Multivariate Time Series Modelling for Urban Air Quality

Hajar Hajmohammadi[a,*1] and Benjamin Heydecker[a]

[a] Centre for Transport Studies (CTS), Department of Civil, Environment and Geomatic Engineering, University College London, Gower Street, London, UK

Abstract: We introduce a spatio-temporal model to represent development of atmospheric pollution in an urban area. An important element of this is that recorded measurements are often incomplete which undermines time-series approaches. We identify the multiple imputation by chained equation (MICE) method as effective to complete data sequences synthetically. Following on from this, we develop a vector autoregressive moving average (VARMA) model for the spatio-temporal development of atmospheric pollution in urban areas. This model was fitted to hourly measurements of four pollutants (NO, NO_2, NO_x and PM_{10}) for the whole of the calendar year 2017 at 30 stations across London, completed by MICE as required. We show by cross-validation that the VARMA model is more effective than other formulations, including the Kriging method of spatial interpolation, and seasonal ARMA models for individual stations with either daily or weekly trends. The resulting model can be used for prediction of air quality in different periods and as the basis for assessment of policy interventions such as increasing vehicle emission standards, and traffic management and control policies such as low and ultra-low emission zones.

Keywords: Air Quality Modelling, Multivariate Time Series Analysis, Spatio-Temporal Analysis, Vector Time-Series Imputation

[1] Corresponding Author, h.hajmohammadi@ucl.ac.uk

1. Introduction

Air pollution is largely known as a crucial issue for management and mitigation which is responsible for over 8.5 million deaths annually worldwide (Orach et al., 2021). A wide range of research has established the strong adverse association between air pollution and public health, such as increases risk of hospital admissions for patients with asthma (Pfeffer et al., 2020), as well as negative impacts on anthropogenic ecosystems (Ochoa-Hueso et al., 2017) and climate (Shindell et al., 2009).

In London, several regulations and policies have been made in the road transport sector, which is recognised as a major source of urban air pollution, to improve air quality. These policies include establishing a low emission zone (LEZ) in 2008 and an ultra-low emission zone (ULEZ) in 2019. These zones in London restrict toll-free access to vehicles that comply with Euro 4 (petrol), Euro 6 (diesel), or better standards. However, concentrations of atmospheric pollutants such as Nitrogen dioxide (NO_2) in several parts of London still exceed the annual mean EU Limit Values of 40 μgm^{-3} (*Air Quality Standards*, 2009). Figure 1 shows the annual average NO_2 concentration in London in 2016 (*London Environment Strategy*, 2018). As is clear from this, most of the areas within and close to central London, as well as major roads to central London, exceed the EU limit.

While controlling air pollution is a costly long-term process, the impacts of different policies on air quality could be investigated relatively easily by modelling air quality: In these models, trends in pollutants, effective factors and interventions are investigated for predicting the future state of air quality.

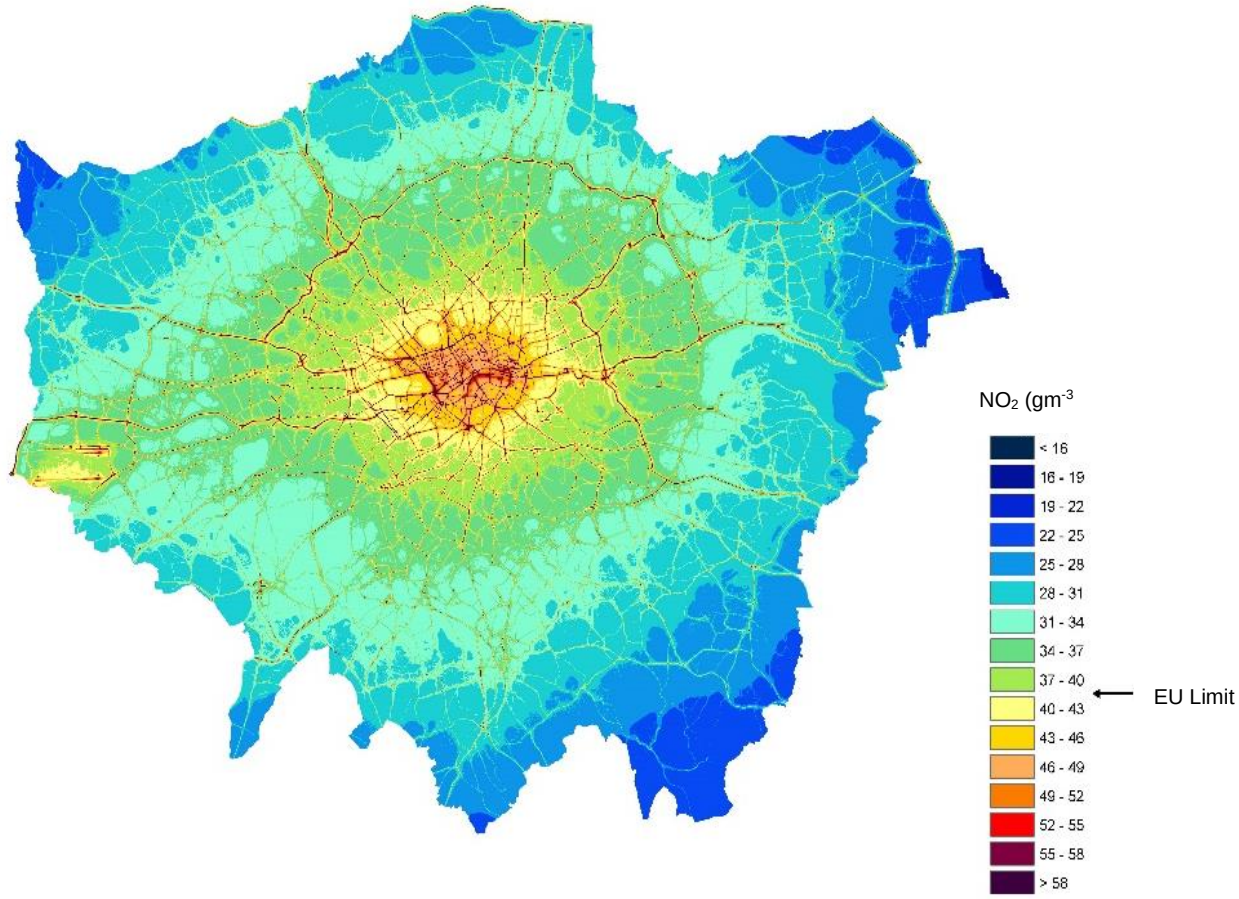


Figure 1: Annual average NO_2 concentrations in 2016 in Greater London (London Environment Strategy, 2018)

Several different statistical modelling approaches have been applied at varying temporal and spatial resolutions of air quality dataset. These range from a simple autoregressive moving average (ARMA) model (Kumar and Goyal, 2011), a non-stationary time series analysis for daily and weekly logarithmically transformed NO concentrations (Romanowicz et al., 2006), and multiple linear regression (Donnelly et al., 2015) to more complicated models such as hierarchical dynamic linear models of daily values of pollutant concentrations (Shaddick and Wakefield, 2002) and (Xu et al., 2016; Young et al., 2015). In these models, the spatio-temporal dependencies for different pollutants were developed using a Bayesian approach (Lee, 2007) and implementation of Markov Monte Carlo simulation (Gamerman, 2006). However, the study duration and observation sites of these researches were limited.

Deep learning modelling (Schmidhuber, 2015) is another approach implemented by Du et al. (2019; 2020) and Xie (2017) for predicting air quality in different cities of China. Freeman et al. (2018) integrated this approach with time series analysis to predict 8 hour averaged surface

ozone (O_3) concentrations in Kuwait. They developed a recurrent neural network (RNN) with long-short term memory that has around 17000 fitted parameters.

The present study develops a vector autoregressive moving average (VARMA) model for the spatio-temporal development of atmospheric pollution in urban areas. This model has the advantage of using a joint vector of hourly observations at multiple stations: we apply this to data from 28 stations across London. Hence, not only the temporal dependency but also the relationship between stations for each pollutant species will be investigated. Time series modelling requires continuous data whilst observed data series typically have breaks for several practical reasons. To synthesise complete sequences, several methods of imputations are investigated and compared, among which the multiple imputation by chained equation (MICE) (Resche-Rigon and White, 2018) method was identified as the best approach for the present application. The performance of the proposed VARMA model was tested by cross-validation in estimating data at each of a set of reserved sites. The results of this were shown to be substantially better compared to the performance of a fully-fitted instantaneous spatial Kriging method.

The rest of the paper is organized as follows: in section 2, the London air quality dataset is described and tested for regional effect by analysis of variance (ANOVA). Imputation of missing values by various methods is compared and the preferred MICE method is presented. Section 3 is devoted to spatio-temporal modelling methodology, including presentation of the VARMA model, its evaluation and validation. In section 4, the results of fitting the VARMA model are presented. Section 5 and 6 are devoted to discussion and conclusion, respectively.

2. Data Analysis and Preparation

In this section, the London air quality dataset used in this study is presented, together with a statistical analysis of its characterises. This includes an analysis of variance (ANOVA) and an imputation method for estimating missing values in the dataset to complete the time series.

2.1. London air quality dataset

A total of 30 stations across London were selected with 13 in the central London congestion charging zone (located within the London inner ring road, thus including the City of London and the West End), 9 in the remainder of the London low emission zone (LEZ) and 8 in greater London (outside of the LEZ). Figure 2 shows the location of each station, along with the boundaries of central London and the LEZ. This dataset contains hourly measurements of Nitric oxide (NO), Nitrogen dioxide (NO_2), oxides of Nitrogen (NO_x) and Particulate Matter with diameter less than 10 micrometres (PM_{10}) for the calendar year 2017 (8760 hours) during which no major air quality measures were implemented.

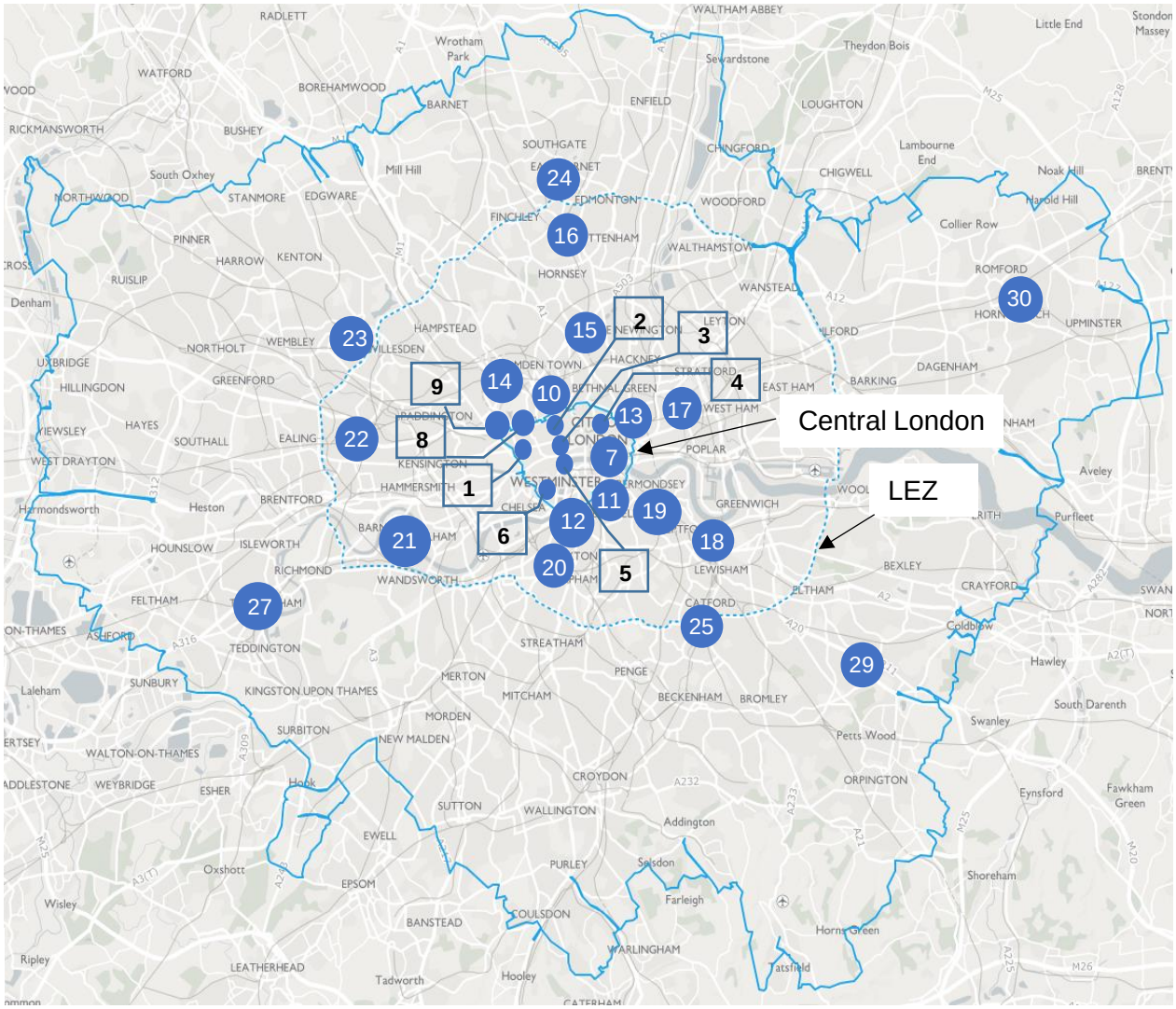


Figure 2: Location of the stations across London with their code

These pollution measurements were extracted from the London Air Quality Network (LAQN) (Environmental Research Group Kings College London, 2016). The code (assigned by

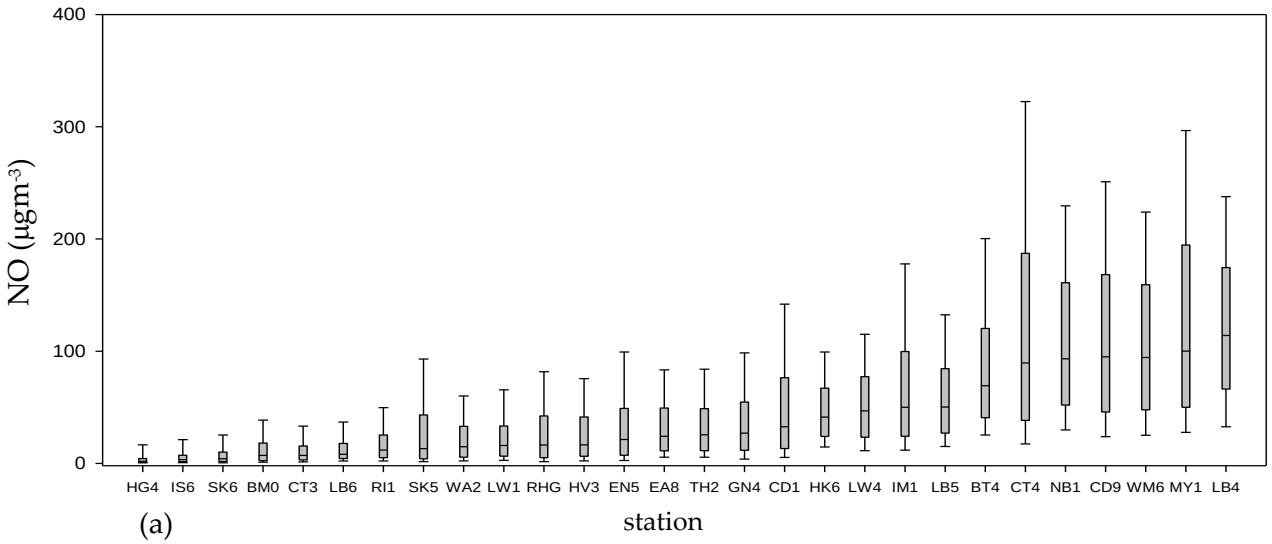
LAQN), location and pollutant measurement availability of each station are presented in Table 1.

Table 1: London Air Quality Network (LAQN) code, location and data availability in the studied stations

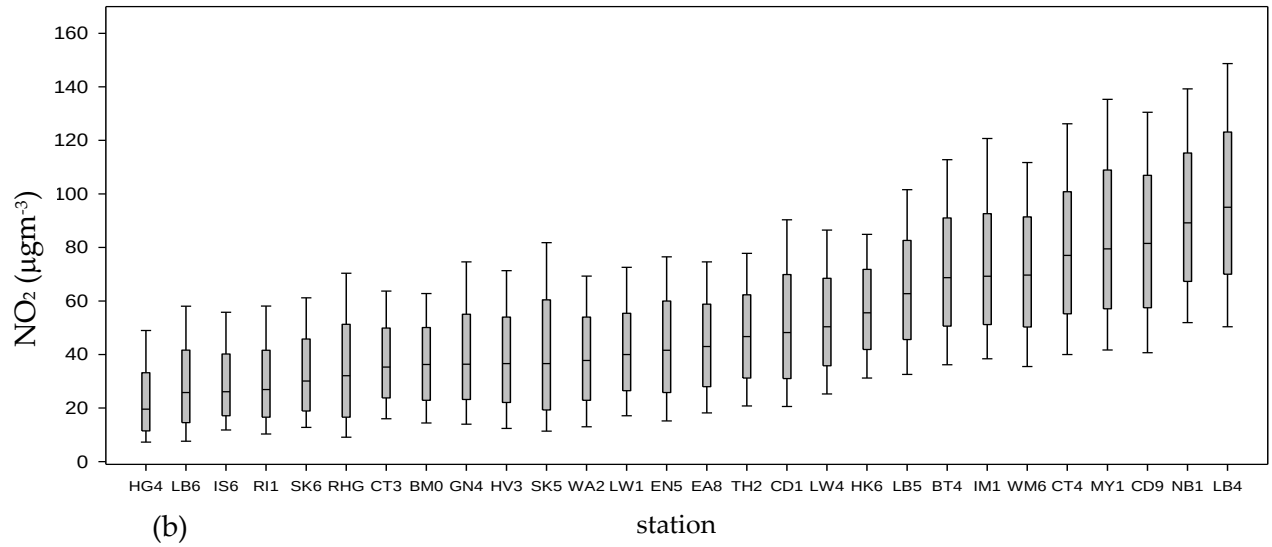
Number	LAQN code	Location	Available Measurements			
			NO	NO_2	NO_x	PM_{10}
1	WM6	Westminster - Oxford Street	✓	✓	✓	-
2	BM0	Camden - Bloomsbury	✓	✓	✓	✓
3	IM1	Camden - Holborn (Bee Midtown)	✓	✓	✓	-
4	CT4	City of London - Beech Street	✓	✓	✓	✓
5	NB1	Westminster - Strand (Northbank BID)	✓	✓	✓	-
6	WM0	Westminster - Horseferry Road	✓	✓	✓	✓
7	CT3	City of London - Sir John Cass School	-	-	-	✓
8	MY1	Westminster - Marylebone Road	✓	✓	✓	✓
9	MY7	Westminster - Marylebone Road FDMS	✓	✓	✓	✓
10	CD9	Camden - Euston Road	✓	✓	✓	✓
11	SK6	Southwark - Elephant and Castle	✓	✓	✓	✓
12	LB5	Lambeth - Bondway Interchange	✓	✓	✓	✓
13	HK6	Hackney - Old Street	✓	✓	✓	✓
14	CD1	Camden - Swiss Cottage	✓	✓	✓	✓
15	IS6	Islington - Arsenal	✓	✓	✓	✓
16	HG4	Haringey - Priory Park South	✓	✓	✓	-
17	TH2	Tower Hamlets - Mile End Road	✓	✓	✓	-
18	LW4	Lewisham - Loampit Vale	✓	✓	✓	✓
19	SK5	Southwark - A2 Old Kent Road	✓	✓	✓	✓
20	LB4	Lambeth - Brixton Road	✓	✓	✓	✓
21	RI1	Richmond Upon Thames - Castelnau	✓	✓	✓	✓
22	EA8	Ealing - Horn Lane	✓	✓	✓	✓
23	BT4	Brent - Ikea	✓	✓	✓	✓
24	EN5	Enfield - Bowes Primary School	✓	✓	✓	✓
25	LW1	Lewisham - Catford		✓	✓	-
26	WA2	Wandsworth - Wandsworth Town Hall	✓	✓	✓	-
27	RHG	Richmond Upon Thames - Chertsey Road	✓	✓	✓	✓
28	LB6	Lambeth - Streatham Green	✓	✓	✓	✓
29	GN4	Greenwich - Fiveways Sidcup Rd A20	✓	✓	✓	✓
30	HV3	Havering - Romford	✓	✓	✓	✓

The distribution of NO, NO_2, NO_x and PM_{10} measurements at each station are presented as box plots in Figure 3. In this figure, the stations are arranged in increasing order of median

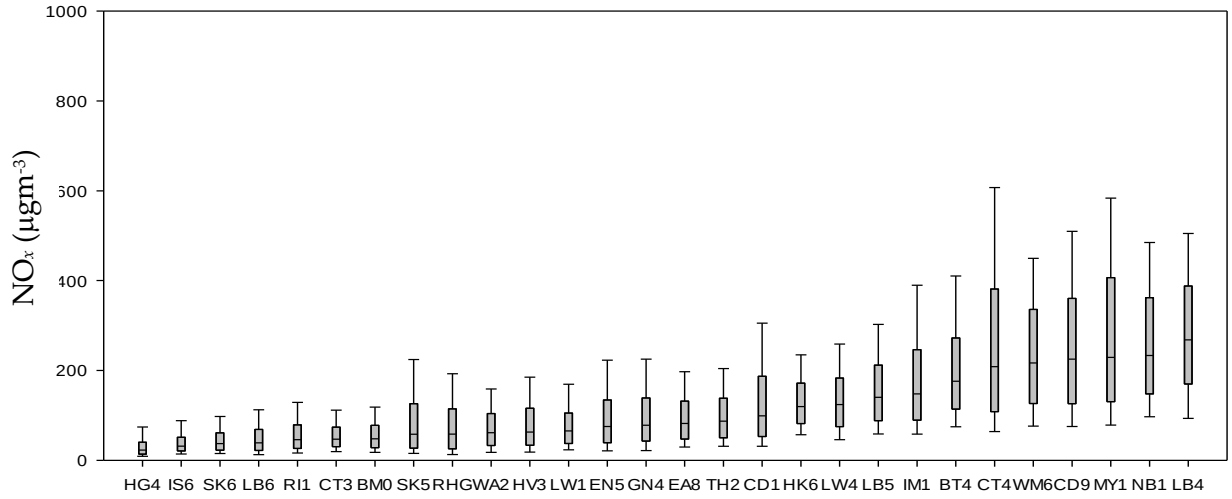
97 value, as indicated by the box midlines. A statistical summary of data at each station is
98 presented in Table 2.



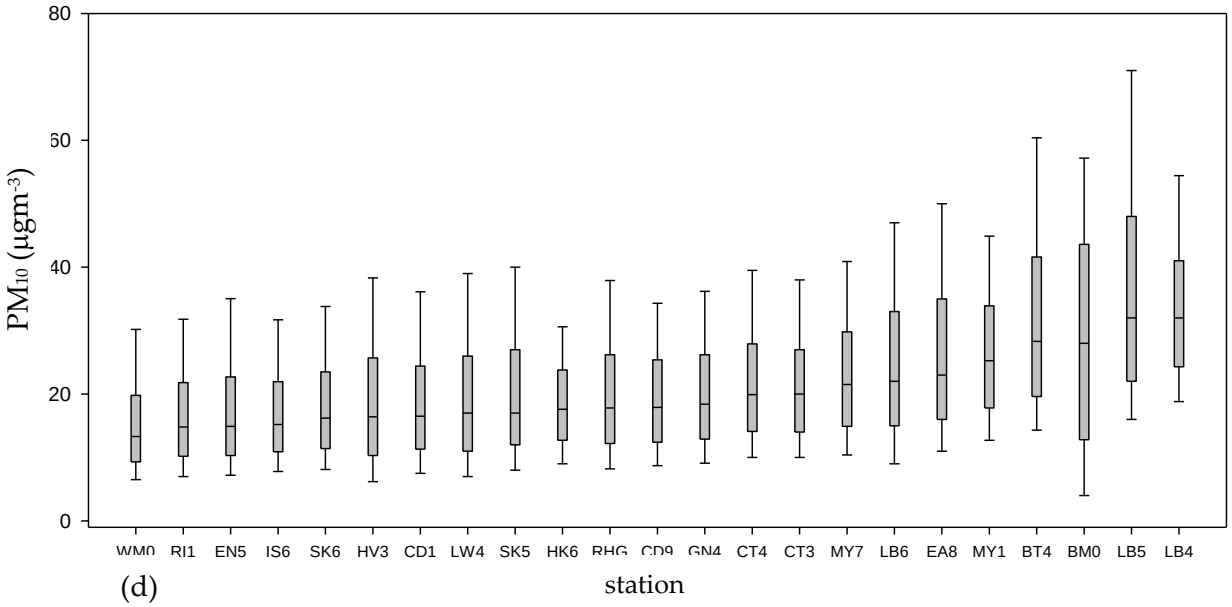
99
100



101



(c)



(d)

Figure 3: distribution of pollutant concentrations at different stations: a) NO, b) NO_2, c) NO_x and d) PM_{10} (μgm^{-3})

Table 2: statistical summary of NO, NO_2, NO_x and PM_{10} data at each station

Zone	Station	Average (μgm^{-3})				Standard Deviation (μgm^{-3})			
		NO	NO_2	NO_x	PM_{10}	NO	NO_2	NO_x	PM_{10}
Central Zone	WM6	111.33	72.07	242.77	-	78.95	29.51	145.59	-
	BM0	15.44	37.70	61.37	29.38	24.52	18.81	51.87	20.48
	IM1	75.44	74.49	190.16	-	75.00	32.56	143.31	-
	CT4	132.32	79.97	282.85	22.87	153.98	36.60	266.89	13.63
	NB1	114.43	91.74	267.62	-	82.75	34.98	156.88	-
	WM0	-	-	-	16.63	-	-	-	12.43
	CT3	15.05	38.04	61.11	22.58	26.68	18.59	54.59	15.32
	MY1	131.73	83.68	285.67	27.41	110.82	37.54	202.18	14.29
	MY7	-	-	-	24.18	-	-	-	14.16
	CD9	117.83	83.15	263.84	20.33	95.25	35.67	177.73	11.79
	SK6	11.80	34.11	52.20	19.20	27.66	19.52	57.02	12.15
	LB5	65.04	65.46	165.19	36.78	55.99	27.69	109.89	24.73
	HK6	51.71	57.32	136.59	19.05	41.14	20.88	79.94	9.56
	Summary*	**76.51**	**64.92**	**182.27**	**23.86**	**70.44**	**28.32**	**131.56**	**14.86**
LEZ	CD1	56.97	52.70	140.04	20.16	67.26	28.32	126.97	14.82
	IS6	11.32	30.92	48.28	18.09	31.72	20.08	64.62	11.55
	HG4	8.91	24.39	38.05	-	28.40	16.91	55.00	-
	TH2	38.23	48.48	107.10	-	42.55	22.79	83.32	-
	LW4	58.40	53.92	143.46	20.97	52.79	25.09	100.99	18.40
	SK5	33.68	42.44	94.08	21.59	49.33	28.22	99.86	15.52
	LB4	127.19	95.48	290.51	34.55	85.87	39.80	166.66	17.44
	RI1	23.77	31.04	67.48	17.68	41.83	19.41	78.74	11.74
	Summary*	**44.83**	**47.43**	**116.16**	**21.86**	**49.86**	**25.05**	**96.82**	**14.74**
Greater London	EA8	38.42	44.98	103.89	27.03	49.89	22.18	93.55	17.22
	BT4	96.38	72.42	220.20	32.96	95.43	32.16	173.80	22.21
	EN5	43.27	44.86	111.21	19.01	72.35	25.21	131.03	15.09
	LW1	29.34	43.08	88.07	-	45.11	22.40	87.16	-
	WA2	26.07	39.78	79.76	-	35.06	21.61	71.08	-
	RHG	34.54	36.61	89.56	21.19	54.14	25.07	103.80	14.42
	LB6	16.65	29.66	55.23	25.59	27.68	20.37	58.40	17.29
	GN4	43.34	41.15	107.79	21.16	54.31	24.22	103.90	12.37
	HV3	32.84	40.09	90.44	19.87	48.52	24.26	94.10	14.69
	Summary*	**39.88**	**43.62**	**104.79**	**24.04**	**53.32**	**24.10**	**101.36**	**16.45**
	Observations	231,326			179,500				

* Summary averages are weighted by the number of observations at each station.
Summary standard deviations are pooled estimates

The average of NO in the central London stations is 76.56 μgm^{-3} while it is about 30% and 40% lower at stations in the LEZ and greater London, respectively. Similar spatial variation occurs in each of NO_x and NO_2. However, there is also substantial variation among the stations within each of the zones. By contrast, the average of PM_{10} does not vary substantially either within or across zones. The standard deviations of concentration of NO and hence of combined NO_x at stations in the central zone are more variable than in the other zones. The mean and standard deviations of PM_{10} values vary little across the stations, showing that this pollutant is spread more evenly across London.

2.2. Analysis of variance (ANOVA)

The concentrations of the oxides of nitrogen summarised in Table 2 vary across the three zones (central London, LEZ and greater London), among the stations within each of these zones, and over the hours of observation at each of the stations. The question arises naturally of whether the variation across zones is greater than would be expected given the variation among stations. If the observations vary across zones are sufficiently large, then allowance would be appropriate for the zone in which each station is located. Otherwise, the variation across zones observed in the data can be considered as consistent with the variation among stations so that no allowance for zone would be justified.

To investigate the variation in air pollution measurements across the zones (central London, LEZ and greater London), a nested two-factor analysis of variance (ANOVA) with random effects was performed (Ledolter and Hogg, 2009). In this analysis, the zones are represented in the upper level, with each station associated uniquely with the zone of its location (hence the nested structure). The stations are considered as a sample of locations in London so that measurements at an additional station would be expected to vary similarly. The results of this analysis for each of the 4 pollutants are presented in Table 3.

The ANOVA procedure compares two estimates of the variance in the observations across stations. The first is a zone-based estimate M_z that is calculated from the sum of the squared (S_z^2) differences between the zonal means and the mean of all estimates:

$$M_z = S_z^2 / \nu_z \quad (1)$$

where ν_z is the degree of freedom associated with the zonal sum of squares. This estimate is compared with M_s that is calculated from the sum S_s^2 of the squared differences between the observations at the stations and the mean at their respective zone, allowing for the residual variation S_r^2 at the stations:

$$M_s = \left(S_s^2 - S_r^2\right) / \tilde{n} \quad (2)$$

In this expression, $\tilde{n}$ is the effective number of observation (because some observations are missing at some of the stations) and is calculated as:

$$\tilde{n} = \frac{1}{(J-1)}\left(\sum_{j=1}^{J} n_j - \sum_{j=1}^{J} n_j^2 \Big/ \sum_{j=1}^{J} n_j \right) \quad (3)$$

where $n_j (1 \le j \le 30)$ is the actual number of observations at station j, and J is total number of stations (30).

The estimate M_z of variance calculated across zones from (1) is compared to the corresponding estimate M_s of variance among stations calculated from (2) in the ratio:

$$F_z = M_z / M_s \ . \quad (4)$$

The probability $P(F_z|H_0)$ is calculated of a value at least this large arising with the degrees of freedom (ν_z, $\tilde{n}$) under the null hypothesis H_0 that variation across zones arises from that among stations. This is compared against the critical value for the required level of statistical significance: if the probability calculated from (4) is less than the critical one, then the null hypothesis is rejected.

The results of these ANOVAs show that the mean concentrations of each of the three measures of oxides of nitrogen vary substantially across the zones (central London, LEZ and greater London), as quantified by the associated sum of squared deviations S_z^2 of the zonal mean from the mean of all measurements of that pollutant. The resulting estimate M_z of variance across zones calculated from (1) for each of the oxides of nitrogen (NO, NO_2, NO_x) is 2-3 times greater than the corresponding estimate M_s of variance among stations calculated from (2), though for the particulates (PM_{10}) these estimates are in reverse magnitude: $M_z < M_s$. However, in each case, the F_z ratio for zones calculated from (4) is less than the critical value at α=0.05 . Accordingly, the null hypothesis that the variation across zones arises from that among stations cannot be rejected at this level of statistical significance.

167 *Table 3: Nested analysis of variance (ANOVA) of NO, NO_2, NO_x and PM_{10} (μgm^{-3}) at each station*

		Sum of squares S^2	Degrees of freedom ν	Mean square $M = S^2/\nu$	F	$P(F\|H_0)$	s_s
NO	Zone	6.53×10^7	2	3.27×10^7	2.59	0.095	
	Station	3.15×10^8	25	1.26×10^7	2876.3	<0.001	39.8
	Residual	1.01×10^9	2.31×10^5	4.39×10^3			$\tilde{n}$ =7967
	Total	1.40×10^9	2.31×10^5				
NO_2	Zone	8.25×10^6	2	4.12×10^6	1.18	0.323	
	Station	8.72×10^7	25	3.49×10^6	4843.9	<0.001	20.9
	Residual	1.67×10^8	2.31×10^5	7.20×10^2			$\tilde{n}$ =7967
	Total	2.62×10^8	2.31×10^5				
NO_x	Zone	2.89×10^8	2	1.45×10^8	2.86	0.076	
	Station	1.27×10^9	25	5.06×10^7	3329.4	<0.001	79.7
	Residual	3.52×10^9	2.31×10^5	1.52×10^4			$\tilde{n}$ =7967
	Total	5.07×10^9	2.31×10^5				
PM_{10}	Zone	5.15×10^4	2	2.58×10^4	0.10	0.909	
	Station	5.40×10^6	20	2.70×10^5	1099.4	<0.001	6.0
	Residual	4.41×10^7	1.79×10^5	2.46×10^2			$\tilde{n}$ =7457
	Total	4.95×10^7	1.79×10^5				

168 The interpretation of this result is that there is insufficient evidence in the data that the variation
169 in concentration of the pollutants across the zones exceeds that expected from the substantial
170 variability among the sites.

171 The categorical models implicit in this analysis show that for each pollutant a significant
172 variation in air pollution arises among sites as quantified by the large values of the F ratios
173 for stations $F_s = M_s / M_r$. However, only a small proportion of the total sum of squared
174 variations S_T^2 in the data arises from differences among the stations, with the remainder arising
175 from variation at the individual stations. This is quantified by the coefficient of determination
176 (R^2) calculated as the proportion of the total sum of squares that is associated with zones and
177 stations:

178 $$R^2 = \frac{S_z^2 + S_s^2}{S_T^2}. \qquad (5)$$

179 The resulting values of R^2 are: 0.27, 0.36, 0.31 and 0.11 respectively for NO, NO_2, NO_x and
180 PM_{10} . These low values of R^2 show that the measured variation is about 3-4 times greater at

the stations than among them so that other effects including spatio-temporal ones should also be considered to represent the variation in concentrations of pollutants.

2.3. Imputation of missing values

Missing observations are a common issue in practical data analysis: this presents particular problems for multivariate time-series analysis in which the series of observations would be interrupted by any missing values at any time within the series. To address this, various methods to impute missing values have been devised to complete series synthetically. These range from simple methods such as assigning the mean or median of each time series (Anzai, 2012; Bishop, 2006); Du et al., 2019), linear interpolation between available observations during short gaps (Freeman et al. 2018), use of data at the corresponding time on the previous day during longer gaps (Freeman et al. 2018) to more sophisticated methods such as spline interpolation (Shaub, 2020) and Kalman filtering (Hyndman and Khandakar, 2008).

In the London air quality dataset analysed here, the proportion of missing values varies across stations and pollutants. Station LB6 is critical in this respect with 52.5% missing data in PM_{10}, and 15.7% in each of the oxides of Nitrogen. Across all stations, an average of 11% of the measurements of PM_{10}, and 5% in each of NO, NO_2 and NO_x are missing.

To identify a suitable imputation approach to complete the data series, six different methods were investigated and their results compared. For this comparison, station HG4 was selected, which has measurements of 3 pollutants (NO, NO_2 and NO_x) and a low missing data rate (0.2%). For each variable at this station, 500 observations were selected for use as validation data using two approaches:

- Single missing gap:

  In this approach, a set of 500 successive hourly observations (about 3 weeks) without any missing values was identified separately for each pollutant.

- Multiple missing gaps:

  Here, multiple gaps with an average of 24 successive observations (1 day of data monitoring) were selected randomly within the series of each pollutant. The total number of validation data is still 500, but they are distributed with smaller durations within the dataset at times when original observations are available at that station.

These two sets of 500 observations were then used as validation data by comparing the values estimated by imputation methods with the true values. The two approaches for selecting validation data intervals are illustrated in Figure 4.

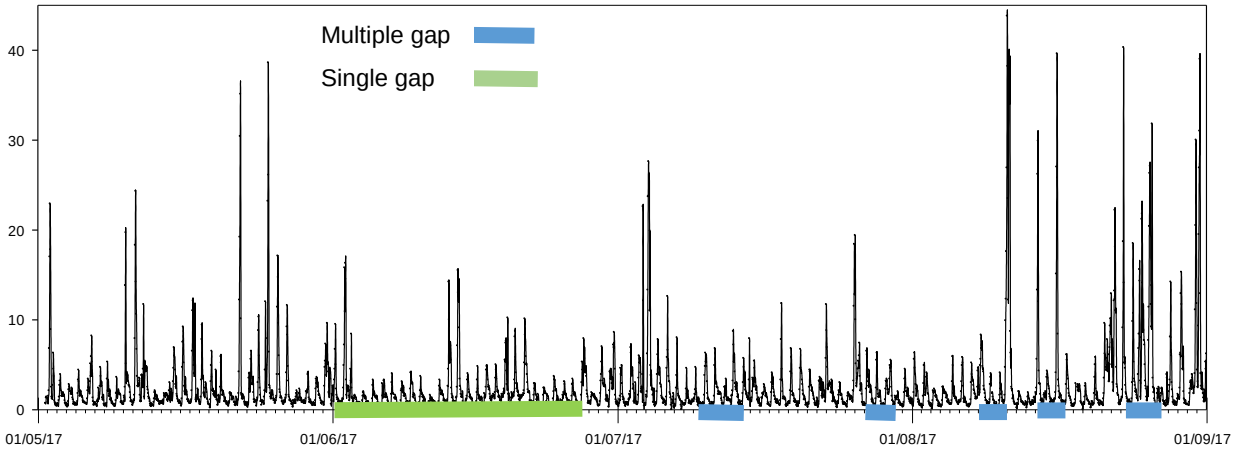


Figure 4: NO concentrations (μgm^{-3}) in station HG4 from May to Sept. 2017, top: all observations, bottom: validation points selected by "single" and "multiple" missing gaps

The process of imputation uses available observations, denoted as the set Ω, to estimate the missing values in the complementary set Ω^C. We denote an interval throughout which values of a series x_t are missing as $I = [a+1, b-1]$ between observations x_a and x_b at times $t = a$ and $t = b$ respectively: the imputation method provides estimates for these missing values in all such intervals and thus completes the series. The methods for this investigated here are described in turn.

227 I. Linear interpolation

228 Linear imputation is a simple and convenient method to impute missing values in a univariate
229 time series (Suresh et al., 2019). This method applies a constant rate of change between
230 values that were observed:

231 $$\tilde{x}_t = x_a + \left(\frac{x_b - x_a}{b - a} \right)(t - a) \quad (a < t < b) \tag{6}$$

232 In this linear interpolation, the imputed values are bounded to lie in the interval $[x_a, x_b]$ between
233 the ones that have been observed. This approach will not bias the mean value of the time
234 series but will usually give estimates that vary from it, and will also tend to reduce both variance
235 and covariance. By its nature, it will minimise the rate of change throughout each interval
236 within which it is applied and so increase serial correlation.

237 II. Spline interpolation

238 Spline imputation is an extension of the linear method that fits a smooth curve to the observed
239 data that is available in Ω and uses that to estimate missing values. The most widely used
240 form of spline is cubic (Wood, 2017), which is therefore considered here. The values imputed
241 for an interval $I = [a+1, b-1]$ are calculated using the cubic function:

242 $$\tilde{x}_t = \sum_{h=0}^{3} \alpha_h \left(t - a\right)^h + \sum_{h=0}^{3} \beta_h \left(b - t\right)^h \quad \left(a < t < b\right) \tag{7}$$

243 where the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are calculated to ensure continuity of finite difference
244 approximations to the first and second derivatives at times $t = a$ and $t = b$ as well as
245 throughout the interval I.

246 III. Exponentially weighted moving average (EWMA)

247 The EWMA approach uses values $x_s \left(s < t\right)$ up to the time t at which is it applied according
248 to the formula:

$$\tilde{x}_t = (1-\lambda)\sum_{j=1}^{\infty}\lambda^j x_{t-j-1} = (1-\lambda)\,x_{t-1} + \lambda\,\tilde{x}_{t-2} \tag{8}$$

with smoothing parameter $\lambda \in (0,1)$.

If no observation x_{t-1} is available at time t-1, then substituting the EWMA estimate $\tilde{x}_{t-1}$ into the recursion relationship gives:

$$\tilde{x}_t = (1-\lambda)\,\tilde{x}_{t-1} + \lambda\,\tilde{x}_{t-2} \Rightarrow \tilde{x}_t = \tilde{x}_{t-1}\ . \tag{9}$$

Accordingly, values calculated using this approach will be constant and equal to the most recent value, so the imputed values will be constant throughout the interval I. This estimate will be sensitive to the weighting parameter λ through the accuracy of estimation of the first missing value and the equality of all subsequent missing ones. This approach will not bias the mean value of the time series but will usually give estimates that vary from it, and will also reduce both variance and covariance. Because it gives estimates that are constant within the interval to which it is applied, it will increase serial correlation.

IV. Multiple imputation

The simplest multivariate method uses correlations observed among the components to estimate deviations of imputed values from their mean (Raghunathan, 2015). Thus let,

$$\mu_k = \frac{1}{\|(k,s)\in\Omega\|}\sum_{s:(k,s)\in\Omega} x_{ks}$$
$$\sigma_{jk} = \frac{1}{\|(j,s)\in\Omega\wedge(k,s)\in\Omega\|}\sum_{s:(j,s)\in\Omega\wedge(k,s)\in\Omega}\left(x_{js}-\mu_j\right)\left(x_{ks}-\mu_k\right) \tag{10}$$

so that μ_k is the mean of observed values of series k and σ_{jk} is the sample covariance of series j and k.

Then the values imputed by this method are:

$$\tilde{x}_{kt} = \mu_k + \sum_{j:(j,t)\in\Omega}\left(x_{jt}-\mu_j\right)\sigma_{jk} \quad \forall (k,t)\in\Omega^{\mathrm{C}} \tag{11}$$

V. Multiple imputation by chained equation (MICE)

The MICE method (Resche-Rigon and White, 2018; van Buuren and Groothuis-Oudshoorn, 2011; Zhang, 2016) adopts a joint distribution function of the multivariate data $\mathbf{x}_t = \left(x_{1t}, x_{2t}, \ldots x_{Kt}\right)'$ at each time t. The multivariate normal (MVN) distribution was used in the present study, though others could be used in its place. The MICE procedure starts with initial estimates for missing data $x_{kt}, \left(k,t\right) \in \Omega^{\mathrm{C}}$ using the mean value of the observations of the corresponding quantity:

$$\begin{aligned} &n = 0, \\ &\tilde{x}_{kt}\left(n\right) = \tfrac{1}{\left\|\left(k,s\right)\in\Omega\right\|} \sum_{s:\left(k,s\right)\in\Omega} x_{ks} \quad \forall\left(k,t\right) \in \Omega^{\mathrm{C}} \end{aligned} \tag{12}$$

Based on this extension to the observations, parameters are specified for distributions of the multivariate mean, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\sigma}$, of the data; parameters $\boldsymbol{\theta}\left(n\right) = \left(\boldsymbol{\mu}\left(n\right), \boldsymbol{\sigma}\left(n\right)\right)$ are then drawn from these distributions. Then the algorithm proceeds with two steps: Imputation (I) and Posterior (P):

- *I-step*:

  Each missing value $\tilde{x}_{kt}\left(n\right), \left(k,t\right) \in \Omega^{\mathrm{C}}$ is synthesised by sampling from the specified distribution with parameters $\boldsymbol{\theta}\left(n\right)$ so that:

$$\begin{aligned} &E\left[\tilde{x}_{kt}\left(n\right)\right] = \mu_k\left(n\right) + \sum_{j:\left(j,t\right)\in\Omega} \left(x_{jt} - \mu_j\left(n\right)\right)\sigma_{jk}\left(n\right) \\ &Var\left[\tilde{x}_{kt}\left(n\right)\right] = \sigma_{kk}\left(n\right) \quad \left(k,t\right) \in \Omega^{\mathrm{C}} \end{aligned} \tag{13}$$

  (using previously imputed values for x_{jt} in the right-hand side where $\left(j,t\right) \in \Omega^{\mathrm{C}}$).

- *P-step:*

  The distributions for the parameters $\boldsymbol{\theta}$ are updated using the observed values complemented with those imputed in the I-step.

290 If updating the distributions of the parameters $\boldsymbol{\theta}$ results in negligible changes, the
291 algorithm terminates yielding the imputed values from the I-step. Otherwise, the process
292 continues with the next I-step.

293 When convergence is achieved, stationary distributions have been achieved. This process
294 completes the multivariate time series by imputing values where observations are missing,
295 resulting in statistical distributions that are consistent with the observed data.

296 These six imputation methods were applied to data from the station HG4, to estimate the
297 values in the validation intervals. Validation data were reserved from the dataset, then the
298 effectiveness of each imputation method was evaluated by comparing each of the imputed
299 values with the corresponding reserved one. The criterion adopted for this was the root mean
300 squared error (RMSE), e, calculated for the test interval(s) of imputation as:

301 $$e = \sqrt{\frac{1}{\left|\Omega^{C}\right|}\sum_{t\in\Omega^{C}}\left(x_t - \tilde{x}_t\right)^2}, \qquad (14)$$

302 where $\left|\Omega^{C}\right|$ is the number of values imputed in the test interval (here $\left|\Omega^{C}\right|$ = 500), x_t is the
303 observed and $\tilde{x}_t$ is the imputed value at time t. Only a small proportion, about 0.2%, of
304 missing values outside the test interval in this dataset were imputed in this procedure, and
305 because no corresponding observed values are available, they were not considered in this
306 test. The results for each of the imputation methods are presented in Table 4.

307 *Table 4: Root mean squared error e for the imputation methods*

308

Imputation method	Selection Method	RMSE *e* (µgm^{-3})		
		NO	**NO_2**	**NO_x**
309 Mean	Single gap	14.21	19.87	21.18
	Multiple gaps	14.21	19.87	21.28
310 Linear	Single gap	4.72	6.73	8.63
	Multiple gaps	3.87	5.80	7.29
311 Spline	Single gap	3.34	5.22	6.58
	Multiple gaps	3.21	5.24	6.40
312 Exponentially weighted moving average (EWMA)	Single gap	3.74	5.52	6.78
	Multiple gaps	3.4	5.23	6.92
313 Multivariate	Single gap	2.23	3.55	5.24
	Multiple gaps	1.83	2.58	3.76
Multiple imputation by chained equation (MICE)	Single gap	1.74	2.52	3.65
	Multiple gaps	1.45	2.02	3.41

The results of imputation with these methods show that the MICE algorithm works substantially better than any of the other approaches. In general, the RMSE values for multivariate approaches are lower than the corresponding ones in the univariate ones. Besides this, when the missing gap is relatively large (500 successive hourly observations, about 3 weeks), the value of e is greater compared to the case with smaller gaps (24 successive hourly observations on average, 1 day). Based on these results, the MICE method was adopted to synthesise complete the multivariate time series in the present study.

3. Methodology

The analysis of variance (ANOVA) presented in Section 2 shows that the majority of variation in atmospheric pollution measurements arises at each of the stations, with relatively little difference among them. To investigate the time dependency of pollutant concentrations, time-series analyses are investigated in this section. The time series models developed here are based on the autoregressive integrated moving average (ARIMA) model, introduced by Box and Jenkins (Box, 2008) with seasonal extensions. Initial investigation showed that there is no need for differencing the data series, so that the simpler autoregressive moving average (ARMA) formulations were pursued.

Two approaches were investigated for modelling atmospheric pollution. In the first one, each station is modelled independently from the others, with the temporal variation of each pollutant modelled as a univariate time-series at that station. The second approach includes the spatial relationship among stations as well as temporal dependency; hence the observations at all stations are modelled jointly as a vector.

Because the response variable (atmospheric concentration of pollutants) cannot be negative, the logarithm transformation of concentrations was used as the dependent variable of these models.

3.1. Seasonal ARMA model (SARMA)

SARMA (Wei et al., 2013) is an extension of ARMA models that estimates the univariate temporal dependency of observations with seasonal trends. The seasonal trend is a repeated pattern over a specified time period, denoted as s. In the SARMA, the autoregressive and moving average orders (p and q, respectively) represent the non-seasonal trends in the time series. To represent the seasonal trend, three further parameters are used, which are denoted as P, Q and s for seasonal autoregressive, seasonal moving average and span of seasonality (seasonal period), respectively. Based on that, the SARIMA model for time series x_t can be expressed in the form:

$$x_t = \frac{\vartheta^Q(B)\theta^q(B)}{\varphi^P(B)\phi^p(B)}\omega_t$$

$$\vartheta^Q(B) = 1 + \sum_{i=1}^{Q} \vartheta_i B^{Si}, \quad \varphi^P(B) = 1 - \sum_{i=1}^{P} \varphi_i B^{Si} \qquad (15)$$

$$\theta^q(B) = 1 + \sum_{i=1}^{q} \theta_i B^i, \quad \phi^p(B) = 1 + \sum_{i=1}^{p} \phi_i B^i$$

where B is the backshift operator ($B^r x_t = x_{t-r}$), ϑ and θ are the moving average coefficients for seasonal and non-seasonal trend, respectively. The parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\phi}$ are the autoregressive coefficients for seasonal and non-seasonal trend, respectively, and $\boldsymbol{\omega}$ is normally distributed and uncorrelated white noise.

This model was developed in R version 3.5.2 using package “forecast”(R. Hyndman et al., 2020).

3.2. Vector ARMA model (VARMA)

VARMA models were developed to improve forecast accuracy by using interrelated variables (Cavicchioli, 2016; Dias and Kapetanios, 2018; Simionescu, 2013). In these models, the temporal relationship of each variable with the vector of others is estimated.

If $\mathbf{X_t}$ is a multivariate time series of K- dimensional vectors, then a VARMA model with autoregressive order p and moving average order q can be expressed in the vector form:

$$\mathbf{X}_t = \frac{\boldsymbol{\theta}^q(B)}{\boldsymbol{\varphi}^p(B)}\boldsymbol{\omega}_t$$

$$\theta^q(B) = 1 + \sum_{i=1}^{q} \theta_i B^i, \quad \varphi^P(B) = 1 - \sum_{i=1}^{P} \varphi_i B^{si} \tag{16}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are $K \times K$ matrices for autoregressive and moving average coefficients, respectively, and $\boldsymbol{\omega}$ is a normally distributed white noise vector of dimension K.

This model was developed in R version 3.5.2 using package "MTS" (Tsay and Wood, 2018).

4. Results

The results of fitting the SARMA and VARMA model to the London air quality, evaluation and validation the models are presented in this section. The performance of these models was evaluated using the Bayes Information Criterion (BIC) (Eilers and Möbus, 2011; Pandis, 2016):

$$\mathrm{BIC} = -2\mathcal{L} + \log_e(n)m \tag{17}$$

where m is the number of free parameters in the model, n is the number of observations and $\mathcal{L}$ is the log-likelihood of the fitted model. Models with smaller values of BIC are preferred, with use of additional parameter justified by sufficient improvement in the likelihood of the fitted model. This criterion provides a balance between improvement in fit (represents by increased log-likelihood) and model complexity (represented by the number of parameters used) whilst respecting the scaling effect of the dataset size (Pandis, 2016).

4.1. SARMA models

To investigate the effects of the seasonal trend for each pollutant type, SARMA models were developed with each of no seasonal ($s = 0$), daily ($s = 24$), and weekly ($s = 7 \times 24 = 168$) seasonal spans. For choosing the best orders for autoregressive and moving average parts, different combinations are tried with $p, q, P, Q \in \{1, 2, 3, 4\}$ and the model with the smallest BIC was selected (Table 5). In this table, the total of BIC values of stations in each zone (central London, LEZ and greater London) are presented for each pollutant as well.

Table 5: SARMA models fitted order and total BIC in each zone

			PM_{10}	NO	NO_2	NO_x
ARMA	Order (p,q)		(4,4)	(3,3)		
	BIC	Central	464367	842898	742294	952844
		LEZ	464365	842757	618585	951278
		Greater	461365	848985	768941	956078
		Total studied area	1390097	2534640	2129820	2860200
SARMA, s=24 (daily)	Order (p,q) (P,Q)		(2,2)(1,1)	(1,1)(1,1)		
	BIC	Central	461861	841147	739640	943704
		LEZ	462401	840281	605478	950101
		Greater	461258	840252	760132	947045
		Total studied area	1385520	2521680	2105250	2840850
SARMA, s=168 (weekly)	Order (p,q) (P,Q)		(1,1)(1,1)	(1,1)(1,1)		
	BIC	Central	460437	840211	733018	940012
		LEZ	460818	837101	605000	947044
		Greater	460355	839208	759012	946054
		Total studied area	1381610	2516520	2097030	2833110

For each pollutant, the same SARMA orders were preferred for all stations, although the coefficients differed: the preferred seasonal models have fewer parameters than the non-seasonal ARMA, so are preferred on grounds of parsimony as well as BIC. Comparing the BIC of SARMA models with different seasonal span shows that for all stations and pollutant types, models with weekly (s= 168) seasonal trends perform better than the corresponding one with daily (s= 24) seasonal trend.

The observed values are plotted against the estimates for station BM0 with the weekly SARMA (1,1)(1,1) model in Figure 5 in natural units for each of the pollutants. The reference red reference line indicates the case in which the observed values would be identical to the fitted ones. The points are spread along the reference line, but there is high variation around this line especially at large values of NO, NO_2 and NO_x. For PM_{10}, greater variation occurs at small values.

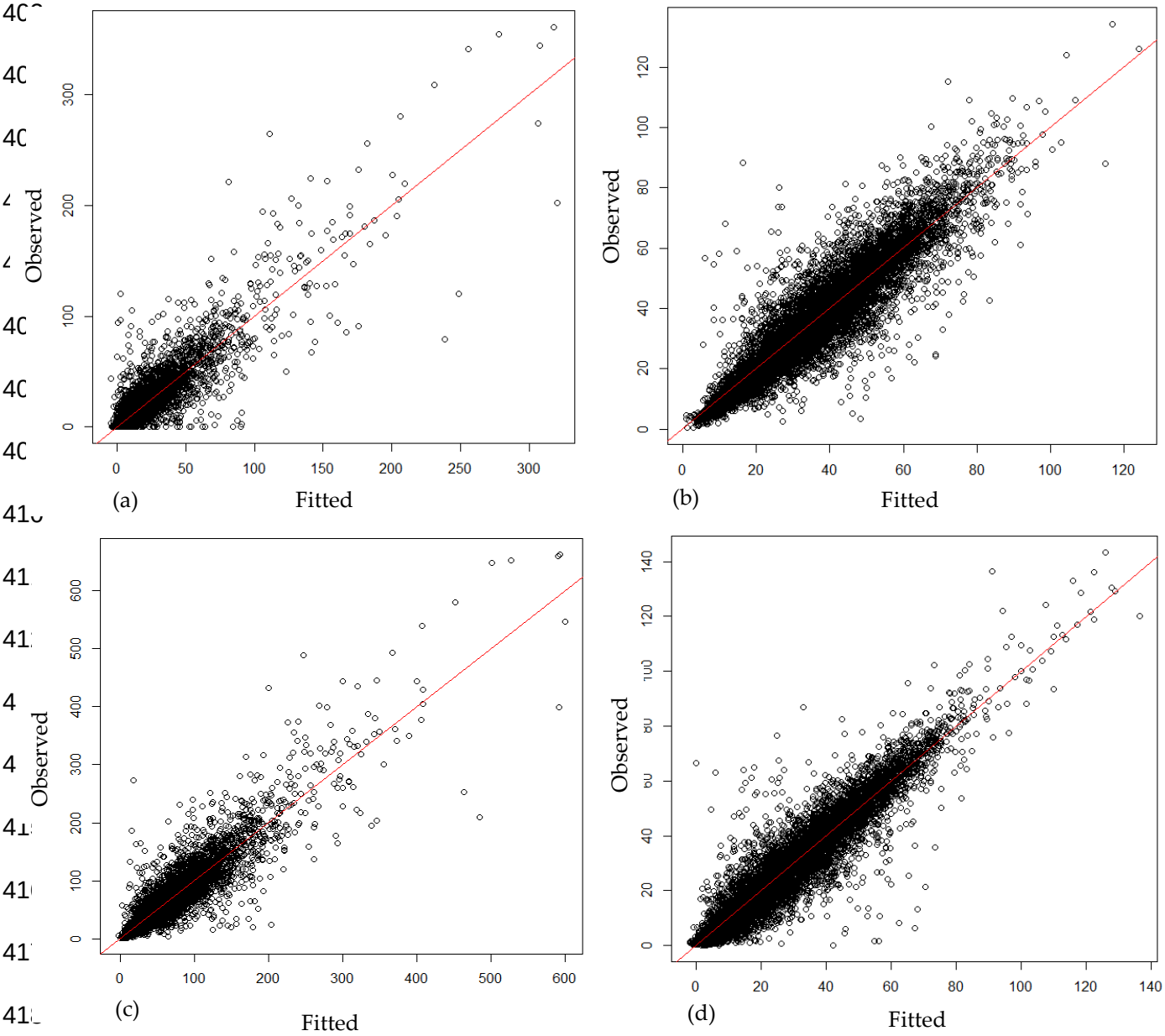


Figure 5: Observed values vs SARMA week-season model estimates: a) NO, b) NO2, c) NOx and d) PM10 (µgm^{-3})

In Figure 6, the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the residuals in the weekly SARMA model for station BM0 are shown. These plots indicate that some temporal structure (significant lags) remains in the residuals of the weekly SARMA model. Hence, this model is not fully successful in representing the temporal variations.

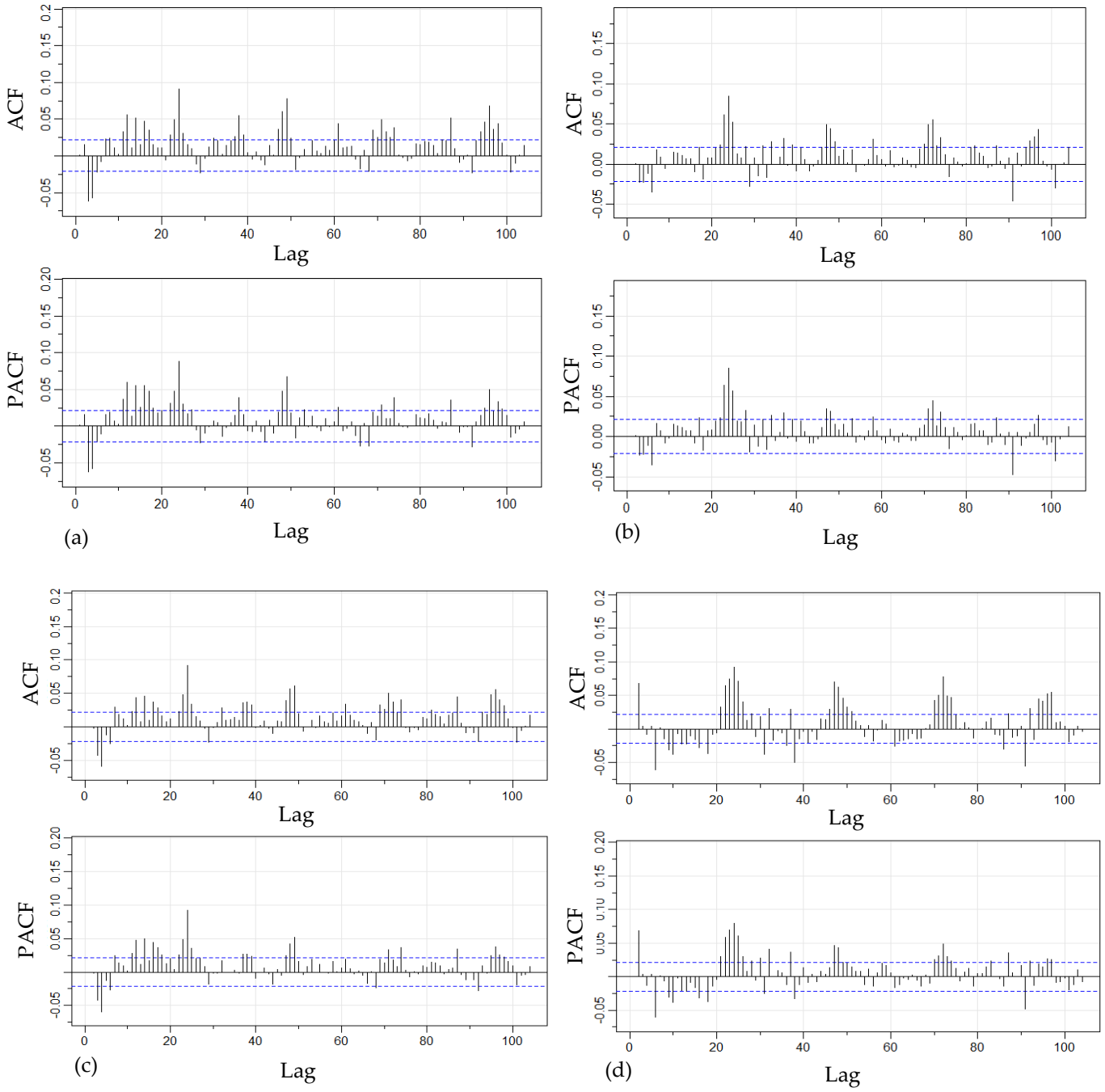


Figure 6: ACF and PACF of residuals, SARMA weekly model for: a) NO, b) NO_2, c) NO_x and d) PM_{10}

4.2. VARMA models

The VARMA models were developed for each pollutant separately over a vector of stations. The preferred order of each of these models is $(p, q) = (1, 1)$. Hence, two 30-dimensional matrixes (one autoregressive and one moving average) were estimated for each pollutant. In these matrixes, the principal is the coefficient of one hour lagged at the same station and the off-diagonal elements are the effects of other stations, again at one-hour time lag. As an example, the autoregressive $\boldsymbol{\theta}$ and moving average $\boldsymbol{\varphi}$ matrix parameters for NO_2 are

summarised in Table 6 and Table 7, respectively. The parameters in these tables are presented as the self-effects autoregressive (θ_{jj}) and moving average (φ_{jj}) for $1 \leq j \leq 30$ (at one hour lag, t-1) and for each station j, the mean and standard deviation of autoregressive (θ_{ji}) and moving average (φ_{ji}) of other stations ($1 \leq i \leq 30$, $i \neq j$). The distribution of these coefficients is plotted in Figure 7 and 8 for autoregressive and moving average, respectively. To summarise the association with other stations, the ratio of mean to standard deviation and the number of other stations with estimates that are more 2 standard deviations are also shown.

Table 6: Summary of the $\boldsymbol{\theta}$ matrix of autoregressive parameters in VARMA model for $\log_e NO_2$

Zone	Station *j*	Lagged effect of same station (*t*-1) θ_{jj}	Effects of other stations (*t*-1) θ_{ji}			
			Mean **μ**	Standard deviation **σ**	**μ/σ**	Number *i* $\theta_{ji} > 2\sigma$
Central Zone	WM6	0.886	0.516	4.452	0.116	5
	BM0	0.909	0.496	2.237	0.222	6
	IM1	0.868	0.488	2.012	0.243	5
	CT4	0.917	0.477	1.843	0.259	3
	NB1	0.907	0.479	4.916	0.097	2
	CT3	0.920	0.495	3.271	0.151	5
	MY1	0.882	0.505	66.340	0.008	4
	CD9	0.851	0.483	9.595	0.050	5
	SK6	0.912	0.516	6.448	0.080	4
	LB5	0.894	0.510	1.489	0.343	3
	HK6	0.888	0.484	1.546	0.313	5
LEZ	CD1	0.914	0.494	4.595	0.108	2
	IS6	0.910	0.521	6.289	0.083	3
	HG4	0.907	0.481	1.633	0.295	2
	TH2	0.931	0.506	3.546	0.143	5
	LW4	0.884	0.481	1.679	0.286	3
	SK5	0.889	0.512	2.039	0.251	5
	LB4	0.896	0.474	1.977	0.240	6
Greater London	RI1	0.899	0.485	0.488	0.994	12
	EA8	0.882	0.485	0.769	0.631	8
	BT4	0.921	0.484	0.434	1.117	16
	EN5	0.890	0.475	0.508	0.936	6
	LW1	0.872	0.488	0.230	2.123	25
	WA2	0.913	0.509	0.271	1.880	24
	RHG	0.924	0.494	0.229	2.160	25
	LB6	0.906	0.436	0.283	1.539	16
	GN4	0.892	0.488	0.219	2.225	18
	HV3	0.927	0.477	0.439	1.087	14

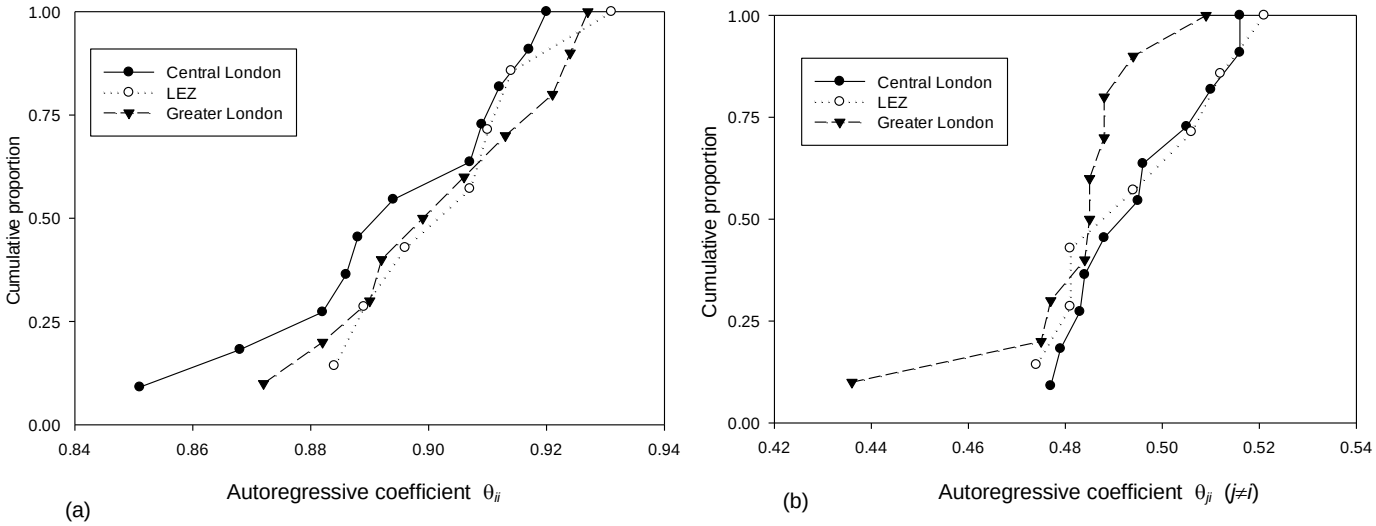


Figure 7: autoregressive coefficients for a) same station (θ_{jj}) and b) average of other stations (θ_{ji})

Table 7: Summary of the $\boldsymbol{\varphi}$ *matrix of moving average parameters in VARMA model for* $\log_e NO_2$

Zone	Station *j*	Lagged effect of same station (*t*-1) φ_{jj}	Effects of other stations (*t*-1) φ_{ji}			
			Mean μ	Standard deviation σ	μ/σ	Number *i* $\varphi_{ji} > 2\sigma$
Central Zone	WM6	0.787	0.014	0.0869	0.159	2
	BM0	0.736	0.054	0.1625	0.318	3
	IM1	0.896	0.020	0.0542	0.349	2
	CT4	0.699	0.102	0.2680	0.369	3
	NB1	0.803	0.043	0.0750	0.138	1
	CT3	0.795	0.006	0.0264	0.224	2
	MY1	0.911	0.034	0.0959	0.011	1
	CD9	0.754	0.007	0.0896	0.073	1
	SK6	0.890	0.002	0.0183	0.113	2
	LB5	0.843	0.022	0.0427	0.496	2
	HK6	0.873	0.017	0.0374	0.430	3
LEZ	CD1	0.862	0.007	0.0437	0.152	1
	IS6	0.964	-0.010	0.0777	-0.121	0
	HG4	0.777	0.040	0.0313	0.416	1
	TH2	0.875	0.027	0.1249	0.210	1
	LW4	0.836	0.033	0.0819	0.392	2
	SK5	0.819	0.007	0.0186	0.364	4
	LB4	0.808	0.014	0.0397	0.352	5
Greater London	RI1	0.801	0.036	0.0244	1.432	9
	EA8	0.813	0.065	0.0394	0.905	8
	BT4	0.831	0.037	0.0219	1.614	12
	EN5	0.801	0.047	0.0339	1.336	6
	LW1	0.725	0.060	0.0191	3.010	22
	WA2	0.934	0.061	0.0214	2.747	24
	RHG	0.903	0.053	0.0160	3.178	22
	LB6	0.878	0.064	0.0286	2.154	13
	GN4	0.747	0.060	0.0186	3.137	18
	HV3	0.881	0.048	0.0294	1.580	10

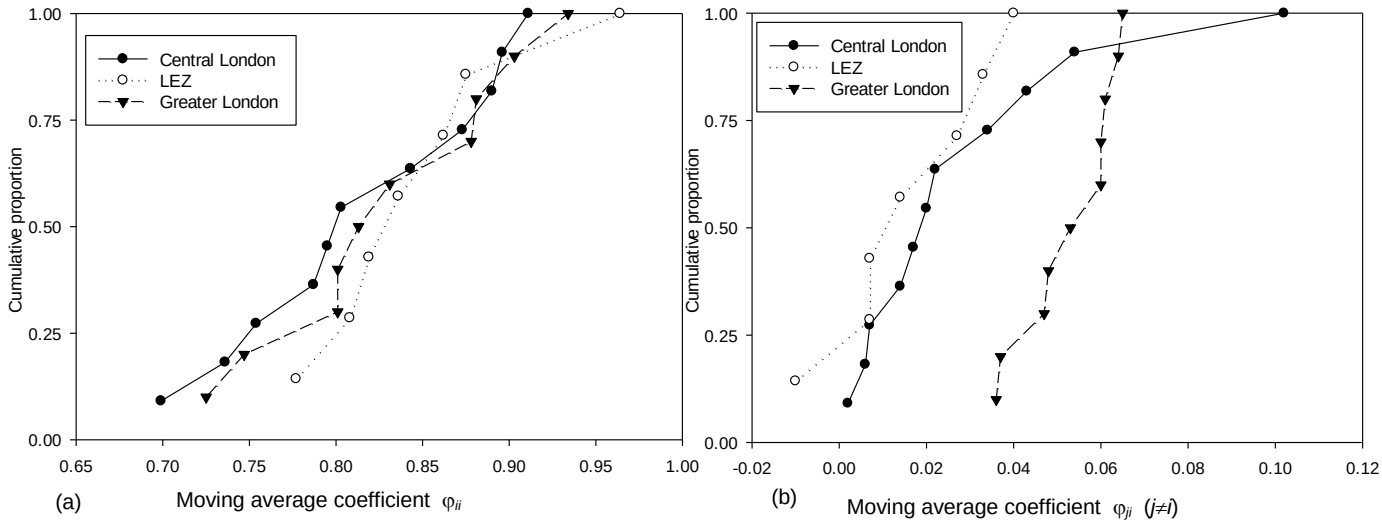


Figure 8: moving average coefficients for a) same station (φ_{jj}) and b) average of other stations (φ_{ji})

The autoregressive self-effect of stations at one-hour lag (t-1) for $\log_e NO_2$ ranges from 0.85 to 0.93 with mean 0.900 and standard deviation 0.019: considering the stations as a sample of locations in London, these can be viewed as random effects. This range is similar for other pollutants, indicating the relevance of the one-hour lagged value in estimating the current value at each station. The corresponding moving average self-effect is slightly weaker but more variable, ranging from 0.70 to 0.96 with mean 0.830 and standard deviation 0.066, a pattern that is repeated for other pollutants. For both autoregressive and moving average parts of the model, the number of other stations with estimates of effects that exceed twice the standard deviation for the associated station is greatest in outer London. For the autoregressive part these are between 2 to 6 (typically around 4) for stations in central London and the LEZ (typically around 4) but between 6 to 25 (typically around 16) for stations outside the LEZ. Hence remarkably, the air quality of the stations located outside of the LEZ (away from central London) is associated with a greater number of other stations, compared to those in central London.

The observed *vs* estimated values from VARMA model for station BM0 are plotted in Figure 9 in natural units of μgm^{-3}. The reference red line is the line of equality. Based on this, there is a good correspondence between the fitted values from the model and the observed ones.

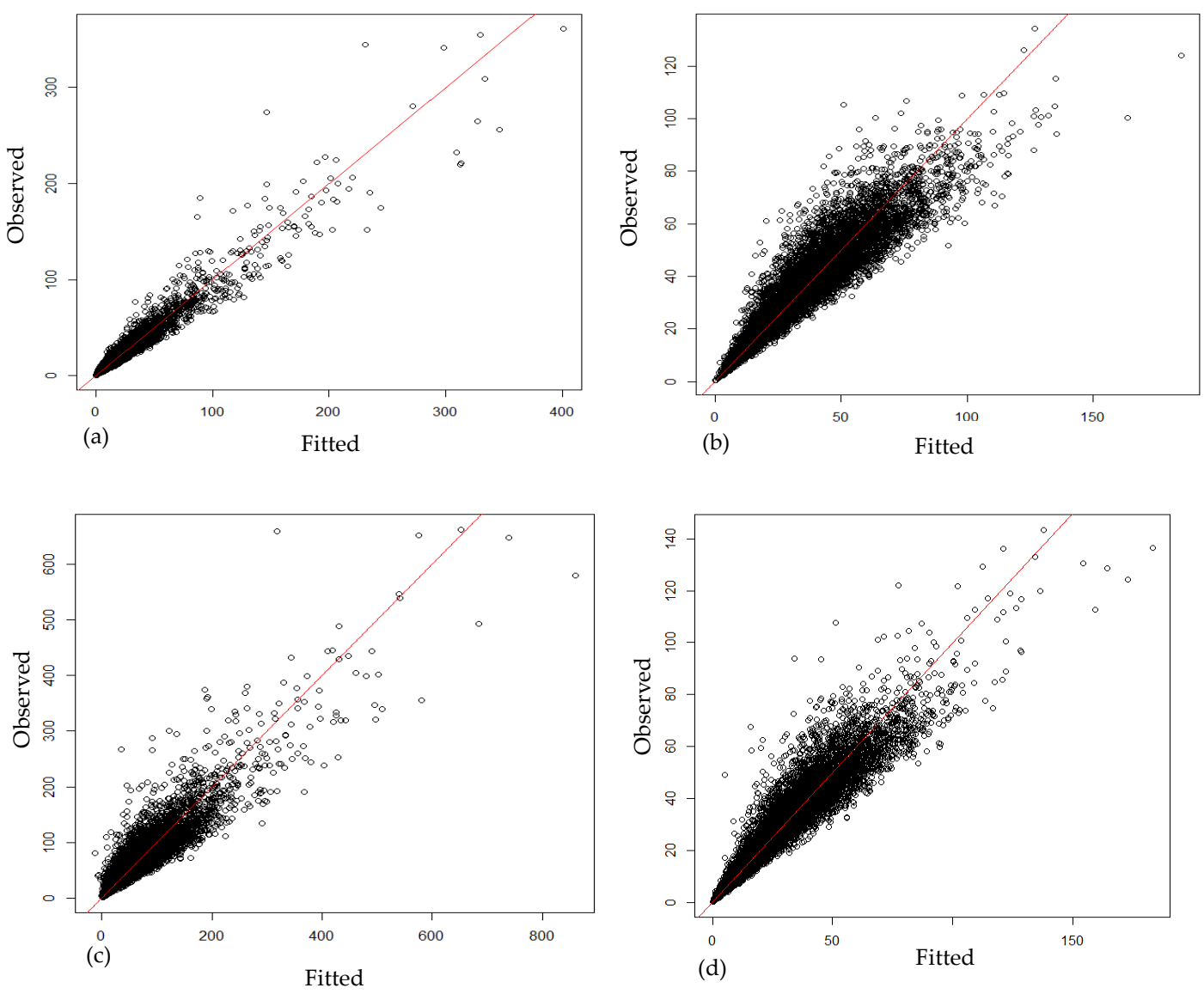


Figure 9: Observed values vs VARMA model estimates: a) NO, b) NO_2, c) NO_x and d) PM_{10} (μgm^{-3})

The ACF and PACF of the residuals in the VARMA model (at station BM0) are shown in Figure 10. Based on these plots, the VARMA model is reasonably successful in representing all the variations in the observations, leaving few significant lags in the residuals.

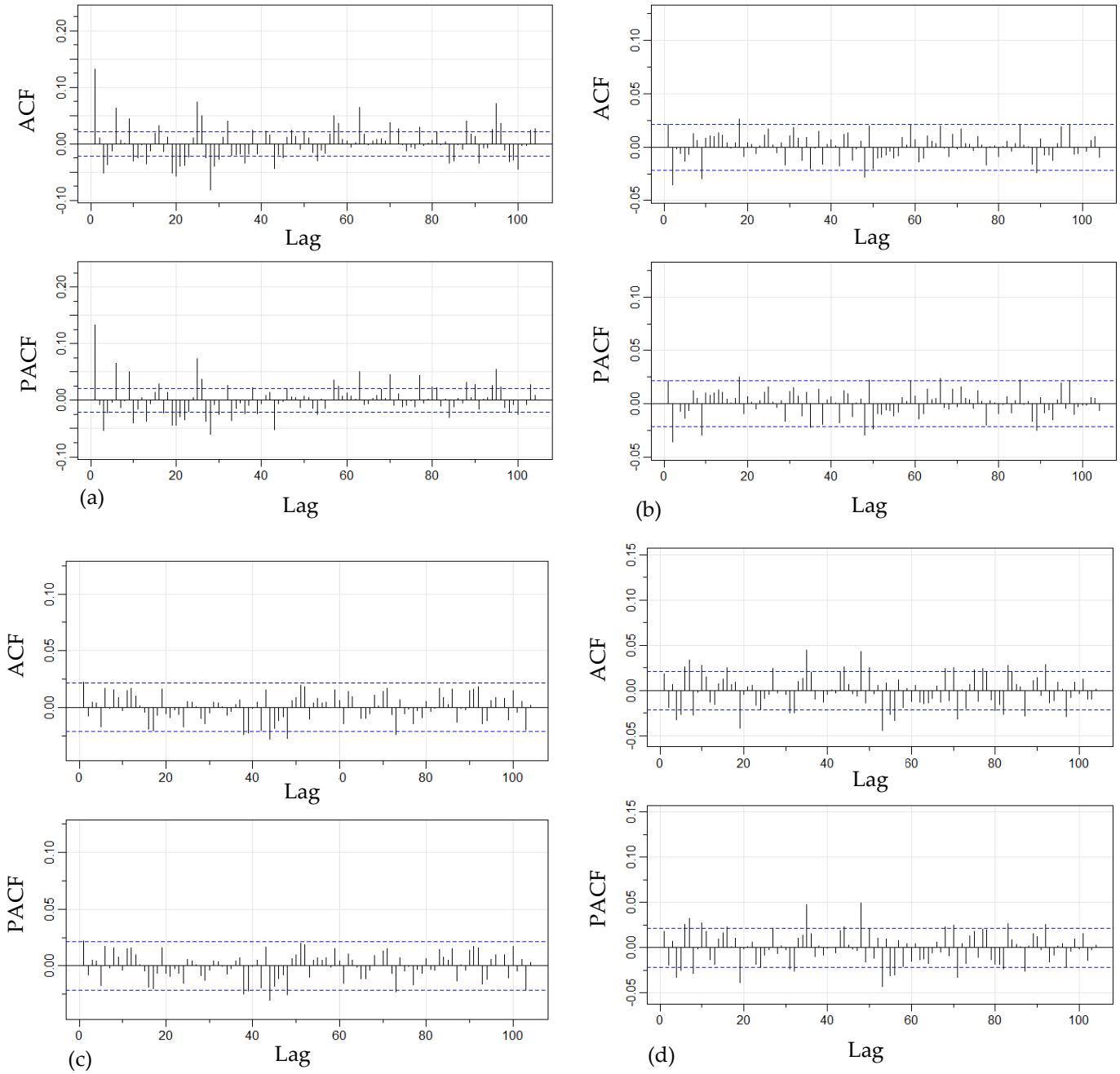


Figure 10: ACF and PACF of residuals, VARMA model station BM0: a) NO, b) NO_2, c) NO_x and d) PM_{10}

4.3. Evaluation

Because the SARMA models are fitted to each station individually whilst VARMA models use a vector of all stations, the BIC values of the SARMA models cannot be compared directly with corresponding ones for VARMA. To address this, an aggregated form of BIC (denoted as BIC') was calculated for the SARMA models as:

$$\mathrm{BIC}' = -2\mathcal{L}' + \log_e(n')m' \tag{18}$$

where the $\mathcal{L}'$ is the sum of model log-likelihoods, m' is the sum of the numbers of free parameters and n' is the sum of the number of observations in the SARMA models across all stations. The BIC' of the SARMA models was calculated by summing the individual values from Equation (18). These results along with the corresponding BIC values of VARMA models are presented in Table 8.

Table 8: BIC' and BIC values (millions) for SARMA and VARMA models

	BIC' (SARMA) *			BIC
Pollutant	**ARMA**	**s=24 (daily)**	**s=168 (weekly)**	**(VARMA)***
PM_{10}	1.37 (e')	1.36	1.34	1.14
NO	2.35	2.34	2.33	1.62
NO_2	1.95	1.94	1.93	1.69
NO_x	2.64	2.63	2.61	2.30

* the values presented are BIC' and BIC divided by 10^6

The values of BIC' in Table 8 show that for each pollutant the seasonal SARMA models perform better than the corresponding non-seasonal one. The weekly models in all pollutant types have better (smaller) BIC' values than the daily ones, indicating that air quality at the current time (t) is estimated more effectively by considering corresponding time one week beforehand (t-168) rather than the corresponding time on the previous day (t-24). However, for each pollutant, the BIC values of the VARMA models are substantially smaller than any of the corresponding SARMA ones, even though the VARMA models have 15 times more parameters. According to this criterion, the VARMA model formulation with 1-step temporal influence alongside broad spatial influence is preferred strongly to the non-spatial ARMA and SARMA models even with 3-step, daily and weekly temporal effects.

In addition, comparing the ACF and PACF plots of these models shows that the VARMA model performs better in representing the temporal structure in the data: statistically significant lags in the residuals of the SARMA models indicate that this model cannot represent all variations in atmospheric pollutants.

In terms of running time, SARMA models are more quickly than VARMA. However, it is worth mentioning that the SARMA models should be run separately for each station, while VARMA

552 models are run with a vector of all stations. VARMA was developed in UCL Myriad High
553 Throughput Computing Facility with CPU @8.60 GHz, in 81 hours. SARMA models are
554 developed for each station in less than 30 minutes with CPU @1.90 GHz.

555 ### 4.4. Validation

556 To test the performance of the preferred VARMA model formulation, internal cross-validation
557 was conducted to assess the effectiveness in estimating values observed at stations that were
558 omitted from the estimation process. The Kriging method (Ogundare, 2018; Warf, 2014; Wiart,
559 2016) of spatial interpolation was selected as a comparison for the results of the VARMA
560 model for all pollutants. In the Kriging method, the value of each location is estimated as a
561 weighted combination of values $(x_1,...,x_W)$ observed at a set of W neighbouring locations:

562
$$\hat{z}(x_0)=\sum_{w=1}^{W}\lambda_w z(x_w)$$
$$\sum_{w=1}^{W}\lambda_w=1 \text{ and } \lambda_w\geq 0,\ (1\leq w\leq W) \quad (19)$$

563 where $\hat{z}(x_0)$ is the estimated value at location x_0, $z(x)$ is the observed value at location x
564 and the λ is the vector of weights of the neighbouring locations.

565 The Kriging method was applied to estimate the values of the four pollutants during the year
566 2017 at the 20 stations across London that have all four pollutants measurements available.
567 For each station, the reciprocal distance between stations was calculated and used as the
568 basis for an initial value of the weighting vector λ, then a constrained optimization was used
569 to calculate weights that minimise the sum of squared deviations from the values at location
570 x_0. As an example, the initial and estimated weights for the four pollutants at station BM0 are
571 presented in Figure 11, with the other stations arranged in order of increasing distance.

572 This shows that the initial values based on reciprocal distance were excessively focused:
573 whilst all but one of the optimal values for each pollutant decreased with distance, they did so
574 less rapidly than the initial ones. Furthermore, the values for NO_2 decrease most sharply with

distance whilst those for PM_{10} decrease least: this shows that NO_2 is more localised whilst PM_{10} more uniformly distributed across London.

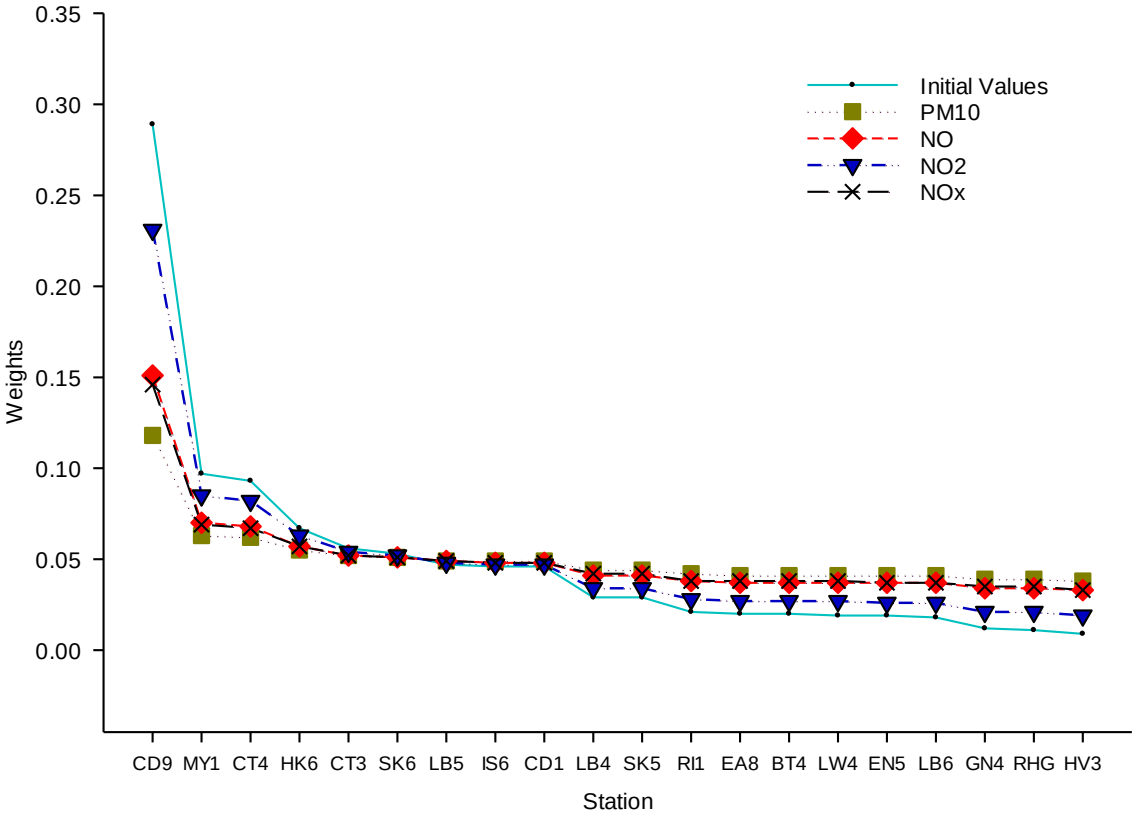


Figure 11: Kriging initial weights (reciprocal distance) and estimated weights for four pollutants

In the validation test, each station was selected in turn as the test station designated by location x_0, then the values there were estimated by each of the Kriging and VARMA models applied to data for the remaining stations. To compare the results of this, the root mean squared error (RMSE), e', for each pollutant type was calculated as:

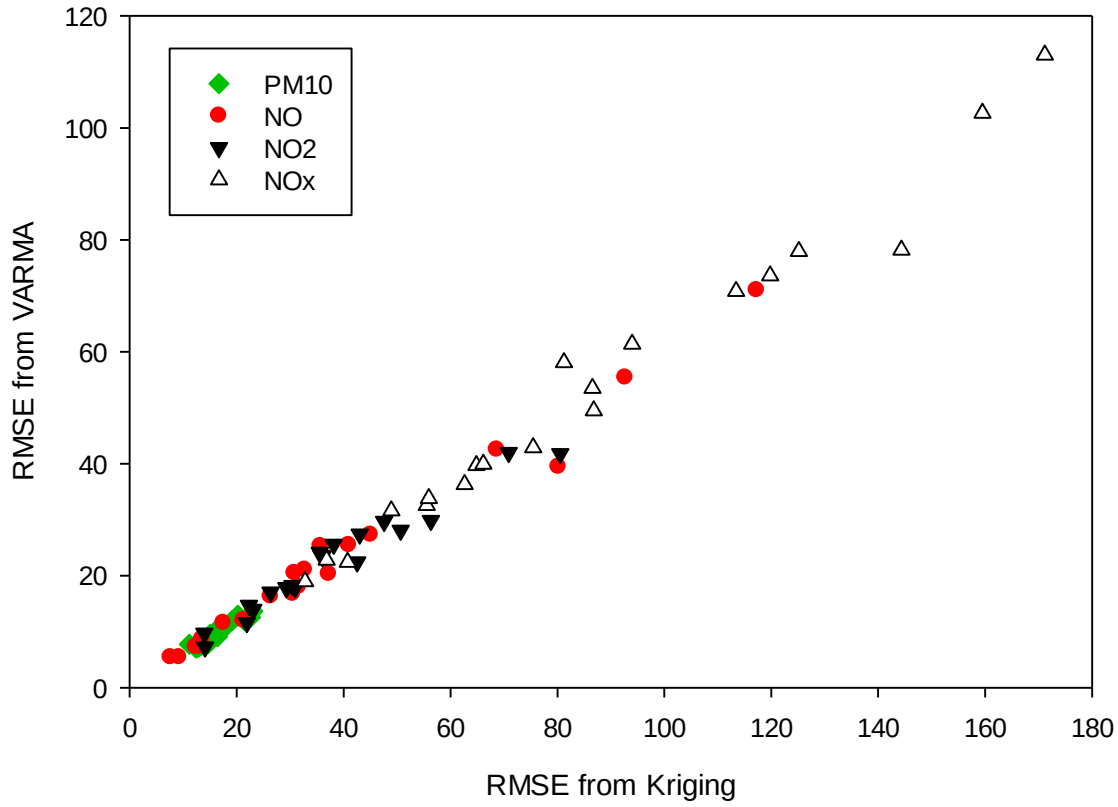
$$e' = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \tilde{y}_t)^2} \qquad (20)$$

where n is the number of observations at the test station, y_i is the observed value (measurement) and $\hat{y}_i$ is the corresponding value estimated by the model under investigation. The results for the Kriging and VARMA models are presented in Table 9 and Figure 12.

588

Table 9: RMSE of estimation for Kriging and VARMA

Station	Model	e' (μgm^{-3})			
		PM_{10}	**NO**	**NO_2**	**NO_x**
BM0	Kriging	20.22	13.55	29.3	62.63
	VARMA	12.96	8.71	17.84	36.29
CT4	Kriging	11.13	92.68	43	125.16
	VARMA	7.76	55.44	27.40	77.94
CT3	Kriging	17.42	12.95	30.96	48.89
	VARMA	10.36	7.25	17.89	31.60
MY1	Kriging	22.59	117.26	56.32	144.33
	VARMA	12.58	71.03	29.84	78.19
CD9	Kriging	14.67	80.16	70.85	171.16
	VARMA	7.86	39.46	42.00	113.03
SK6	Kriging	16.4	9.2	21.89	36.8
	VARMA	9.09	5.42	11.62	22.80
LB5	Kriging	19.22	40.95	42.54	64.81
	VARMA	12.03	25.47	22.45	39.70
HK6	Kriging	14.95	37.2	50.68	113.41
	VARMA	8.44	20.30	28.09	70.83
CD1	Kriging	13.84	45.03	22.3	93.96
	VARMA	7.63	27.30	14.70	61.38
IS6	Kriging	12.91	7.61	14.08	40.72
	VARMA	8.17	5.44	7.26	22.42
LW4	Kriging	15.03	31.6	23.08	86.54
	VARMA	9.52	18.06	13.98	53.49
SK5	Kriging	12.41	26.32	35.56	55.92
	VARMA	7.09	16.31	24.16	33.80
LB4	Kriging	21.45	30.8	80.52	159.49
	VARMA	11.97	20.48	41.77	102.65
RI1	Kriging	13.32	21.22	13.96	55.52
	VARMA	7.50	12.02	9.72	32.55
EA8	Kriging	15.97	17.51	36.02	66.11
	VARMA	10.07	11.58	23.39	39.94
BT4	Kriging	23.04	68.62	47.58	119.8
	VARMA	13.67	42.50	29.68	73.58
EN5	Kriging	14.99	32.72	38.14	86.79
	VARMA	8.75	21.07	25.60	49.47
RHG	Kriging	12.81	30.46	30.39	75.44
	VARMA	7.72	16.79	18.24	42.87
LB6	Kriging	15.41	12.34	26.34	32.77
	VARMA	9.15	7.27	17.07	18.95
GN4	Kriging	12.84	35.65	35.85	81.21
	VARMA	7.22	25.27	23.60	58.10
Average	**Kriging**	**16.03**	**38.19**	**37.47**	**86.07**
	VARMA	**9.48**	**22.86**	**22.32**	**52.98**



589

590 *Figure 11: VARMA vs Kriging RMSE (µgm^{-3})*

591 The RMSE values e' for the VARMA model are smaller than the corresponding ones for
592 Kriging in each case, with typical reduction of about 40%. This indicates the substantially better
593 performance of VARMA compared to Kriging.

594 5. Discussion

595 Continuous time series of data are important in estimating statistical models with temporal
596 components. However, available measurements often have missing values for various
597 reasons stemming from detector or recording issues. Any such interruption will fragment the
598 time series of data and so impede modelling processes: this is especially influential in
599 multivariate analyses where a value from any of the variables will be influential. For this
600 reason, imputation methods are valuable to complete the time series by estimating suitable
601 values in place of missing ones. The investigation presented here considered a range of
602 approaches to this, of which multiple imputation by chained equations (MICE) was preferred.

603 This method synthesises complete time series that preserve the mean and covariance of the
604 available data. It was found to outperform the other methods tested in estimating values for
605 data withheld from estimation, both in the case of a single interval of 500 hours (about 3 weeks)
606 or multiple intervals of mean duration 24 hours (1 day) with total duration 500 hours. The
607 analysis of variance (ANOVA) of the London air quality dataset (Table 3) shows that the air
608 pollution measurements vary across the zones of London (central London, LEZ and greater
609 London) but with greater variability among the individual stations. Statistical analysis of this
610 variation established that it can be represented efficiently by associating it solely with the
611 individual sites. Beyond the substantial variation of the pollutants among the sites, the greater
612 amount that occurs over time at each of them shows the need for a time-series approach.
613 Hence, the temporal autoregressive moving average models (ARMA, and SARMA with
614 seasonal component) and the spatio-temporal vector autoregressive moving average
615 (VARMA) models were developed for comparison. The effectiveness of the temporal (ARMA
616 and SARMA) models was assessed using a form of the Bayes information criterion BIC'
617 shown in Table 8. This confirms that using a seasonal component (either weekly or daily)
618 improves the performance of the model, reflecting repeated weekly and diurnal variation in
619 generation of these pollutants. Of these, the weekly SARMA model for each of the pollutant
620 types performs better than the corresponding daily one, indicating that atmospheric pollution
621 concentrations at each time t can be estimated more accurately by considering its value at
622 the corresponding time one week beforehand (t-168) rather than on the previous day (t-24).

623 The spatio-temporal VARMA model developed here uses data from all sites in the study area,
624 but at the previous time only. This was found to have better performance than either of the
625 SARMA models forms even though they use data from earlier times in addition to the previous
626 one. This preference for the VARMA formulation is supported by substantially better goodness
627 of fit to the data as assessed by the BIC values shown in Table 8. This outcome is further
628 supported by the ACF and PACF plots of the residuals of the SARMA and VARMA models
629 (Figure 6 and 10, respectively). The presence of statistically significant correlations at many

lags for the SARMA models in Figure 6 shows they are not effective in describing the temporal structure of the data. By contrast, the absence of most such correlations for the VARMA models in Figure 10 shows that by including a spatial element, this model is effective in capturing the temporal structure.

The performance of the spatio-temporal VARMA and the purely spatial Kriging methods can be compared by estimating pollution at sites not used in their estimation. This was undertaken in turn for each of the 20 stations across London that have all four pollutants measurements available. The resulting root mean square error (RMSE) values e' in Table 9 show that the VARMA model is typically about 40% more accurate than the corresponding Kriging one. This indicates the substantially better performance of the VARMA model compared to the Kriging technique.

6. Conclusion

The present study considered the development of atmospheric pollution in London in the calendar year 2017 during which no major air quality measures were implemented. The pollutants analysed were Nitric oxide (NO), Nitrogen dioxide (NO_2), oxides of Nitrogen (NO_x) and Particulate Matter with diameter less than 10 micrometres (PM10). The data were based on hourly measurements of atmospheric concentrations at 30 stations, each in one of three zones (central London, LEZ, greater London), throughout the year (8760 hours). The investigation developed spatio-temporal models of the pollution and compared these with corresponding temporal (time-series) and spatial formulations.

Statistical analysis of the London 2017 dataset showed that the atmospheric concentrations of pollutants varied across the three zones. However, there is greater variation among the stations. Treating the stations as sited at a sample of locations in London, the variability across the zones is consistent with the combined variability at the stations within each of them. From this, the conclusion is that measurements at an additional station would be expected to vary similarly to the stations irrespective of its zone of location.

656 The absence from the datasets of some observations is usual for several reasons. The dataset
657 used here and hence the analysis was found to be affected by this, thus hampering the present
658 time-series analysis. Several methods were investigated to impute values for the missing
659 observations and so complete the dataset. Of these, the multiple imputation by chained
660 equations (MICE) approach was found to be more effective than others. The methods were
661 tested by withholding observed data artificially from test intervals to compare with imputed
662 ones: the MICE approach generated the most accurate estimates. The MICE approach has
663 the additional benefit of preserving the mean and covariance of the dataset that it completes.

664 The goodness of fit of a spatio-temporal vector autoregressive moving average (VARMA) and
665 purely temporal seasonal autoregressive moving average (SARMA) models were compared
666 using the Bayes information criterion (BIC). According to the results of this, the VARMA model
667 performs substantially better than any of the SARMA ones. This shows that the distribution of
668 atmospheric pollutants is strongly influenced by the previous hourly observation at the same
669 and nearby locations: this leads to more accurate estimation than use of purely temporal data
670 from the site even including observations from the previous day or week.

671 In addition to tests on goodness of fit using BIC, diagnostic tests were undertaken on the time
672 series of model residuals using autocorrelation (ACF) and partial auto correlation (PACF). The
673 results of this show that the spatio-temporal VARMA model is more successful in representing
674 temporal variations in atmospheric pollution than any of the purely temporal SARMA models.
675 This is indicated by absence of residual temporal structure in VARMA while this is substantial
676 in both the daily and weekly SARMA models.

677 The strength of the spatial influence identified by the VARMA model is greater at the periphery
678 of London (greater London zone) than in the central London zone. This spatial influence is
679 more extensive for particulates PM_{10}, showing that it is more diffused, and least for NO_2,
680 indicating that this is more localised.

681 Validation tests were undertaken of the VARMA model in estimating values of data reserved
682 at each of several sites in turn that were not used in its estimation. The results of this showed

683 that the spatio-temporal VARMA performs substantially better than the purely spatial Kriging
684 method even when that was fitted optimally to instantaneous values at the test site: the root
685 mean squared error (RMSE) for the VARMA model in estimation is about 40% less than that
686 of the corresponding fitted Kriging one.

687 The conclusion from this investigation is that spatio-temporal modelling is effective for
688 atmospheric concentrations of pollution. Comparison with purely temporal models shows
689 presence of spatial influences, thus verifying the distribution of pollution due to dispersion
690 within the atmosphere and mobility of some sources. Comparison with purely spatial models
691 shows presence of temporal influences, thus verifying the persistence of pollutants in the
692 atmosphere and continuing emission from their sources. Whilst the strength of these
693 influences varies among atmospheric pollutants, the importance of spatio-temporal modelling
694 and the structure of the VARMA used for this applies to all those investigated here.

695 This investigation has established the VARMA model as appropriate for use as a basis for the
696 evaluation of interventions such as the Ultra-Low Emissions Zone on air quality across
697 London. This approach will need further parameters to represent the effects of interventions.

698 **Acknowledgments:** The authors acknowledge the use of the UCL Myriad High Throughput
699 Computing Facility (Myriad@UCL), and associated support services in the completion of this
700 work. This research was funded by the Lloyd's Register Foundation (LRF), which helps to
701 protect life and property by supporting engineering-related education, public engagement and
702 the application of research.

703 References

704 *Air quality standards*. (2009). Springer Series in Reliability Engineering.
705 https://doi.org/10.1007/978-1-84882-602-1_4
706 Anzai, Y. (2012). *Pattern Recognition and Machine Learning (1st edition).*
707 Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.
708 Box, G. (2008). *Time Series Analysis: Forecasting and Control (4th edition).*
709 Cavicchioli, M. (2016). Statistical Analysis Of Mixture Vector Autoregressive Models.
710 *Scandinavian Journal of Statistics*, *43*(4), 1192–213.

Dias, G. F., and Kapetanios, G. (2018). Estimation and forecasting in vector autoregressive moving average models for rich datasets. *Journal of Econometrics*, *202*(1), 75–91.

Donnelly, A., Misstear, B., and Broderick, B. (2015). Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmospheric Environment*, *103*, 53–65. https://doi.org/10.1016/j.atmosenv.2014.12.011

Du, S, Li, T., Yang, Y., and Horng, S.-J. (2019). Deep Air Quality Forecasting Using Hybrid Deep Learning Framework. *IEEE Transactions on Knowledge and Data Engineering*, 1. https://doi.org/10.1109/TKDE.2019.2954510

Du, S, Li, T., Yang, Y., and Horng, S.-J. (2020). Multivariate time series forecasting via attention-based encoder–decoder framework. *Neurocomputing*, *388*, 269–79. https://doi.org/10.1016/j.neucom.2019.12.118

Eilers, M., and Möbus, C. (2011). Learning the Relevant Percepts of Modular Hierarchical Bayesian Driver Models Using a Bayesian Information Criterion. In *Digital Human Modeling,* 6777, 463–72. Springer: Berlin, Heidelberg.

Environmental Research Group Kings College London. (2016). *London Air*. https://www.londonair.org.uk/london/asp/nowcast.asp

Freeman, B. S., Taylor, G., Gharabaghi, B., and Thé, J. (2018). Forecasting air quality time series using deep learning. *Journal of the Air & Waste Management Association*, *68*(8), 866–86. https://doi.org/10.1080/10962247.2018.1459956

Gamerman, D. (2006). *Markov chain Monte Carlo : stochastic simulation for Bayesian inference (2nd edition).*

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., R Core Team, Ihaka, R., Reid, D., Shaub, D., Tang, Y., and Zhou, Z. (2020). forecast: forecasting functions for time series and linear models. R package version 8.13. *Http://CRAN.R-Project.Org/Package=forecast*, 144. https://cran.r-project.org/package=forecast

Hyndman, R. J., and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(3), 1–22.

Kumar, A., and Goyal, P. (2011). Forecasting of daily air quality index in Delhi. *Science of the Total Environment*, *409*(24), 5517–5523. https://doi.org/10.1016/j.scitotenv.2011.08.069

Ledolter, J. and Hogg, R. V. (2009) *Applied Statistics for Engineers and Physical Scientists (3rd edition). Prentice Hall*

Lee, S.-Y. (2007). *Structural equation modeling a Bayesian approach. Wiley*.

London Environmet Strategy. (2018). https://www.london.gov.uk/sites/default/files/les_appendix_2_-_evidence_base_0_0.pdf

Ochoa-Hueso, R., Munzi, S., Alonso, R., Arróniz-Crespo, M., Avila, A., Bermejo, V., Bobbink, R., Branquinho, C., Concostrina-Zubiri, L., Cruz, C., Cruz de Carvalho, R., De Marco, A., Dias, T., Elustondo, D., Elvira, S., Estébanez, B., Fusaro, L., Gerosa, G., Izquieta-Rojano, S., Theobald, M. R. (2017). Ecological impacts of atmospheric pollution and interactions with climate change in terrestrial ecosystems of the Mediterranean Basin: Current research and future directions. *Environmental Pollution*, *227*, 194–206. https://doi.org/10.1016/j.envpol.2017.04.062

Ogundare, J. O. (2018). Introduction to Least Squares Collocation and the Kriging Methods. In *Understanding Least Squares Estimation and Geomatics Data Analysis,* 6*13–33. Wiley.* https://doi.org/10.1002/9781119501459.ch16

Orach, J., Rider, C. F., and Carlsten, C. (2021). Concentration-dependent health effects of

air pollution in controlled human exposures. *Environment International*, *150*, *106424*. https://doi.org/10.1016/j.envint.2021.106424

Pandis, N. (2016). Multiple linear regression analysis. In *American Journal of Orthodontics and Dentofacial Orthopedics, 149(4), 581.* https://doi.org/10.1016/j.ajodo.2016.01.012

Pfeffer, P. E., Mudway, I. S., and Grigg, J. (2020). Air Pollution and Asthma: Mechanisms of Harm and Considerations for Clinical Interventions. *Chest*. https://doi.org/10.1016/j.chest.2020.10.053

Raghunathan, T. (2015). *Missing data analysis in practice.*

Resche-Rigon, M., and White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research , 27(6), 1634-49.*

Romanowicz, R., Young, P., Brown, P., and Diggle, P. (2006). A recursive estimation approach to the spatio-temporal analysis and modelling of air quality data. *Environmental Modelling and Software*, *21*(6), 759–69. https://doi.org/10.1016/j.envsoft.2005.02.004

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, *85–117.* https://doi.org/10.1016/j.neunet.2014.09.003

Shaddick, G., and Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *51(3), 351–72.* https://doi.org/10.1111/1467-9876.00273

Shaub, D. (2020). Fast and accurate yearly time series forecasting with forecast combinations. *International Journal of Forecasting*, *36(1), 116–20.*

Shindell, D. T., Faluvegi, G., Koch, D. M., Schmidt, G. A., Linger, N., and Bauer, S. E. (2009). Improved attribution of climate forcing to emissions. *Science*, *326(5953), 716–8.* https://doi.org/10.1126/science.1174760

Simionescu, M. (2013). The use of VARMA models in forecasting macroeconomic indicators. *Economics & Sociology*, *6(2), 94–102.*

Suresh, M., Taib, R., Zhao, Y., and Jin, W. (2019). Sharpening the BLADE: Missing Data Imputation Using Supervised Machine Learning. In *AI 2019: Advances in Artificial Intelligence, 215–27.*

Tsay, R. S., and Wood, D. (2018). MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models. R package version 1.0. In *https://CRAN.R-project.org/package=MTS*.

van Buuren, S., and Groothuis-Oudshoorn, C. G. M. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45(3).*

Warf, B. (2014). Spatial Statistics. In *Encyclopedia of Geography, 429–53. Wiley.* https://doi.org/10.4135/9781412939591.n1074

Wei, W. W. S., Watts, A. S., Wang, L. L., Little, T. D., and Anderson, R. A. (2013). Time Series Analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: 2*(1). Oxford University Press.

Wiart, J. (2016). Stochastic Dosimetry. In *Radio-Frequency Human Exposure Assessment, 119–55. Wiley.*

Wood, S. N. (2017). *Generalized additive models : an introduction with R (2nd edition).*

Xie, J. (2017). Deep Neural Network for PM2.5 Pollution Forecasting Based on Manifold Learning. *2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), 236–40.* https://doi.org/10.1109/SDPC.2017.52

Xu, L., Williams, L. R., Young, D. E., Allan, J. D., Coe, H., Massoli, P., Fortner, E., Chhabra, P., Herndon, S., Brooks, W. A., Jayne, J. T., Worsnop, D. R., Aiken, A. C., Liu, S., Gorkowski, K., Dubey, M. K., Fleming, Z. L., Visser, S., Prévôt, A. S. H., and Ng, N. L. (2016). Wintertime aerosol chemical composition, volatility, and spatial variability in the greater London area. *Atmospheric Chemistry & Physics*, *16(2), 1139–60.* https://doi.org/10.5194/acp-16-1139-2016

Young, D. E., Allan, J. D., Williams, P. I., Green, D. C., Harrison, R. M., Yin, J., Flynn, M. J., Gallagher, W. M., and Coe, H. (2015). Investigating a two-component model of solid fuel organic aerosol in London: processes, PM1 contributions, and seasonality. *Atmospheric Chemistry and Physics*, *15(5), 2429–43.* https://doi.org/10.5194/acp-15-2429-2015

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of Translational Medicine*, *4(2), 30.*