Supplement of Atmos. Meas. Tech., 12, 5161–5181, 2019
https://doi.org/10.5194/amt-12-5161-2019-supplement

Atmospheric Measurement Techniques
Open Access
EGU

Supplement of

Gaussian process regression model for dynamically calibrating and surveilling a wireless low-cost particulate matter sensor network in Delhi

Tongshu Zheng et al.

Correspondence to: Tongshu Zheng (tongshu.zheng@duke.edu)

Reasoning behind step four of the schema for the simultaneous GPR and simple linear regression calibration model

Once the optimum $\boldsymbol{\Theta}$ for the (initial) GPR was found, we used the learned covariance function to find the mean of each low-cost node *i*'s Gaussian Distribution conditional on the remaining 30 nodes within the network (i.e., $\mu_{A|B}^{it}$) on day *t* as described mathematically in Eq. (S1)–(S4) and repeatedly did so until all 59 days' $\mu_{A|B}^{it}$ (i.e., $\boldsymbol{\mu_{A|B}^{i}}$) were found and then re-calibrated that low-cost node *i* based on the $\boldsymbol{\mu_{A|B}^{i}}$. This procedure was performed iteratively for all low-cost nodes one at a time.

$$p\left(\begin{bmatrix} r_A^{it} \\ \boldsymbol{r_B^{it}} \end{bmatrix}\right) = N\left(\begin{bmatrix} r_A^{it} \\ \boldsymbol{r_B^{it}} \end{bmatrix}; \begin{bmatrix} \mu_A^{it} \\ \boldsymbol{\mu_B^{it}} \end{bmatrix} \begin{bmatrix} \Sigma_{AA}^{it} & \boldsymbol{\Sigma_{AB}^{it}} \\ \boldsymbol{\Sigma_{BA}^{it}} & \boldsymbol{\Sigma_{BB}^{it}} \end{bmatrix}\right) \quad \text{(S1)}$$

$$r_A^{it} \left| \boldsymbol{r_B^{it}} \right. \sim N\left(\mu_{A|B}^{it}, \Sigma_{A|B}^{it}\right) \quad \text{(S2)}$$

$$\mu_{A|B}^{it} = \mu_A^{it} + \boldsymbol{\Sigma_{AB}^{it}} \boldsymbol{\Sigma_{BB}^{it}}^{-1} (\boldsymbol{r_B^{it}} - \boldsymbol{\mu_B^{it}}) \quad \text{(S3)}$$

$$\Sigma_{A|B}^{it} = \Sigma_{AA}^{it} - \boldsymbol{\Sigma_{AB}^{it}} \boldsymbol{\Sigma_{BB}^{it}}^{-1} \boldsymbol{\Sigma_{BA}^{it}} = a\ constant\ for\ low-cost\ node\ i\ regardless\ of\ day\ t = \Sigma_{A|B}^{i} \quad \text{(S4)}$$

where r_A^{it} and $\boldsymbol{r_B^{it}}$ are the daily $PM_{2.5}$ measurement(**s**) of the low-cost node *i* and the remaining 30 nodes on day *t*; μ_A^{it}, $\boldsymbol{\mu_B^{it}}$, and $\mu_{A|B}^{it}$ are the mean (**vector**) of the partitioned Multivariate Gaussian Distribution of the low-cost node *i,* the remaining 30 nodes, and the low-cost node *i* conditional on the remaining 30 nodes, respectively, on day *t*; and Σ_{AA}^{it}, $\boldsymbol{\Sigma_{AB}^{it}}$, $\boldsymbol{\Sigma_{BA}^{it}}$, $\boldsymbol{\Sigma_{BB}^{it}}$, and $\Sigma_{A|B}^{it}$ are the covariance between the low-cost node *i* and itself, the low-cost node *i* and the remaining 30 nodes, the remaining 30 nodes and the low-cost node *i*, the remaining 30 nodes and themselves, and the low-cost node *i* conditional on the remaining 30 nodes and itself, respectively, on day *t*.

The reasoning behind recalibrating each low-cost node *i* based on the $\boldsymbol{\mu_{A|B}^{i}}$ is given as follows:

The conditional log-likelihood under the Univariate Gaussian distribution on day *t* is:

$$\log p\left(r_A^{it} \left| \boldsymbol{r_B^{it}} \right.\right) = constant - 0.5 {\Sigma_{A|B}^{it}}^{-2} (r_A^{it} - \mu_{A|B}^{it})^2 \quad \text{(S5)}$$

Then the complete log-likelihood over all 59 days is therefore given by:

$$\sum_{t=1}^{59} \log p\left(r_A^{it} \left| \boldsymbol{r_B^{it}} \right.\right) = constant - positive\ constant \cdot \sum_{t=1}^{59} (r_A^{it} - \mu_{A|B}^{it})^2 \quad \text{(S6)}$$

The objective is to maximize the complete log-likelihood over all 59 days (i.e., S6), that is equivalent to minimizing the term of $\sum_{t=1}^{59} (r_A^{it} - \mu_{A|B}^{it})^2$:

$$\max_{\boldsymbol{r_A^i}} \sum_{t=1}^{59} \log p\left(r_A^{it} \left| \boldsymbol{r_B^{it}} \right.\right) = \min_{\boldsymbol{r_A^i}} \left\| \boldsymbol{r_A^i} - \boldsymbol{\mu_{A|B}^i} \right\|_2^2 \quad \text{(S7)}$$

$$\text{and } \boldsymbol{r_A^i} = \boldsymbol{Y_i \beta_i} \quad \text{(S8)}$$

where $\boldsymbol{Y_i} = \begin{bmatrix} 1 & y_{it} \\ \vdots & \vdots \\ 1 & y_{i59} \end{bmatrix}$ and $\boldsymbol{\beta_i}$ is a vector of the intercept and slope (to be learned) of the simple linear regression calibration equation for low cost node *i*.

And to minimize $\left\|\boldsymbol{Y_i\beta_i} - \boldsymbol{\mu^i_{A|B}}\right\|_2^2$ is then equivalent to optimizing a simple linear regression model to re-calibrate the raw low-cost node signals based on the mean of each node's Gaussian Distribution conditional on the remaining 30 nodes within the network (i.e., $\boldsymbol{\mu^i_{A|B}}$).

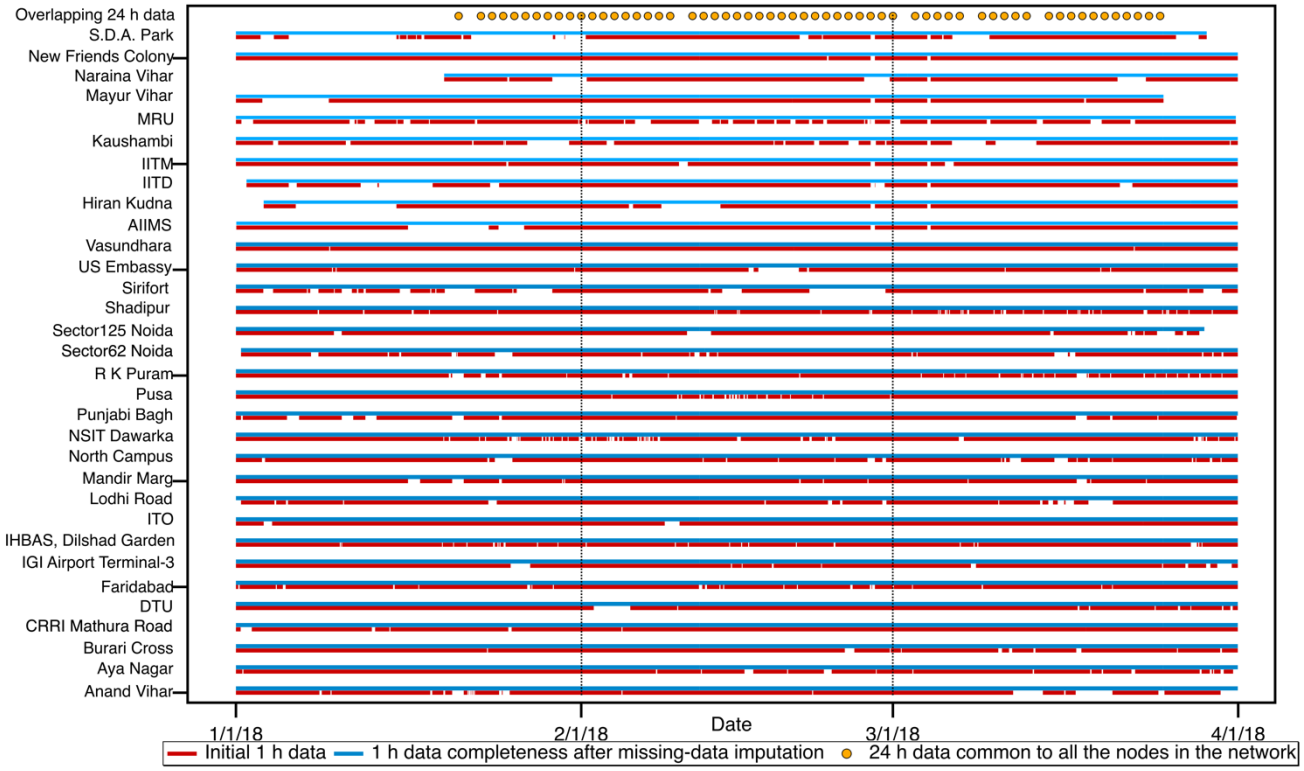

Figure S1: Periods over which 1 h data were available for each individual site before and after missing-data imputation and a total of 59 24 h aggregated observations common to all the nodes in the network used for the on-the-fly calibration feasibility test. The top 10 sites (i.e., from S.D.A. Park to AIIMS) are the low-cost sites and the remaining sites (i.e., from Vasundhara to Anand Vihar) are the reference sites. Note that there is no obvious pattern in the data missingness.

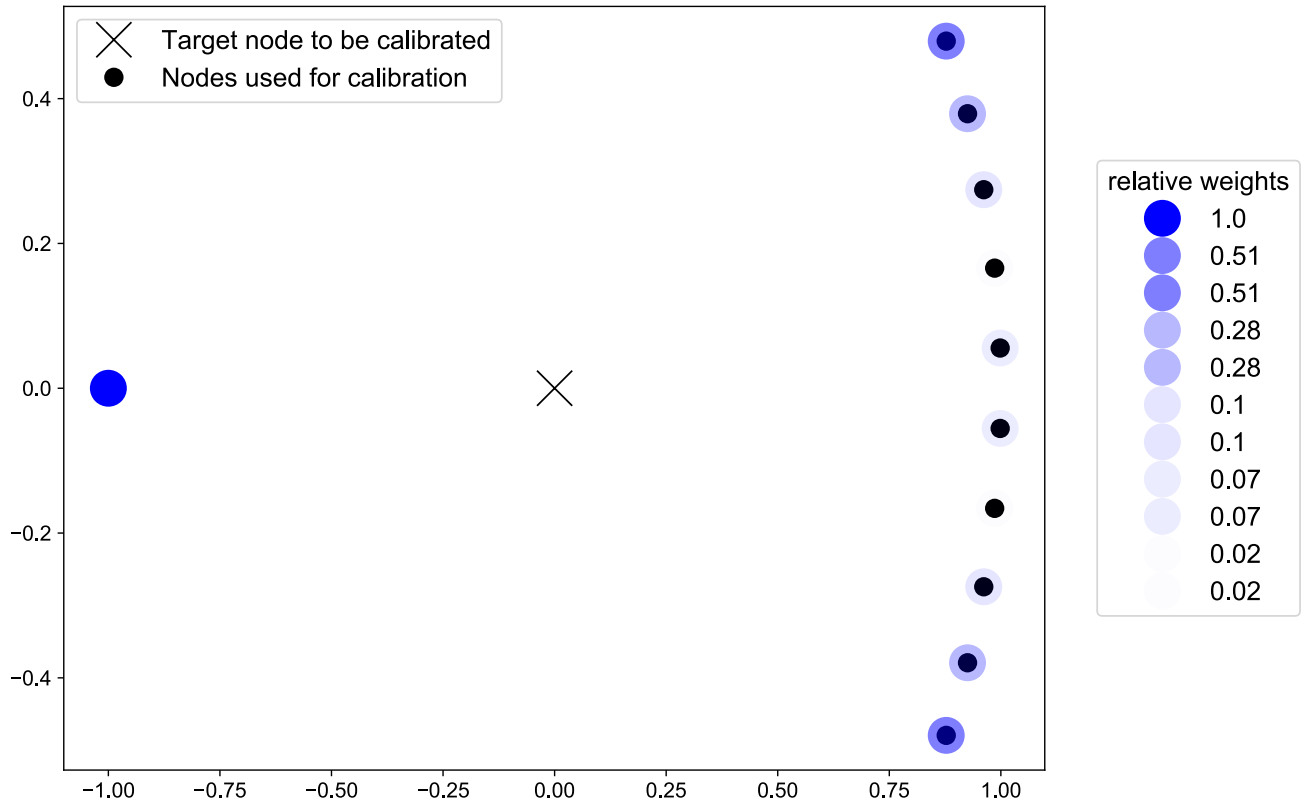

Figure S2: Simplified illustration of the relative importance (i.e., importance normalized by the max value) of each node within the network when using GPR to calibrate the target low-cost node and when all the nodes used for calibration are equally distant from the target node.

5

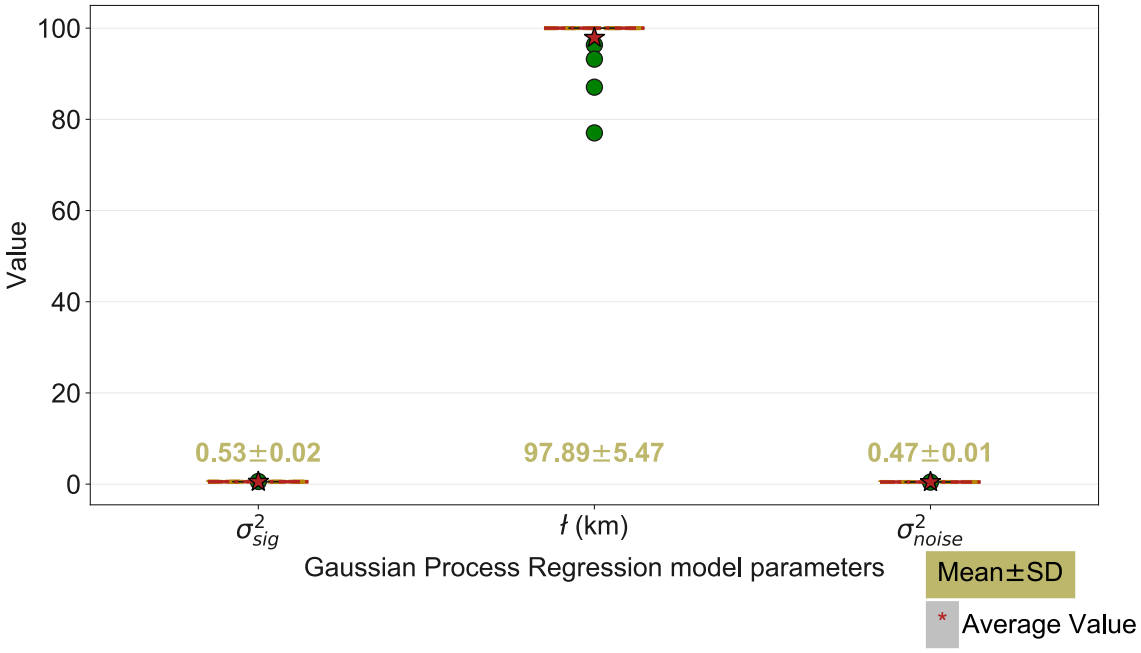

Figure S3: Box plots of the learned optimum Gaussian Process Regression model parameters including the signal variance (σ^2_{sig}), the characteristic length scale (l), and the noise variance (σ^2_{noise}) from the 22-fold leave-one-out cross-validation. The mean and SD of each parameter are superimposed on the box plots.

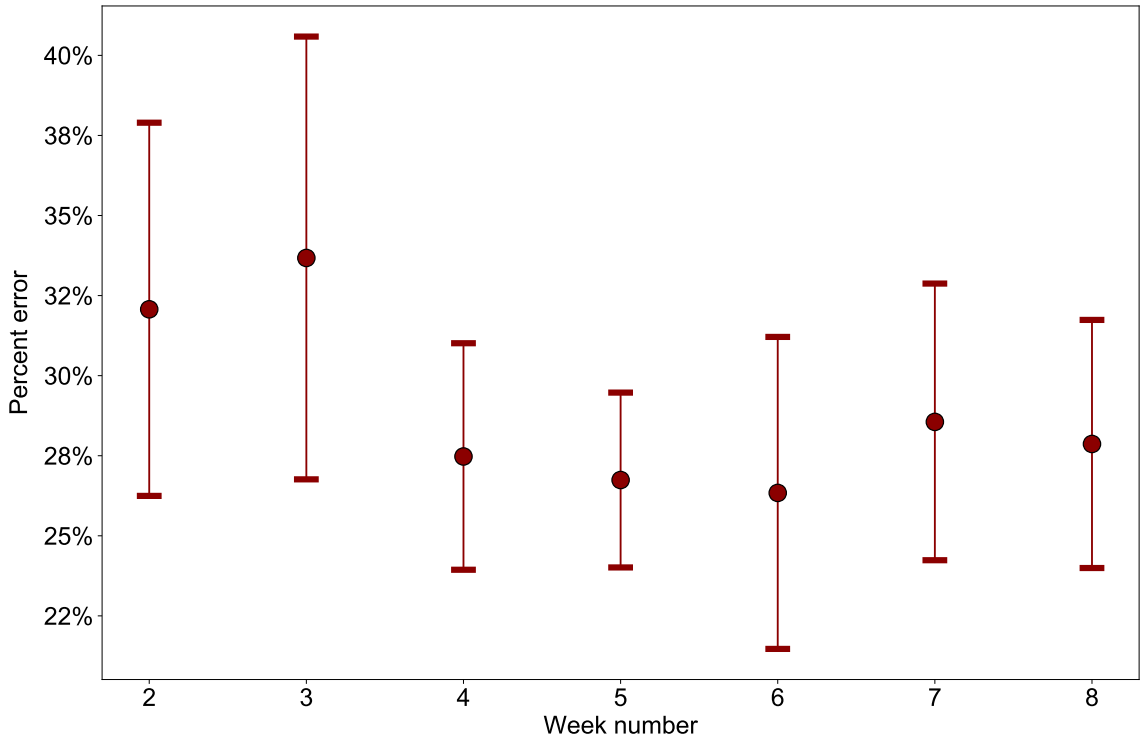

Figure S4: The 1 week-ahead prediction error of the GPR models (which were pre-trained on the current week's data) as a function of the week being predicted. The error bars represent the standard error of the mean (SEM) of the GPR prediction errors of the 22 reference nodes.

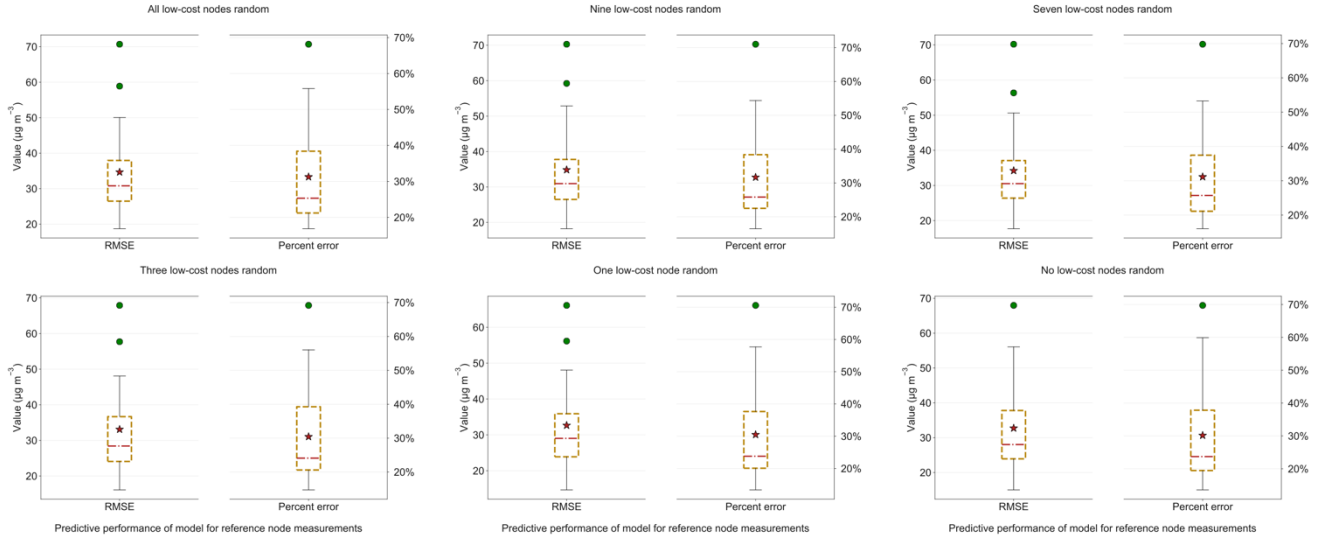

Figure S5: Gaussian Process Regression model 24 h performance scores (including RMSE and percent error) for predicting the measurements of the 22 holdout reference nodes across the 22-fold leave-one-out cross-validation using the full sensor network, when measurements of all (top left), nine (top center), seven (top right), three (bottom left), one (bottom center), and zero (bottom right) of the low-cost nodes are replaced with random integers bounded by the min and max of the true signals reported by the corresponding low-cost nodes.

Figure S6: Gaussian Process Regression model 24 h performance scores (including RMSE and percent error) for predicting the measurements of the 22 holdout reference nodes across the 22-fold leave-one-out cross-validation using the full sensor network, when measurements of two (bottom/1st row), four (2nd row), six (3rd row), eight (4th row), and all ten (top/5th row) of the low-cost nodes developed significant (11 %–99 %, left column), marginal (1 %–10 %, right column), and a balanced mixture of significant and marginal drifts. Note the sensors that drifted, the percentages of drift, and which sensors drifted significantly or marginally are randomly chosen. The results reported under each scenario are based on averages of 10 simulation runs.

Table S1: Comparison of the GPR model 24 h prediction percent errors for the 22 reference nodes across the 22-fold leave-one-out CV with and without interpolating the missing 1 h $PM_{2.5}$ values for all the reference and low-cost stations.

Reference nodes	Percent error	
	with interpolation	without interpolation
Anand Vihar	32 %	31 %
Aya Nagar	38 %	37 %
Burari Cross	39 %	38 %
CRRI Mathura Road	21 %	21 %
DTU	36 %	35 %
Faridabad	18 %	17 %
IGI Airport Terminal–3	32 %	32 %
IHBAS, Dilshad Garden	41 %	42 %
ITO	14 %	12 %
Lodhi Road	41 %	39 %
Mandir Marg	14 %	13 %
North Campus	24 %	24 %
NSIT Dawarka	19 %	20 %
Punjabi Bagh	20 %	20 %
Pusa	70 %	69 %
R K Puram	20 %	20 %
Sector125 Noida	23 %	21 %
Sector62 Noida	60 %	60 %
Shadipur	22 %	22 %
Sirifort	18 %	16 %
US Embassy	18 %	18 %
Vasundhara, Ghaziabad	44 %	34 %
Delhi-wide mean	30 %	29 %
SD	14 %	15 %

Table S2: Comparison of pre-determined percentages of drift to those estimated from the Gaussian Process Regression model for intercept and slope, respectively, for each individual low-cost node, assuming eight and four of the low-cost nodes developed various degrees of drift such as significant (11 %–99 %), marginal (1 %–10 %), and a balanced mixture of significant and marginal. Note the sensors that drifted, the percentages of drift, and which sensors drifted significantly or marginally are randomly chosen. The results reported under each scenario are based on averages of 10 simulation runs.

Drift category	Low-cost nodes	Eight low-cost nodes drift				Four low-cost nodes drift			
		Intercept drift (%)		Slope drift (%)		Intercept drift (%)		Slope drift (%)	
		True	Estimated	True	Estimated	True	Estimated	True	Estimated
Significant	AIIMS	55 %	54 %	55 %	55 %	0 %	-2 %	0 %	0 %
	Hiran Kudna	57 %	43 %	54 %	56 %	47 %	42 %	54 %	54 %
	IITD	68 %	70 %	61 %	61 %	0 %	-1 %	0 %	-1 %
	IITM	0 %	-2 %	0 %	-1 %	0 %	-2 %	0 %	-1 %
	Kaushambi	0 %	-1 %	0 %	-1 %	0 %	-1 %	0 %	-1 %
	MRU	45 %	46 %	52 %	51 %	0 %	-4 %	0 %	1 %
	Mayur Vihar	56 %	59 %	48 %	47 %	42 %	44 %	57 %	56 %
	Naraina Vihar	63 %	61 %	57 %	57 %	51 %	51 %	48 %	48 %
	New Friends Colony	53 %	53 %	57 %	57 %	70 %	71 %	39 %	38 %
	S.D.A. Park	55 %	50 %	55 %	56 %	0 %	-4 %	0 %	2 %
	Mean absolute difference	**3 %**		**1 %**		**2 %**		**1 %**	
50 % significant and 50 % marginal	AIIMS	0 %	-1 %	0 %	-1 %	0 %	-1 %	0 %	-1 %
	Hiran Kudna	47 %	40 %	58 %	58 %	0 %	-9 %	0 %	3 %
	IITD	57 %	62 %	58 %	57 %	0 %	0 %	0 %	-2 %
	IITM	6 %	5 %	6 %	3 %	4 %	3 %	7 %	6 %
	Kaushambi	4 %	4 %	5 %	1 %	0 %	0 %	0 %	-2 %
	MRU	47 %	54 %	55 %	53 %	0 %	-1 %	0 %	-1 %
	Mayur Vihar	56 %	62 %	46 %	43 %	44 %	48 %	70 %	68 %
	Naraina Vihar	5 %	3 %	4 %	3 %	58 %	56 %	46 %	47 %
	New Friends Colony	6 %	7 %	6 %	2 %	5 %	6 %	6 %	3 %
	S.D.A. Park	0 %	-3 %	0 %	1 %	0 %	-3 %	0 %	2 %
	Mean absolute difference	**3 %**		**2 %**		**2 %**		**2 %**	
Marginal	AIIMS	5 %	6 %	4 %	3 %	0 %	0 %	0 %	-1 %
	Hiran Kudna	6 %	6 %	7 %	6 %	0 %	0 %	0 %	0 %
	IITD	6 %	7 %	6 %	4 %	0 %	1 %	0 %	-1 %
	IITM	5 %	5 %	5 %	4 %	0 %	0 %	0 %	-1 %
	Kaushambi	5 %	5 %	5 %	4 %	5 %	6 %	7 %	6 %
	MRU	7 %	9 %	4 %	2 %	7 %	8 %	5 %	4 %
	Mayur Vihar	0 %	1 %	0 %	-1 %	6 %	7 %	4 %	3 %
	Naraina Vihar	6 %	7 %	6 %	5 %	0 %	0 %	0 %	-1 %
	New Friends Colony	0 %	1 %	0 %	-2 %	0 %	1 %	0 %	-1 %
	S.D.A. Park	5 %	6 %	4 %	3 %	7 %	7 %	5 %	4 %
	Mean absolute difference	**1 %**		**1 %**		**1 %**		**1 %**	