Imputation of missing $PM_{2.5}$ observations in an air-quality monitoring network of air quality monitoring stations by a new *k*NN method

Formatted: Font: Italic, Complex Script Font: Italic

Supplementary Material

Idit Belachsen and David M. Broday*

Faculty of Civil and Environmental Engineering, Technion, Israel Institute of Technology, Haifa, Israel

Corresponding author:

David Broday, Faculty of Civil and Environmental Engineering, Technion, Israel
E-mail: dbroday@technion.ac.il

Table S1. The ID numbers and names of the 59 AQM stations in the imputed dataset. In bold are the 36 AQM stations that have accumulated missing observations of ≤4 years.

AQM station ID and name			
1: **Afula**	2: **Antokolsky**	3: Ehad-Haam	4: **Holon**
5: **Ironi D**	6: **Kvish 4**	7: **Petah-Tikva rd.**	8: Rakevet Hagana
9: Rakevet Hashalom	10: Rakevet Komemyut	11: Rakevet Yoseftal	12: **Remez**
13: Rishon-Lezion	14: Tahana Merkazit	15: **Yad Lebanim**	16: **Yefet Yafo**
17: **Ahuza G**	18: **Atzmaut B**	19: Begin	20: Igud
21: Kakal	22: **Kiryat Ata**	23: **Kiryat Bialik**	24: **Kiryat Binyamin**
25 :**Kiryat Tivon**	26: **Nave-Shaanan**	27: **Nesher**	28: Park Hacarmel
29: **Bar-Ilan**	30: **Efrata**	31: Nave-Ilan	32: **Ashdod Igud**
33: **Ashkelon South**	34: **Dalya**	35: **Gedera**	36: **Gvaraam**
37: Hayovel	38: **Kiryat Malahi**	39: **Nir-Israel**	40: Ofek
41: **Ort**	42: **Rova TV**	43: **Sderot**	44: **Sde-Yoav**
45: **Yahalom**	46: **Beer-Sheva**	47: **East Negev**	48: **Kfar-Masarik**
49: Barkai	50: Hefziba	51: Kfar-Saba	52: **Pardes-Hana**
53: **Raanana**	54: Shfeya	55: Bnei-Atarot	56: Modein
57: Ashalim	58: Gush-Ezion	59: Neot-Hakikar	

Table S2. The performance metrics used in this work, with y_i the true measured value and $\hat{y}_i$ the imputed value.

Abbreviation	Term	Formula
NRMSE	Normalized root mean squared error	$\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}}{\frac{1}{N}\sum_{i=1}^{N}y_i}=\frac{RMSE}{mean(Y)}$
R^2	Coefficient of determination	$1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}$
NMAE	Normalized mean absolute error	$\frac{\sum_{i=1}^{N}\lvert y_i-\hat{y}_i\rvert}{\frac{1}{N}\sum_{i=1}^{N}y_i}=\frac{MAE}{mean(Y)}$
NMB	Normalized mean bias	$\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)}{\frac{1}{N}\sum_{i=1}^{N}y_i}=\frac{MB}{mean(Y)}$
Normalized CRMSE	Normalized centered root mean squared error	$\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}[(y_i-\bar{y})-(\hat{y}_i-\bar{\hat{y}})]^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2}}=\frac{CRMSE}{SD(Y)}$

Table S3. The tuned hyperparameters of the w*k*NN*r* model (see section 2.4.1). *k*- the number of weights accounted for, *q*- the power in Eq. 4.

Hyperparameter	*k*	*q*	Correlation type
Search space	5-100	1-2.5	Spearman/Pearson
Optimal values:			
very short	10	2.5	Spearman
short	25	2	Spearman
medium-length	30	1.5	Spearman
long	35	1.5	Spearman

Table S4. The tuned hyperparameters of the iiET model *(see section 2.4.2)*. *n_estimators*- the number of trees in the ensemble, *min_samples_leaf*- the minimum number of samples in each leaf node, *min_samples_split*- the minimum number of samples required to split a node, and *max_iter*- the number of iterations.

Hyperparameter	*n_estimators*	*min_samples_split*	*min_samples_leaf*	*max_iter*
Search space	40-100	2-10	1-10	1-10
Optimal value:				
very short	70	4	1	2
short	80	5	3	2
medium-length	80	4	2	2
long	100	9	2	2