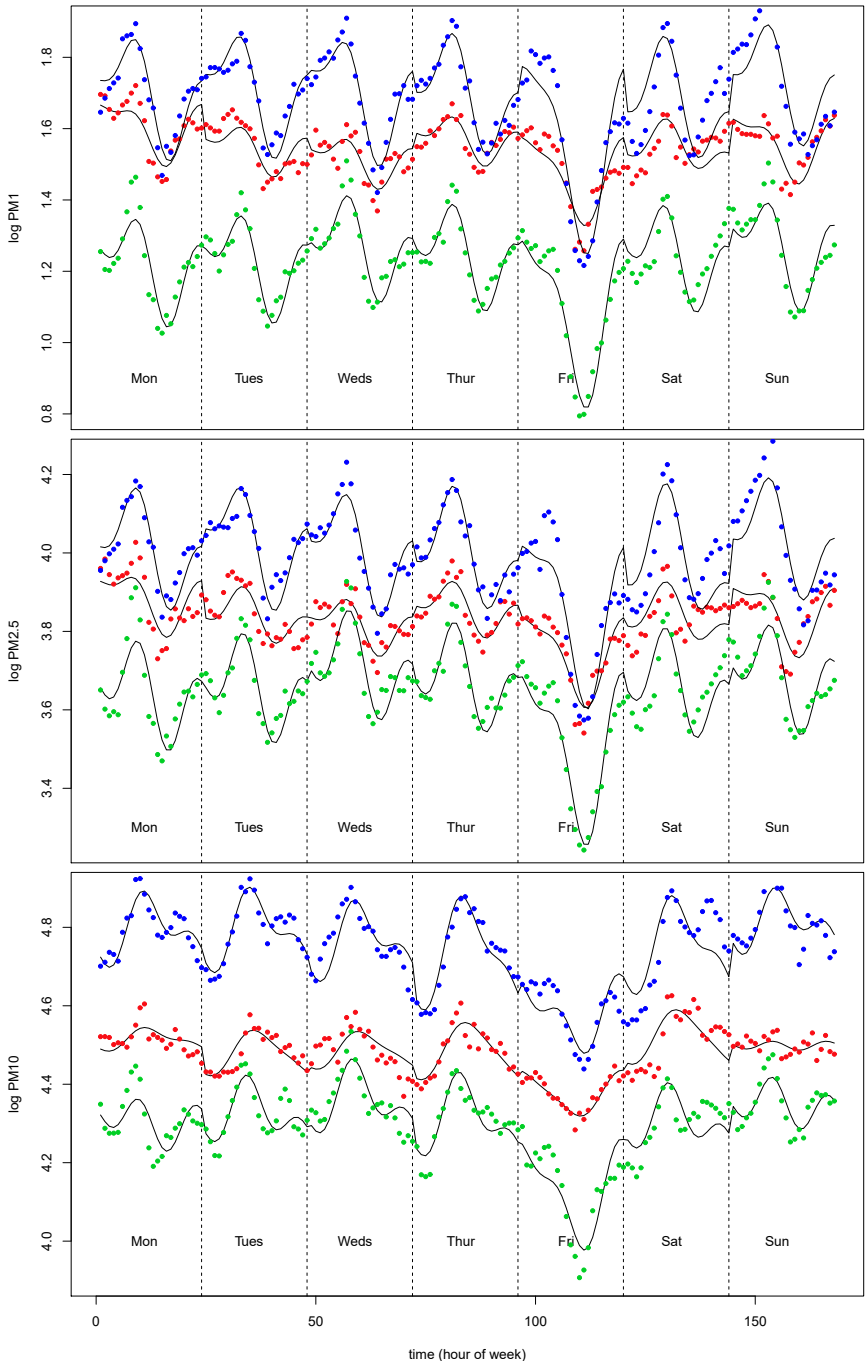

Figure 2: Estimations of log(PM) (top: PM1; middle: PM2.5; bottom: PM10) using additive trigonometric model (lines) with averages (points) shown by time of day and day of week for Corniche (red), Aspire (green) and QU (blue).

variables tend to be much lower. Also, we note that many of the variables are correlated with the time of day, so including this may reveal *partial* correlations which are nevertheless useful.

Throughout the rest of this paper we consider the task of forecasting 24 hours ahead, using current and previous data. In Section 4.1 we carry out predictions at each location using only data from that location, initially using only using the PM data and meteorological data. Obviously we know the time of day, and the day of the week for which the forecast is being made. In real life, there may also be more accurate information about weather forecasts, but here we suppose that nothing else is known, and use only "current" information to ascertain the relationships. In Section 4.2 we include all other pollution data as explanatory variables, still focusing on one location at a time. In Section 4.3 we also use information (both PMx and pollution data) from other locations, and in Section 4.4 we use all the data and compare some of the selected models.

Throughout this section we consider selecting amongst the variables, in order to try to obtain a parsimonious solution which can lead to interpretation and insights. However, the missing data makes this tricky; standard methods would simply remove all rows with any missing data, and then use backward selection procedures. This would ignore the eventuality that after a variable has been removed, then more observations may become complete. The solution is to reconsider the completeness (or missingness) after each variable removal, which slows down the process, but is otherwise quite tractable. When using multiple regression, we fixed the threshold of the p-value for removal/inclusion at 0.001 in order to combat the multiple testing pitfalls encountered when choosing amongst so many models.

4.1 Within location using meteorological data

Since we have removed all meteorological variables related to the Aspire location, here the model selection was restricted only to the PMx variables and time $\times$ day information. In this case, prediction of any of the PMx variables used current values of all three of the PMx data (with PM2.5 being the least useful, and with a negative coefficient, whereas both the other variables were positive) as well as including time, day and their interaction. The resulting multiple R^2 values for all three PMx models were around 40%.

At the Corniche and QU locations, the most important meteorological factor was temperature, with higher temperatures associated with higher PMx readings – this is after allowing for the time of day (which will already give an indication of temperature). Time of day, and day of the week were also very important (though the interaction term was not present at Corniche for PM2.5 and PM10). At QU, humidity and air pressure were also selected in the final models for all three PMx models – these were both negatively associated with the response variable. Wind direction was also seen to be important: in Fig. 3 we show the wind directions at which the maximum PMx values were obtained, along with a rose diagram indicating the general distribution of wind directions at Corniche and QU. It is perhaps fortuitous that the "bad" wind directions occur infrequently. As may be expected, higher wind speeds are observed with lower PMx values.

4.2 Within location using all data

The (possible) inclusion of more variables does improve the fit: there was an increase in the (adjusted) R^2 value by about 5.2% (as a ratio), and the predictive performance (as measured by mean squared error) was smaller by nearly 4%. However, the correlations between the explanatory variables make the final models hard to interpret. It is clear that the most

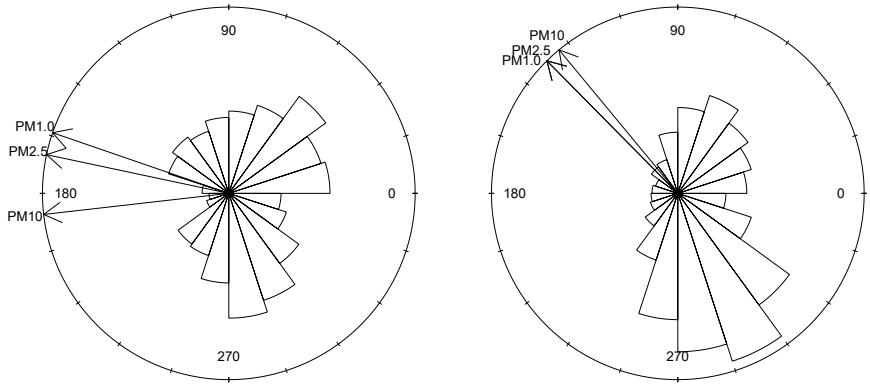

Figure 3: Rose diagram of wind directions at Corniche (left) and QU (right), with arrows indicating those directions at which the corresponding PMx values were maximized (with all other variables held fixed). The radii scales are linear.

important pollution variables are NO and NO2, and that – even after inclusion of relevant pollution data – the effect of temperature, and wind speed and direction remain the same. However, the fact that ozone (O3) is negatively correlated with the nitride variables leads to a change in the sign of some coefficients dependent on whether O3 is included or not.

4.3 Using meteorological data from other sites

We extend Section 4.1 to include explanatory variables from all three sites. In Fig. 4 we indicate which explanatory variables (in addition to time of day and day of the week) were included in the final linear regression models. In this figure the explanatory variables are grouped by location and type; recall that the meteorological data was mostly missing from Aspire (and so not considered). We can note that more variables were selected for log transformed response variable models, and in general the meteorological data are often included, even from other locations.

4.4 Using all meteorological and pollution data

We show which variables were selected in Fig. 5, continuing to use a threshold of 0.1% significance level for inclusion. For all 9 models, the PM variables were seen to be most useful, particularly within location. With the exception of PM1 at QU, the "same" variable as the response variable was always selected as explanatory variable, and day of the week and time of day were included in each model. Meteorological variables were still present, with temperature and air pressure most prevalent. Amongst the pollution variables, SO2 at Aspire was always picked, with Aspire variables being generally more useful than those from QU. This could have implications for optimising the location of monitoring sites in the future.

4.5 Comparing models, including log transformation

Comparing two methods when there are missing values in the data can be problematic. Firstly, when working on a day-to-day basis, if one of the explanatory variables is missing, it is

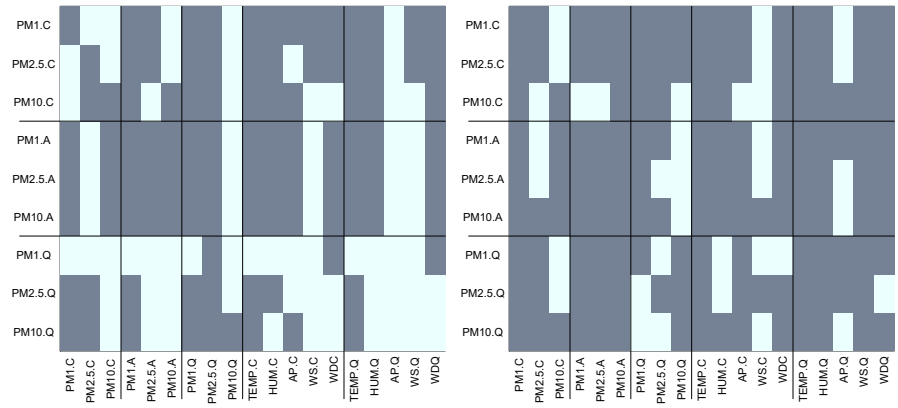

Figure 4: Variables selected in 9 linear regression models based on PMx, and meteorological data only. y-axis shows response variables, with shade corresponding to variables selected. Left uses raw data, and right uses log transformed data.

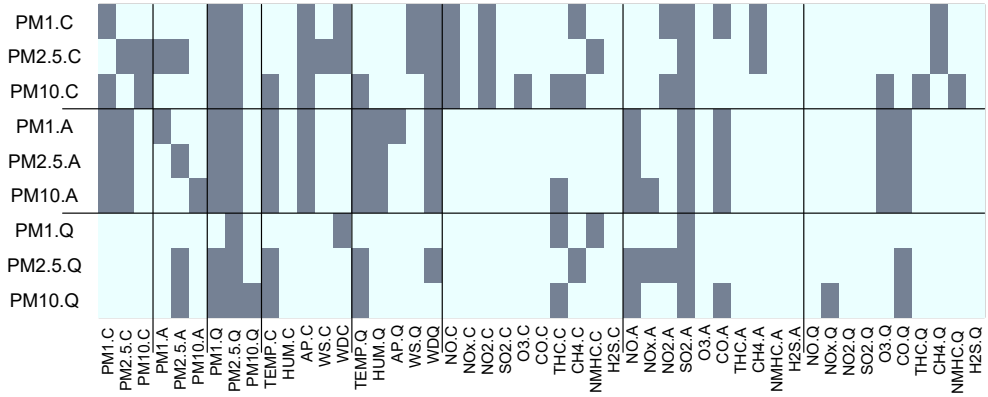

Figure 5: Variables selected in 9 linear regression models based on PMx, and all variables. The y-axis shows response variables, with shade corresponding to variables selected. No transformations used.

unsatisfactory to declare that a forecast for tomorrow is not possible. Also, when determining which method to use (from the training data) it is important to try to make *fair* comparisons.

To illustrate the issue we use, as an example, the problem of model selection when forecasting PM1 at the QU site, when pollution data is also available from other sites as explanatory variables. Using the log transformed data (as described in Section 2.3) to fit the model, and carry out variable selection, the mean-squared prediction error (MSPE) – on the original scale – is 370.8. In this case, 25 meteorological and pollution variables were selected, as well as time of day, day of the week and their interaction term. This large number of selected variables resulted in 38,613 (65%) observations being removed due to missingness. Conversely, using the original data (without any transformations) to fit the model, the mean squared prediction error is 25,710.2 – more than 50 times larger. Here, only 5 pollution variables were selected, as well as time of day, day of the week and their interaction. With fewer variables selected, only 26,882 (45%) observations were deleted due to missingness.

Table 1: ER values – see eqn (1) – for each selected model according to whether the original data (a), or log-transformed data (b) was used. Rows are numbered (in pairs) with explanatory variables selected from: 1 meteorological data within location; 2 all data within location; 3 meteorological data from all sites; and 4 all data from all sites.

		Corniche			Aspire			QU		
	PM	1.0	2.5	10	1.0	2.5	10	1.0	2.5	10
1	(a)	0.05	0.04	0.06	0.06	0.06	0.06	0.01	0.03	0.07
	(b)	0.03	0.02	0.04	0.08	0.08	0.08	0.00	0.02	0.07
2	(a)	0.05	0.04	0.06	0.08	0.07	0.06	0.01	0.04	0.07
	(b)	0.03	0.02	0.04	0.09	0.08	0.09	0.00	0.02	0.07
3	(a)	0.06	0.05	0.07	0.11	0.11	0.11	0.01	0.04	0.07
	(b)	0.03	0.02	0.04	0.09	0.09	0.10	0.00	0.02	0.05
4	(a)	0.07	0.07	0.08	0.13	0.12	0.12	0.01	0.04	0.07
	(b)	0.03	0.03	0.05	0.10	0.09	0.10	0.05	0.08	0.09

At first, it appears that the log transformation has been hugely beneficial to the prediction problem, but it could be that the additional observations which were deleted were (in some sense) harder to predict. A suggested way to compensate for this is to also measure the variability of the response variable over the observations which were not removed. In a sense, this variance is the same as simply using the mean value for predictions, with no explanatory variables (intercept only model), and so provides a benchmark. In this example, the variance of the non-deleted observations in log-transformed model is 391.1, and in the untransformed data model is 25,884.6 – again more than 50 times larger. This motivates a measure – to be maximized across models – given by

$$\text{ER} = 1 - \frac{\text{MSPE}}{\text{var}(y)} = 1 - \frac{\sum_{i \in L}(y_i - \hat{y}_i)^2}{\sum_{i \in L}(y_i - \bar{y})^2}, \tag{1}$$

where both the numerator and denominator are evaluated using the non-deleted (and non-missing) observations (L). In this example, the above ratio is 0.05 using the log transformed data, and 0.01 for the original data. Although the first is still better, the difference is now much less. Note that eqn (1) is very similar to the usual R-squared summary obtained in a regression model, one difference being that the above is consistently applied to the original scale, and so any transformation must be back-transformed before evaluation.

We have found that using the log transformation helped with the regression diagnostics, but with regard to prediction (on the original scale) – arguing as above – we found that using the raw data was slightly better overall; see Table 1 for a comparison across the 9 fitted models with restrictions of the explanatory variables according to Sections 4.1–4.4. None of these models does very well.

5 REGRESSION TREES

Regression (and classification) trees [4] have been developed to provide predictions for regression data, which can work particularly well when the explanatory variables are a mix of numerical and categorical, or with missing data, and can lead to models which are much easier to interpret than linear regression models. Conversely, compared with linear regression

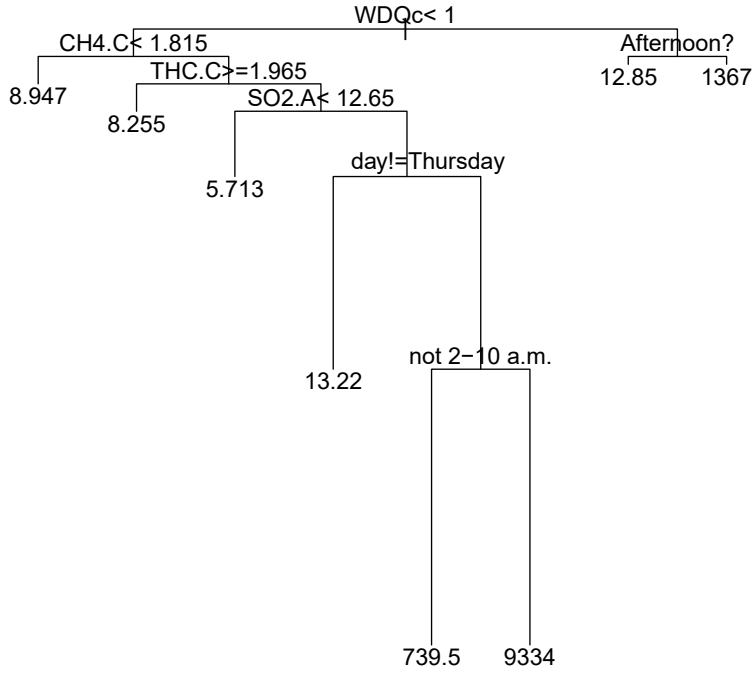

Figure 6: Regression tree for PM1 at QU. Left branches are used when the condition at the node is TRUE. Values at the final steps (leaves) indicate the fitted values (of PM1) according the conditions specified. WDQc indicates the cosine of the wind direction at QU, and the length of the branches indicates the improvement made (in reducing the variance) as a result of that split. No data transformations used.

models, the distributional theory is much less mature and obtaining meaningful prediction confidence intervals is not straightforward.

The procedure recursively splits the data by partitioning the variable space. In this case, it considers using "split points" between each value of the ordered data for each variable. So, for distinct data, this would initially consider a total of $(n-1) \times p$ split points, where n is the number of observations and p the number of variables. The choice of split is taken to maximize some measure of homogeneity (for example variance) of the response variable at the two resulting "nodes". The process continues until the resulting nodes are not worth splitting further, and the average value of the response variable is then used as a prediction value for any (new) data which meets all the conditions fulfilled by this terminal node (leaf).

We illustrate a fitted regression tree model in Fig. 6 for the case of predicting PM1 at QU, using all the explanatory variables (as in Section 4.4). There are 6 variables selected here, with time of day, and day of the week both featuring prominently. The first split (WDQc < 1) is rather strange; only 33 observations failed to meet this condition (wind direction not in the interval $(-0.3°, 0.3°)$) and the interpretation is not clear. However, the remainder of the splits (many of which have more impact) make much more sense, and this interpretability is often seen as an advantage of regression trees. A second advantage, which is useful for our many missing values, is that – in the event that the required explanatory variable is missing at a given split point – other "surrogate" variables can be used. This means that predictions are available far more often than in our regression models. In this example, the

Table 2: ER values – see eqn (1) – for each selected model using a regression tree. Rows are numbered with explanatory variables selected from: 1 meteorological data within location; 2 all data within location; 3 meteorological data from all sites; and 4 all data from all sites.

	Corniche			Aspire			QU		
PM	1.0	2.5	10	1.0	2.5	10	1.0	2.5	10
1	0.25	0.28	0.06	0.06	0.10	0.09	0.32	0.24	0.07
2	0.28	0.32	0.13	0.14	0.10	0.09	0.28	0.21	0.28
3	0.31	0.40	0.08	0.33	0.17	0.12	0.25	0.26	0.24
4	0.33	0.42	0.20	0.36	0.17	0.15	0.85	0.56	0.34

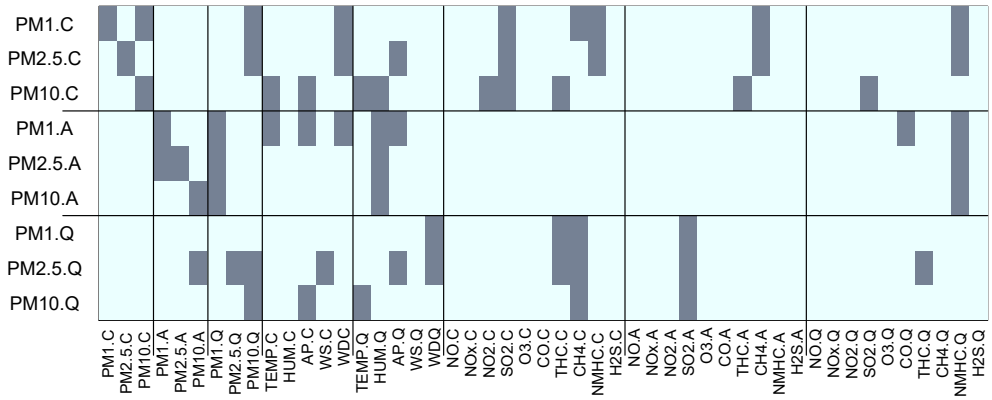

Figure 7: Variables selected in 9 regression tree models based on PMx, and all data. y-axis shows response variables, with shade corresponding to variables selected using the R package `rpart`. No data transformations used.

mean-squared prediction error (MSPE) is 3194, the variance of the actual response values is 21,873, resulting in a value of 0.85 using eqn (1).

By comparing Table 2 with Table 1, it can be seen that, overall, regression trees have better predictive capability than the linear models for these data. Moreover, as can be seen in Fig. 7 (which can be compared with Fig. 5) fewer variables are required, and the resulting models are easier to interpret. The main drawback remains that confidence intervals for the predictions are harder to obtain.

6 MISSING DATA

Our anaylses have been hampered by missing data, which have made methods such as principal component regression (PCR) or partial least squares regression (PLSR) harder to use. In this final section we consider the effect of using multiple imputation methods on the model selection process, and the resulting goodness of fit. Various methods exist for imputing missing data, but here we use multivariate imputation using chained equations [5, 6] and its implementation in R [7]. Briefly, this method uses Gibbs sampling, in which missing data for each variable (in turn) is generated (by "predictive mean matching") using other columns in the data. For predictors that have missing data, the most recently generated imputations are

used to complete the predictors before imputation of the target variable, and the process is iteratively applied leading to as many multiple imputation datasets as desired.

We continue to exclude variables for which more than half of the observations are missing, and consider imputation only of the remaining data. For each prediction problem, we generated 5 imputed sets for the missing values, and for each resultant (full) dataset, we used stepwise regression, with AIC as the model selection criterion. Then, we selected the variables which appeared in *all* 5 models and used these for prediction, finally taking the average of the predicted values. We are now able to obtain predictions at every time point, though obviously can only make comparisons with observed data.

This approach is intended to work well when the missing data values are *missing at random*. If a missing value is caused by a malfunction in equipment, this is more likely to occur in more extreme situations, so missing values will not occur at random. We believe it is for this reason that the results (as measured by ER) using models fitted with the imputed data were somewhat *worse* than models based on only observed data. This suggests that any attempts to use PCR or PLSR are also unlikely to be successful.

Finally, we note that our imputations are based on a cross-sectional approach (effectively treating the data as i.i.d.), and a potential alternative is to impute missing values based on forecasting (or backcasting) fitted time series models. However, in our case, with so many missing values – and so many of these being contiguous – this would be difficult to carry out in practice. An area for future research would be to consider a hybrid approach.

ACKNOWLEDGEMENT

This research was supported by the Qatar National Research Fund (NPRP 7–897–1–165).

REFERENCES

[1] World Health Organization Global Urban Ambient Air Pollution Database (update 2016). www.who.int/phe/health_topics/outdoorair/databases/cities/en/. Accessed on: 16 Feb. 2018.

[2] Wan Mahiyuddin, W.R., Sahani, M., Aripin, R., Latif, M.T., Thach, T.-Q. & Wong, C.-M., Short-term effects of daily air pollution on mortality. *Atmos. Environ.*, **65**, pp. 69–79, 2013.

[3] Khan, F., Latif, M.T., Juneng, L., Amil, N., Mohd Nadzir, M.S. & Syedul Hoque, H.M., Physicochemical factors and sources of particulate matter at residential urban environment in Kuala Lumpur. *J. Air Waste Manage. Assoc.*, **65**, pp. 958–69, 2015, doi: 10.1080/10962247.2015.1042094.

[4] Breiman, L., Friedman, J., Stone, C.J. & Olshen, R.A., *Classification and Regression Trees*, Chapman & Hall/CRC Press: Boca Raton, FL, 1984.

[5] van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn C.G.M. & Rubin, D.B., Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.*, **76**, pp. 1049–1064, 2006.

[6] van Buuren, S., *Flexible Imputation of Missing Data*, Chapman & Hall/CRC Press: Boca Raton, FL, 2012.

[7] van Buuren, S. & Groothuis-Oudshoorn, K., Mice: multivariate imputation by chained equations in R. *J. Stat. Softw.*, **45**, pp. 1–67, 2011.