Electronic Supplementary Material (ESI) for Environmental Science: Processes & Impacts.

Supplementary Information (SI)

for

Outlier detection and gap filling methodologies for low-cost air quality measurements

Thor-Bjorn Ottosen, Prashant Kumar[1]

Global Centre for Clean Air Research (GCARE), Department of Civil and Environmental Engineering, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, United Kingdom

[1] Corresponding author. Address as above. Tel.: +44 1483 682762; Fax: +44 1483 682135; E-mail addresses: P.Kumar@surrey.ac.uk, Prashant.Kumar@cantab.net

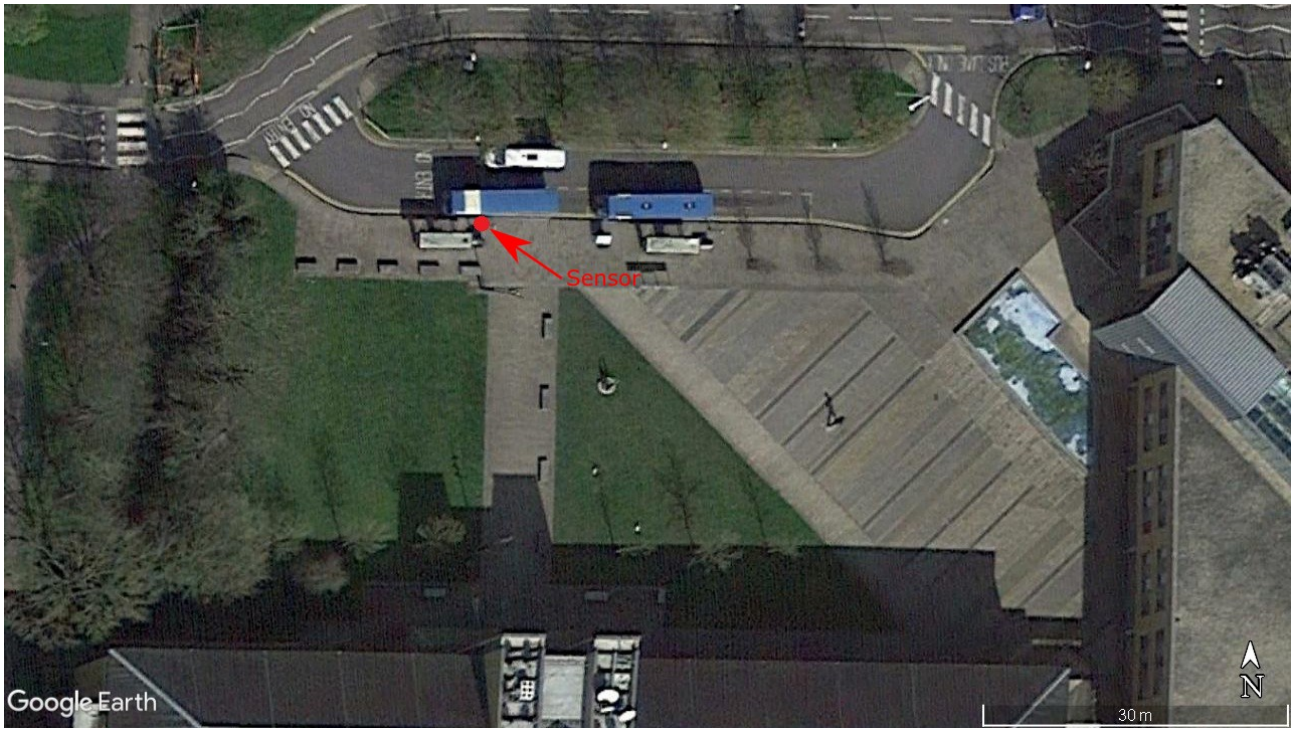

Figure S1. Aerial photo of the bus stop, where the measurements have been made. The location of the AQMesh pod is marked with a red dot. Data source: Google Earth.

S1. Pruned Exact Linear Time algorithm

Given an ordered sequence of data (in this case air quality measurements) $y_{1:n} = (y_1,\ldots,y_n)$, the aim of the algorithm is to find the number of changepoints m with positions $\tau_{1:m} = (\tau_1,\ldots,\tau_m)$. To this end, the following expression is minimised:

$$\sum_{i=1}^{m+1}\left[\mathrm{C}(y_{(\tau_{i-1}+1):\tau_i})\right] + \beta f(m) \qquad (1)$$

Where C is a cost function for segment i and $\beta f(m)$ is a penalty function to prevent overfitting. In the PELT algorithm twice the negative log likelihood is used as cost function[1] and the penalty function is set equal to the number of changepoints. Calculating this expression for all possible combinations would yield 2^{n-1} calculations[2]. The PELT algorithm is a faster way to minimise equation 1, and the interested reader is referred to[1] for details.

S2. *k*-Nearest Neighbor outlier detection

In k-Nearest Neighbor outlier detection, "The anomaly score of a data instance is defined as its distance to its k[th] nearest neighbor in a given data set"[3]. In the present study the Euclidian distance is used:

$$d(\vec{p},\vec{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} \quad (2)$$

Where $\vec{p}$ and $\vec{q}$ are two air quality measurements with the species representing the components of the vectors.

S3. Regression based outlier detection

In the present implementation of regression based outlier detection, an Auto-Regressive Integrated Moving Average (ARIMA) model is fitted to the data. The equation for the ARIMA model is[4]:

$$\phi(B)(1 - B^d)y_t = c + \theta(B)\varepsilon_t \quad (3)$$

Where y_t is the time series of air quality measurements, B is the backshift operator, $\phi(z)$ and $\theta(z)$ are polynomials of order p and q respectively. The outlier score is subsequently defined as the residual between the fitted model and the measurements.

S4. Multivariate Linear regression

Given an air quality measurement with multiple species $\vec{p} = (p_1,\ldots,p_n)$, where p_1 is missing, the multivariate linear regression fits a function for $p_1$as a function of the remaining species:

	$p_1 = \alpha_1 p_2 + \alpha_2 p_3 + \ldots + \alpha_{n-1} p_n + \beta$	(4)

Where α and β are regression coefficients.

Table S1. Summary statistics for the multivariate gap filling methods. Δ is the difference between the parameter for the distribution without gaps and the parameter for the filled distribution. The following statistics are presented: (μ) mean, (σ) standard deviation, (γ) skewness, (κ) kurtosis, (R) correlation, (d) index of agreement, (RMSE) Root Mean Square Error, (MAE) Mean Average Error.

	Δμ	Δσ	Δγ	Δκ	R	d	RMSE	MAE
NO								
No gap filling	1.35	1.48	0.08	0.63				
Linear interpolation	0.36	0.15	0.02	0.17	0.79	0.88	14.25	8.35
Linear regression	0.82	0.54	0.03	0.46	0.76	0.83	16.02	12.44
Neural Networks	1.57	0.73	0.15	0.96	0.72	0.80	22.13	15.80
NO_2								
No gap filling	0.96	0.53	0.05	0.19				
Linear interpolation	0.23	0.10	0.01	0.05	0.82	0.90	8.32	5.91
Linear regression	0.73	0.77	0.05	0.49	0.11	0.39	16.53	13.34
Neural Networks	1.02	0.30	0.09	0.30	0.22	0.55	21.71	16.37
SO_2								
No gap filling	0.12	0.10	0.04	0.48				
Linear interpolation	0.04	0.03	0.02	0.08	0.96	0.98	1.69	1.07
Linear regression	0.13	0.24	0.04	0.67	0.86	0.86	3.42	2.97
Neural Networks	0.22	0.13	0.04	0.39	0.87	0.89	2.98	1.90
CO								
No gap filling	3.26	1.86	0.08	0.92				
Linear interpolation	0.21	0.28	0.01	0.05	0.96	0.98	23.81	15.90
Linear regression	2.34	6.31	0.11	1.59	0.68	0.66	76.36	55.58
Neural Networks	2.10	2.94	0.08	0.74	0.81	0.85	53.95	38.04
O_3								
No gap filling	0.82	0.47	0.08	0.19				
Linear interpolation	0.13	0.11	0.01	0.05	0.94	0.97	7.17	5.18
Linear regression	0.51	0.79	0.04	0.42	0.78	0.84	15.67	12.47
Neural Networks	0.61	0.35	0.05	0.17	0.86	0.91	12.79	9.40

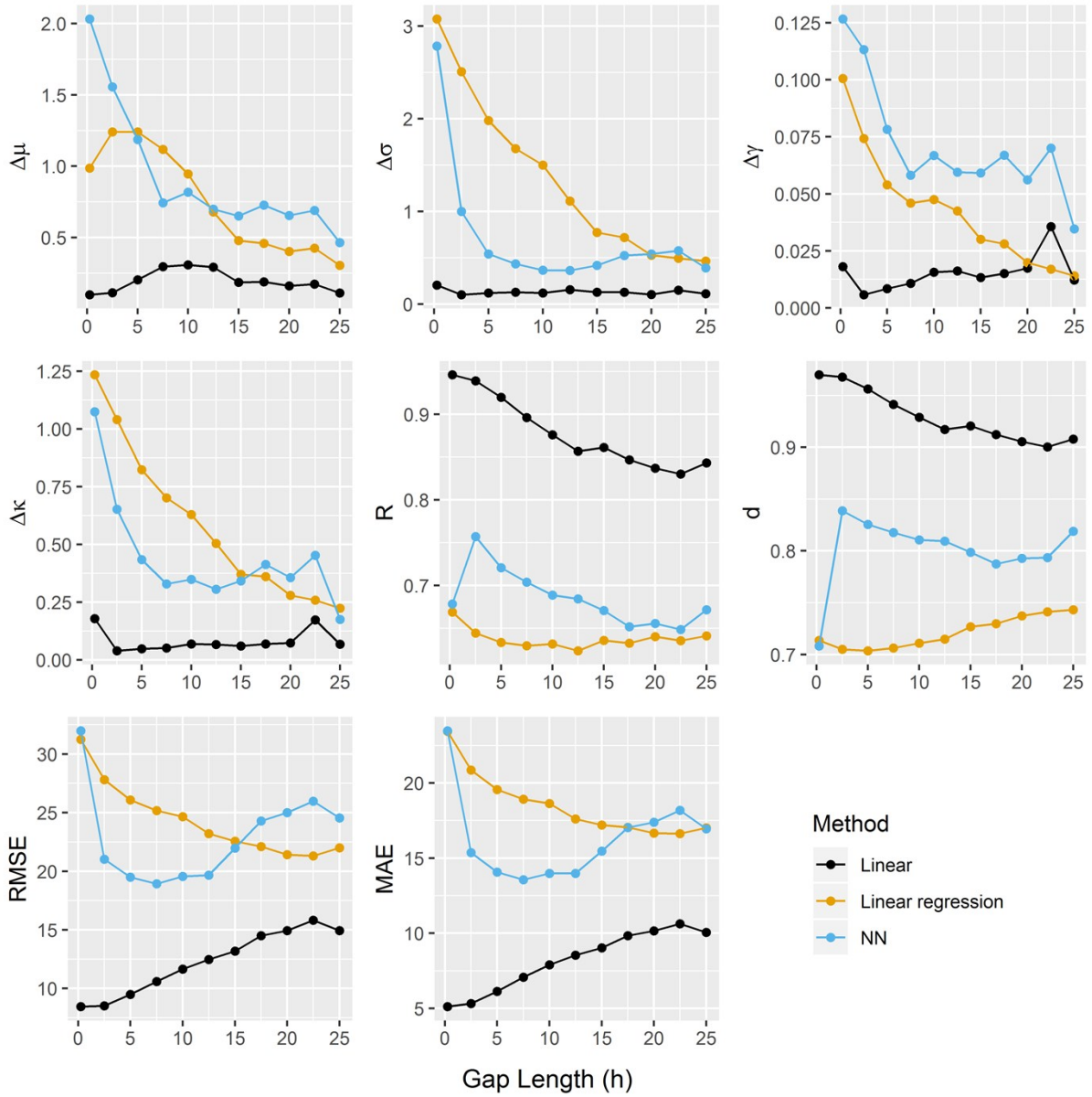

Figure S2. Performance statistics for the two multivariate gap filling methods and linear interpolation as a function of gap length in hours. Δ is the difference between the parameter for the distribution without gaps and the parameter for the filled distribution. The following statistics are presents: (μ) mean, (σ) standard deviation, (γ) skewness, (κ) kurtosis, (R) correlation, (d) index of agreement, (RMSE) Root Mean Square Error, (MAE) Mean Average Error.

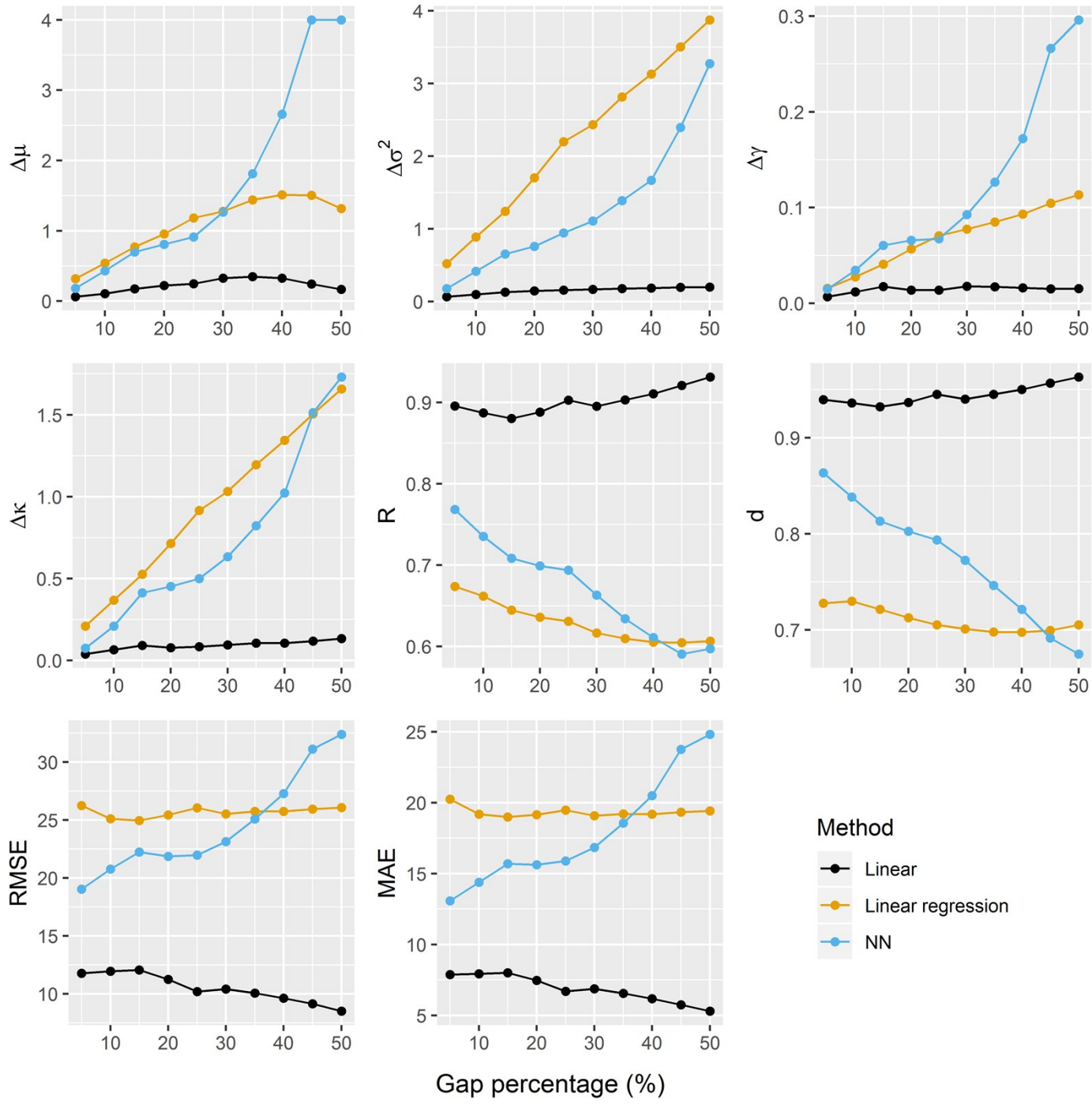

Figure S3. Performance statistics for the two multivariate gap filling methods and linear interpolation as a function of gap percentage. Δ is the difference between the parameter for the distribution without gaps and the parameter for the filled distribution. The following statistics are presents: (μ) mean, (σ) standard deviation, (γ) skewness, (κ) kurtosis, (R) correlation, (d) index of agreement, (RMSE) Root Mean Square Error, (MAE) Mean Average Error.

References

1. R. Killick, P. Fearnhead and I. A. Eckley, Optimal Detection of Changepoints With a Linear Computational Cost, *J. Amer. Statistical Assoc.*, 2012, **107**, 1590-1598.
2. R. Killick and I. A. Eckley, changepoint: An R Package for Changepoint Analysis, *Journal of Statistical Software*, 2014, **58**, 19.

3. V. Chandola, A. Banerjee and V. Kumar, Anomaly detection: A survey, *ACM Comput. Surv.*, 2009, **41**, 1-58.
4. R. J. Hyndman and Y. Khandakar, Automatic Time Series Forecasting: The forecast Package for R, *Journal of Statistical Software*, 2008, **27**, 22.