Data from Small Monitoring Networks is Unreliable

Case: Indian Cities

Sai Krishna Dammalapati & Sarath K. Guttikunda (UEinfo)
Nipun Batra & Zeel Patel (IIT, Gandhi Nagar)
SIM-air working paper series # 48-2024

UrbanEmissions (UEinfo) was founded in 2007 with the vision to be a repository of information, research, and analysis related to air pollution. UEinfo has four objectives: (1) sharing knowledge on air pollution (2) science-based air quality analysis (3) advocacy and awareness raising on air quality management and (4) building partnerships among local, national, and international airheads.

All our publications are accessible @ www.urbanemissions.info/publications

All the doodles are made using the open software @ www.excalidraw.com

Air Quality Index (AQI) data from Indian cities, utilized in this study, is available (open-access) as part of SIM-Series working paper #47-2024

Send your questions and comments to simair@urbanemissions.info

Key Messages

An ambient monitoring network in a city requires a minimum of 4-5 stations to truly represent the spatial and temporal trends of emission intensities in an urban airshed. These locations must include representation from residential, commercial, industrial, traffic, and background activities.

Operating less than the minimum number of ambient air monitoring stations will misrepresent the ground realities. Larger sample size is also necessary to capture the heterogeneity in the landuse activities and source mixes across an urban airshed.

Comparing a city represented by only one monitoring station with a city represented by at least 5 monitoring stations will lead to biased interpretations.

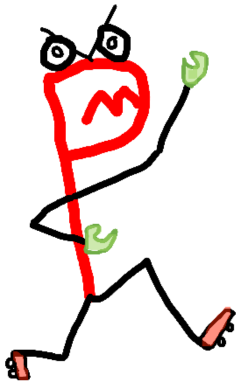
With more (and at least minimum number of) monitors, the confidence intervals are narrower, helping in definite attribution of air quality, air quality index value, and air quality index category for a city.

With more monitors, sensitivity to the type of statistical inference reduces.

1. Problem Statement

Operating less than the minimum number of ambient air monitoring stations will misrepresent the ground realities.

Is it right to compare air quality data in a city with one only monitor with as a city with 40 representative monitors?



Global rankings for most polluting cities in 2023 listed 9 cities from India in the top 10, 21 in the top 25, and 83 in the top 100[1]. Delhi remains the most polluted capital city in the world with an annual average of 102.1 $\mu g/m^3$ for $PM_{2.5}$ - this is a 10% increase from the 2022 average of 92.6 $\mu g/m^3$ and a 21% increase from the 2020 average of 84.1 $\mu g/m^3$. The 2020 average includes a drop in the annual average concentrations from multiple COVID19 lockdowns, which observed some of the strictest regulations cutting down passenger and freight traffic from the roads and shutting down several commercial and industrial activities.

Overall, India is ranked third in 2023, behind Bangladesh and Pakistan, with an annual average of 54.4 $\mu g/m^3$, 11-times more than the World Health Organization guideline of 5 $\mu g/m^3$ for $PM_{2.5}$. Chronic exposure of 1.4 billion people to these $PM_{2.5}$ concentrations in India results in an estimated 1 million premature deaths[2].

The most polluted city in the world in 2023 is Begusarai, a rural district in the state of Bihar (India), located 120 km east of Patna, the state capital. This is evidence that the pollution trends are equally worse in the rural areas, across the Indo-Gangetic plain (IGP) from Punjab in the west to West-Bengal in the east. The urban-rural nexus can be explained only by expanding the monitoring network beyond the urban centres and to discuss air quality in the areas other than big cities Delhi, Mumbai, Chennai, Kolkata, Pune, Hyderabad, and Bengaluru.

IGP experiences the worst levels of air pollution starting from post-monsoon in October and through the winter months due to an increasing demand for space heating which is supported by in situ combustion of coal, biomass, crop residue, and waste[3]. The second most polluted city is Guwahati in the Northeast, followed by Delhi. In general, the Northeastern states host more clean-air-n-blue-sky days

---

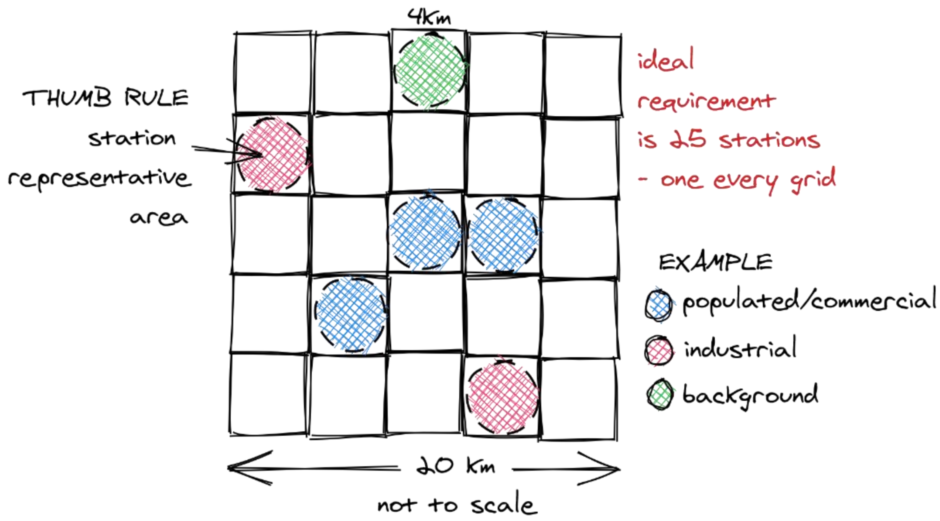
[1] https://iqair.com

[2] State of the Global Air (SoGA) portal summarizes the health impacts due to outdoor $PM_{2.5}$ and ozone and household air pollution @ https://www.stateofglobalair.org/resources/report/state-global-air-report-2024

[3] Summary of reanalysed annual and monthly $PM_{2.5}$ concentrations using a combination of emission inventories, global chemical transport model results, satellite observations, and ground measurements, for the period covering 1998 to 2022 is available @ https://urbanemissions.info. Data is available as gridded files covering the Indian Subcontinent at 0.1° resolution, state level averages, and district level averages.

than the rest of the country. However, a steady increase in the demand for urban amenities is shifting this trend in their cities.

The world rankings report was received with scepticism, because the cities with only one monitoring station and multiple monitoring stations were treated in the same order, irrespective of the representativeness of the stations. For example, there is only one monitoring station operational in Begusarai versus 40 stations in Delhi.



At city scale, we need a minimum number of monitoring stations to spatially and temporally represent the various landuse types, commercial activities, industrial facilities, traffic density, and population layout. At the least, we require five (5) monitoring stations, one each at a traffic junction, industrial site, residential site, commercial junction, and a background site, to represent the mix of activities

In this working paper, we are demonstrating methods to evaluate uncertainty associated with operating small monitoring networks to represent heterogeneity in the emission sources and landuse types in an urban airshed.

2. Data Source and Gaps

Statistical and uncertainty analysis presented in this working paper is based on air quality index (AQI) data extracted from the official daily AQI bulletins issued by the Central Pollution Control Board (CPCB), New Delhi, India, between 2015 and 2023[4].

Air Quality Index (AQI) is an important tool for communicating the quality of air pollution as health-related alerts. AQI unifies all this complicated science of pollution composition, exposure rates-based health severity, ambient standards, measurements, and standard protocols, into simple colour coded bins for everyone to see how good or bad or severe the pollution levels are[5].

AQI calculations is often based on the ambient monitoring data for 6 pollutants – particulates (as $PM_{2.5}$ and PM_{10} size fractions), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and ozone.

Key messages from India's AQI bulletins

Between 2015 and 2023 (a) the number of unique cities increased 12-fold from 22 to 271 (b) the average number of reporting stations increased 15-fold from 31 to 469 (c) and the average number of stations per unique city increased from 1.4 to 1.7– an overall 20% increase.

	Number of unique cities listed	Number of reporting stations (avg.)	Number of reporting stations (max.)	Number of stations per unique city
2015	22	31	37	1.4
2016	33	53	54	1.6
2017	54	80	90	1.5
2018	75	129	137	1.7
2019	115	188	206	1.6
2020	135	238	258	1.8
2021	170	300	326	1.8
2022	209	338	396	1.6
2023	271	469	514	1.7
Number of stations recommended			4094	
Minimum number of stations per city recommended				5.0

[4] A cleaned database of AQI data from all Indian cities, some statistical analysis, and visualizations were released as SIM-air Working Paper Series # 47-2024 @ https://urbanemissions.info and a library of python scripts used to tabulate the data from PDF bulletins is available @ www.github.com/urbanemissions

[5] An example AQI calculator comparing approved methodologies from six countries and two instructional videos is available @ https://urbanemissions.info/tools

While the number of cities and overall monitoring capacity increased between 2015 and 2023, 80% (215 out of 271) of the cities had only one monitoring station and 92% (249 out of 271) had three or less monitoring stations.

Number of cities with… # stations →	1	2	3	4	5-10	10-20	20+
in 2015	17	2	1	0	2	0	0
in 2016	28	1	2	1	1	0	0
in 2017	47	1	2	2	1	1	0
in 2018	66	3	2	1	2	0	1
in 2019	99	2	5	4	4	0	1
in 2020	111	9	7	2	4	1	1
in 2021	139	9	8	4	8	1	1
in 2022	170	14	9	6	7	2	1
in 2023	215	18	16	7	11	2	2

In 2023, only metropolitan and some Tier-1 cities, reported data from more than five (5) monitoring stations – which is a representative sample size for any city.

These 15 cities are – Agra (6), Ahmedabad (9), Bengaluru (13), Chennai (8), Delhi (39), Hyderabad (14), Jaipur (6), Jodhpur (5), Kolkata (7), Lucknow (6), Moradabad (6), Mumbai (28), Navi Mumbai (7), Patna (6), and Pune (8).

CPCB guidelines suggests a minimum of four (4)

CPCB approved the following guidelines[6] to calculate the minimum number of monitoring stations required to operate in an airshed, based on airshed's population and commercial density. The guideline for particulate pollution monitoring start with a minimum of four (4) stations for any airshed. Similar guidelines exist for gaseous pollutants – SO_2, NO_2, CO and Ozone.

Based on total population (TP) for PM monitoring.

For TP under 100,000 -- 4 units
For TP under 1 million - 4 + 0.6 per 100,000
For TP under 5 million – 7.5 + 0.25 per 100,000
For TP above 5 million - 12 + 0.16 per 100,000

[6] "Guidelines for ambient air quality monitoring", by the Central Pollution Control Board (CPCB), New Delhi, India, April-2003. Full document is available @ https://urbanemissions.info (under resources)

3. Margin of Error in Small Samples

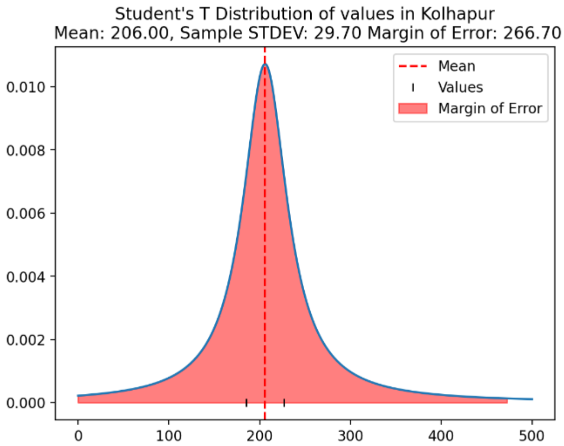
Sampling Bias: When there are fewer monitoring stations, they may not be a representative sample of the entire city, and the placement of these monitoring stations may not be "random" (in a statistical sense). Especially when the city is operating only one station, it is often located at the premises of state or regional pollution control board. Hence, any inference made on the air quality of the entire city based on this unrandom sample will be biased.

Wide Confidence-Intervals: Even if the assumption of "randomness" in the placement of air quality monitors is considered, there is an issue of wide confidence intervals. CI of the mean air quality built using the student's t-distribution function will be wide for small sample sizes. For instance, if a city only has 2 monitors, the margin of error would be 12.7 times the standard error of the mean (SEM for a 95% CI). More monitoring stations would be needed to address this issue.

We examined the margin of errors for four case studies (a) Kolhapur with 2 data points (b) Jabalpur with 4 data points (c) Hyderabad with 10 data points and (d) Delhi with 37 data points.

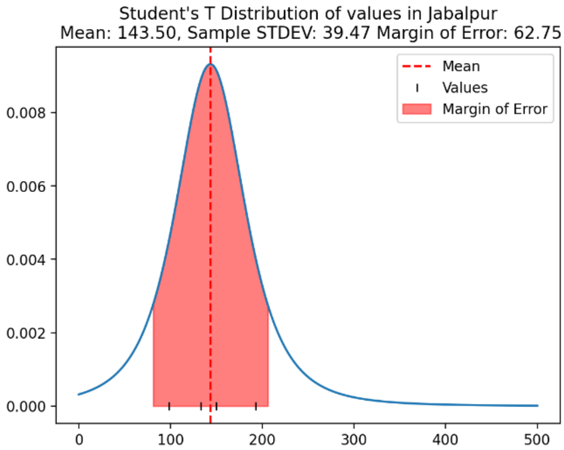
Example 1:

Kolhapur reported AQI from two monitoring stations (N=2). As April 1st, 2024: 185 and 227.
The mean AQI is 206 and the standard deviation (s) is 29.7.
SEM is 21 (s/√N).
For N=2 (dof = 1), the margin of error is 12.7 times SEM for a 95% CI, which is 267. So, the true AQI value of Kolhapur would be anywhere between 0 to 473 - a very large band.

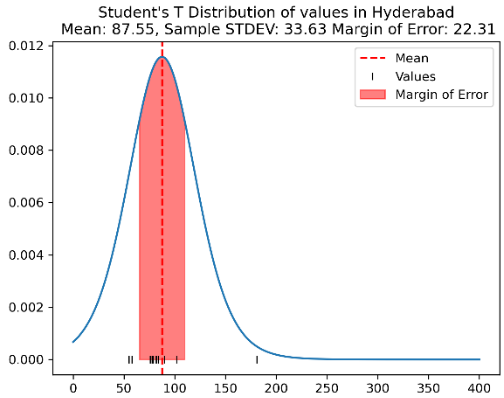


Example 2:

Jabalpur reported AQI from four monitoring stations (N=4). As of April 1st, 2024: 98, 133, 150, and 193.
The mean AQI is 143 and the standard deviation (s) is 39.5.
SEM is 19.7 (s/√N).
For N=4 (dof = 3), the margin of error is 3.2 times SEM for a 95% CI, which is 63. So, the true AQI value of Jabalpur would be anywhere between 80 to 206 - a medium size band.

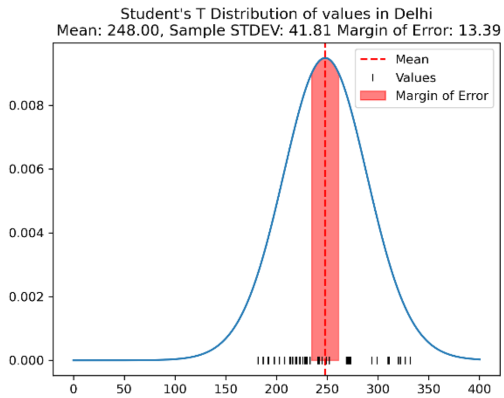


Example 3:



Hyderabad reported AQI from 10 monitoring stations (N=10). As of April 1st, 2024: 90, 78, 181, 79, 78, 76, 55, 82, 84, 58, and 102.
The mean AQI is 88 and the standard deviation (s) is 33.6.
SEM is 10.6 (s/√N).
For N=10 (dof = 9), the margin of error is 2.3 times SEM for a 95% CI, which is 25. So, the true AQI value of Hyderabad would be anywhere between 53 to 113 - a medium size band.

Example 4:



Delhi reported AQI from 37 monitoring stations (N=37) on March 31st, 2024: 182, 322, 252, 269, 245, 214, 230, 223, 229, 327, 219, 216, 272, 208, 320, 187, 230, 192, 332, 213, 198, 269, 226, 242, 299, 241, 249, 203, 270, 273, 228, 294, 233, 220, 208, 245, and 271.
The mean AQI is 245 and the standard deviation (s) is 40.
SEM is 6.6 (s/√N).
For N=37 (dof = 36), the margin of error is 2.0 times the SEM for a 95% CI, which is 13. So, the true AQI value of Delhi would be between 232 to 258 - a narrower band.

Reference:
SEM: Standard error of the mean
See the annexure of information on how to calculate margin of error

Python codes to make the plots presented in this section (and the following section) are included in the Annexure. Codes are also accessible @ https://github.com/sustainability-lab/SparseSensorsStudy

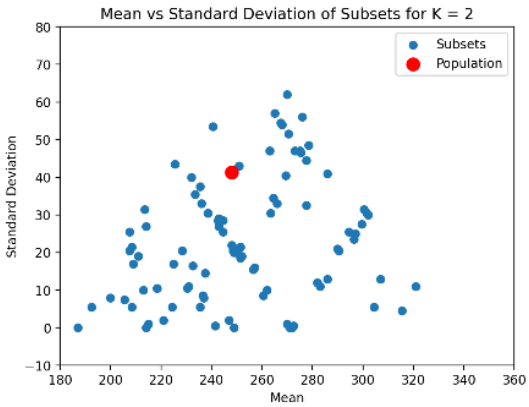
4. Heterogeneity in Monitoring Data

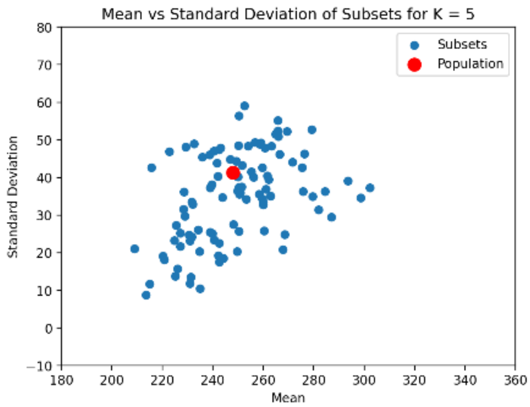
Why we need data from all representative stations?

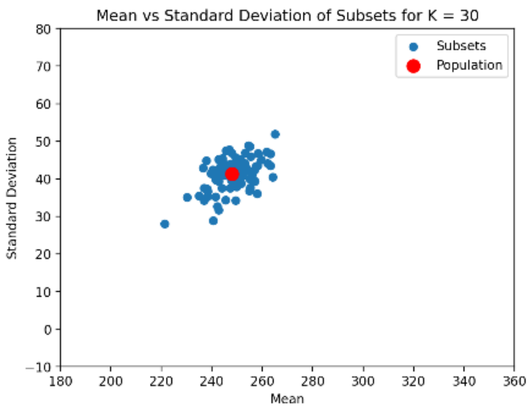
Indian urban airsheds are diverse with a mix of landuse types representing overlapping features of residential, commercial, industrial, and transport activities. This means that the heterogeneity in the airshed is very strong, and no two stations represent the same mix of emission sources.

As an experiment, for the same 37 data points from Delhi, as we randomly choose different sample sizes (2, 5, and 30), we get different means and variances, and as the sample size increases (to N=30), they close the gap to the mean and variance of the population (N=37). Even at N=30, variance in the samples is significant.

This example demonstrates the likely variation in the interpretations when data from some of the stations is not available, which is often the case at Indian monitoring stations with 80% or less data availability from the continuous ambient air quality monitoring stations.







The three panels are presenting variance in means and standard deviations for sample sizes (N=2, 5, and 30) with data selected randomly from a population of 37 AQI data samples from Delhi. The red dot indicates the population average (N=37)

To truly spatially represent the emission and landuse mix, Delhi's airshed needs to operate at least 101 continuous monitoring stations[7].

[7] Airshed level estimates for minimum number of monitoring stations for Indian cities is tabulated @ https://urbanemissions.info/india-air-quality/india-ncap-cities. A summary of required minimum number of stations in India's non-attainment cities under NCAP is included in the annexure.

5. Statistical Inference of Averages

True air quality value of a city using ambient monitoring techniques can only be determined by installing a station every 9 sq.km, assuming that a regulatory monitoring equipment can represent the activities up to 2 km radius from this location. For many reasons, often financial and personnel, we do not operate monitors at this density.

What is the statistical inference of air quality in a city with limited (small sample) monitoring?

Do we get the same conclusion using various statistical methods with limited (small sample) monitoring?

There are two ways to perform this statistical inference: non-parametric and parametric (see the Annexure for methods).

- Non-parametric methods like bootstrap estimation are performed when we are unaware of the underlying population distribution.
- Parametric estimations are performed when we have knowledge of underlying population distributions from prior research. Prior research indicates that the pollution concentration data and AQI data is in a log-normal distribution.

We evaluated these methods for 4 cities – Kolhapur, Jabalpur, Hyderabad and Delhi. Key messages from this exercise

- With more monitors, the confidence intervals are narrower, helping in definite attribution of AQI category for the city.
- With more monitors, sensitivity to the type of statistical inference reduces.

Example 1: Kolhapur

Kolhapur in Maharashtra has only two air quality monitoring stations (N=2). As of April 1st, 2024, the AQI values reported by these two stations are: 185 and 227. The official AQI bulletin reported an average AQI of 206 and attributed “Poor” AQI category accordingly.

A non-parametric bootstrap statistical inference on such a small sample would estimate that the true AQI mean value would lie between 185 and 227. A parametric inference of Kolhapur’s true AQI value can be performed considering that AQI data is a log-normal distribution. Since the sample size is small, Student-t distribution was used. This is because we consider that the sampling distribution of log-means would converge to log-normal distribution at higher sample sizes. But at smaller sample sizes, it would converge to log Student-t distribution. Inference with this assumption would give an extremely wide 95% confidence interval for the true AQI of Kolhapur – (55, 751).

Table of statistical inference done for Kolhapur with various assumptions.

Statistical Inference	95% percentile confidence interval of mean	AQI categories
Non-parametric bootstrap	185-227	Moderate-Poor
Parametric: log Student-t distribution	55-751	Satisfactory-Moderate-Poor-Very Poor-Severe
Parametric: log Normal distribution	167-250	Moderate-Poor
Parametric: Normal distribution	164-247	Moderate-Poor
Parametric: Student-t distribution	0-472	Good-Satisfactory-Moderate-Poor-Very Poor-Severe

Example 2: Delhi

Delhi reported AQI from 36 stations (N=36) on April 1st, 2024. The AQI values reported are: 105, 144, 148, 150, 118, 179, 120, 156, 147, 87, 133, 83, 158, 109, 288, 94, 104, 118, 195, 170, 97, 123, 116, 119, 120, 130, 139, 136, 120, 118, 108, 199, 112, 106, 111, 131. The official AQI bulletin reported an average AQI as 133 and attributed “Moderate” AQI category accordingly.

A non-parametric bootstrap statistical inference on this sample estimated that the true AQI mean value would lie between (121, 146) interval with 95% confidence. A parametric inference considering log-normal distribution estimated that the true AQI value of Delhi would be between (118, 139) interval with 95% confidence. This is a narrower band compared to that of Kolhapur (with N=2). It also helps in placing Delhi’s AQI category deterministically in the “Moderate” category.

Table of statistical inference done for Delhi with various assumptions.

Statistical Inference	95% percentile confidence interval of mean	AQI categories
Non-parametric bootstrap	121-146	Moderate
Parametric: log Student-t distribution	118-140	Moderate
Parametric: log Normal distribution	118-139	Moderate
Parametric: Normal distribution	120-145	Moderate
Parametric: Student-t distribution	120-146	Moderate

Example 3: Jabalpur

Jabalpur reported AQI data from four air quality monitoring stations (N=4) on April 1st, 2024. The AQI values reported are: 98, 150, 133, 193. The official AQI bulletin reported the average AQI as 144 and attributed "Moderate" AQI category accordingly

Table of statistical inference done for Jabalpur with various assumptions.

Statistical Inference	95% percentile confidence interval of mean	AQI categories
Non-parametric bootstrap	111-178	Moderate
Parametric: log Student-t distribution	89-218	Satisfactory-Moderate-Poor
Parametric: log Normal distribution	105-183	Moderate
Parametric: Normal distribution	104-182	Moderate
Parametric: Student-t distribution	80-206	Satisfactory-Moderate-Poor

Example 4: Hyderabad

Hyderabad reported data from 11 monitoring stations (N=11) on April 1st, 2024. The AQI values reported are: 90, 78, 181, 79, 78, 76, 55, 82, 84, 58, 102. The official AQI bulletin reported an average AQI of 88 and attributed "Satisfactory" AQI category accordingly.

Table of statistical inference done for Hyderabad with various assumptions.

Statistical Inference	95% percentile confidence interval of mean	AQI categories
Non-parametric bootstrap	72-108	Satisfactory-Moderate
Parametric: log Student-t distribution	67-102	Satisfactory-Moderate
Parametric: log Normal distribution	69-100	Satisfactory
Parametric: Normal distribution	67-107	Satisfactory-Moderate
Parametric: Student-t distribution	64-110	Satisfactory-Moderate

6. Annexure: Methods

Non-parametric bootstrap

The non-parametric bootstrap method is a resampling technique used to estimate the distribution of a statistic by repeatedly sampling with replacement from the observed data. This method is particularly useful for making statistical inferences when the underlying distribution is unknown.

Let X = {x_1, x_2,…,x_n} be the original sample consisting of n observations. Then we resample with replacement from this original sample several times, say 10,000 times. Thus, we obtain 10,000 resampled samples and thus 10,000 means or any other statistic of interest θ. The collection of these statistics (θ) is then used to infer the true statistic. 95% Confidence Interval is estimated by building the interval from 2.5 percentile to 97.5 percentile of the collection of these statistics.

Parametric inference using Normal Distribution and Students' t-Distribution

Parametric inference involves making statistical inferences about population parameters based on assumptions about the underlying distribution of the data. When the data is assumed to follow a normal distribution, parametric inference is performed by first estimating the parameters of the normal distribution (mean, standard deviation) using the sample data.

Let X = {x_1, x_2,…,, x_n} be the original sample consisting of n observations. Then, the maximum likelihood estimates of the mean ($\underline{x}$) and standard deviation (s) are:

$$\underline{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{1=n}(x_i - \underline{x})^2}$$

Once the estimates are calculated, then confidence interval of the mean can be calculated by

$$\underline{x} \pm z\frac{s}{\sqrt{n}}$$

where z is the confidence level value.

When the sample size is small, the sampling distribution of means doesn't converge to a normal distribution and thus a Students' t-distribution is used. The confidence interval of the mean can then be calculated by

$$\underline{x} \pm t\frac{s}{\sqrt{n}}$$

where t is the critical value of t-distribution at desired confidence level.

Parametric inference using log-Normal Distribution and log-Normal Students' t-Distribution

The log-normal distribution also has parameters like a normal distribution – mean, standard deviation. However, these are calculated after log transformation of the original sample data.

Let X = {x_1, x_2,..., x_n} be the original sample consisting of n observations. Then this sample data is transformed by applying natural logarithm. Y = {ln(x_1), ln(x_2), ..., ln(x_n)}. Then, the maximum likelihood estimates of the mean ($\underline{y}$) and standard deviation (s) are:

$$\underline{y} = \frac{1}{n}\sum_{i=1}^{i=n} \ln{(x_i)}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{1=n}\left(ln(x_i) - \underline{y}\right)^2}$$

Once the estimates are calculated, then confidence interval of the log-mean can be calculated by

$$\underline{y} \pm z\frac{s}{\sqrt{n}}$$

where z is the confidence level value.

The confidence interval of the mean can be calculated by applying exponential transformation to the lower and upper bounds.

$$(e^{\underline{y} - z\frac{s}{\sqrt{n}}}, e^{\underline{y} + z\frac{s}{\sqrt{n}}})$$

When the sample size is small, the sampling distribution of log-means doesn't converge to a normal distribution and thus a Students' t-distribution is used. The confidence interval of the log-mean can then be calculated by

$$\underline{y} \pm t\frac{s}{\sqrt{n}}$$

where t is the critical value of t-distribution at desired confidence level.

The confidence interval of the mean can be calculated by applying exponential transformation to the lower and upper bounds.

$$(e^{\underline{y} - t\frac{s}{\sqrt{n}}}, e^{\underline{y} + t\frac{s}{\sqrt{n}}})$$

Margin of Errors

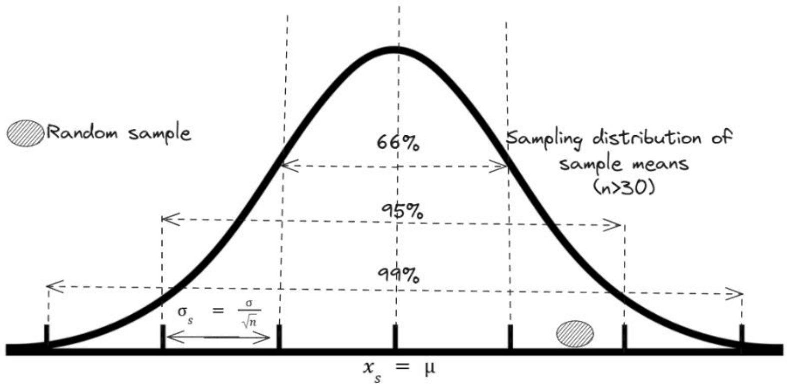
Given a small sample size, a Student's t-distribution would be used for the purposes of statistical inference.

When the sample sizes are large (generally >30), then according to the Central Limit Theorem (CLT), the sampling distribution of sample means would be normally distributed.

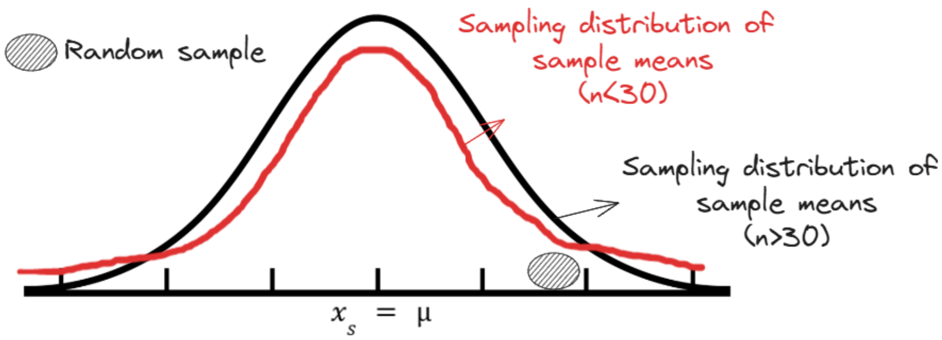
This 'normal' sampling distribution would have a mean equal to the true mean of the population and a standard deviation (standard error of mean - SEM) equal to the standard deviation of the population divided by the square root of the sample size.

In such a scenario, the mean of any random sample would be within 2 standard deviations (1.96 to be precise) away from the mean of the sampling distribution 95% of times. The margin of error would then be 2 times the standard error of the mean at 95% confidence.

$$x_s = \mu; \quad \sigma_s = \frac{\sigma}{\sqrt{n}}$$



But when sample sizes are smaller, the sampling distribution of sample means would not be normal in distribution. There would be fatter tails in the distribution.



In such a scenario, the mean of a random sample would be further away from the mean of the sampling distribution. This standard error of the mean can be computed from a Student’s t-Table.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
					Confidence Level						

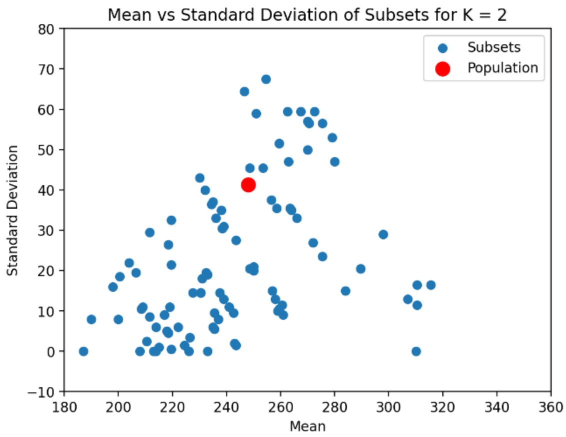
For instance, if the sample size is 2 (degrees of freedom = 1) then the margin of error would be 12.71 times the standard error of the mean at 95% Confidence. If the sample size increases to 4 (dof = 3), the margin of error would reduce to 3.18 times the standard error of the mean.

7. Annexure: Python Codes

Codes are also accessible @ https://github.com/sustainability-lab/SparseSensorsStudy

All the codes are authored by Nipun Batra and Zeel Patel

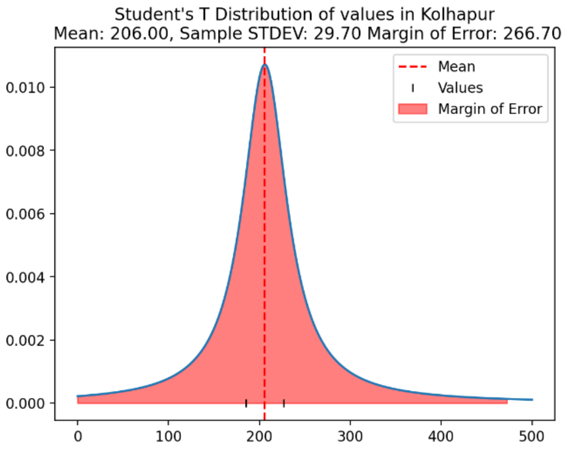
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
%config InlineBackend.figure_format = 'retina'
import torch
import torch.nn as nn
import torch.nn.functional as F
from einops import rearrange, reduce, repeat
from scipy import stats

# enter data values
vals_Delhi = np.array([182, 322, 252, 269, 245, 214, 230, 223,
229, 327, 219, 216, 272, 208, 320, 310, 187, 230, 192, 332, 213,
198, 269, 226, 242, 299, 241, 249, 203, 270, 273, 228, 294,
233, 220, 311, 208, 245, 271])
#vals_Kolhapur = np.array([90, 150])
vals_Kolhapur = np.array([185,227])
vals_Hyderabad = np.array([90, 78, 181, 79, 78, 76, 55, 82,
84, 58, 102])
vals_Jabalpur = np.array([98, 150, 133, 193])
```



 # for heterogeneity plots

```
pop_mean = np.mean(vals_Delhi)
pop_stdev = np.std(vals_Delhi)

# Now, consider different subsets of the data of size K and find the mean and standard deviation of
each subset
# and plot the mean and standard deviation of each subset.

K = 5

def plot_subsets(vals, K):
  means = []
  stdevs = []
  num_subset = 100
  for i in range(num_subset):
    subset = np.random.choice(vals_Delhi, K)
    means.append(np.mean(subset))
    stdevs.append(np.std(subset))
  plt.scatter(means, stdevs, label = 'Subsets')
  plt.xlabel('Mean')
  plt.ylabel('Standard Deviation')
  plt.scatter([pop_mean], [pop_stdev], color='red', label = 'Population', s = 100)
  plt.xlim(180, 360)
  plt.ylim(-10, 80)
  plt.legend()
  plt.title(f'Mean vs Standard Deviation of Subsets for K = {K}')

# for K =2
plot_subsets(vals_Delhi, 2)
# for K =5
plot_subsets(vals_Delhi, 5)
```

```
# for K =30
plot_subsets(vals_Delhi, 30)
```



 # for margin of error plots

```
def plot_distribution (vals, city='Delhi'):
    # Fit a normal distribution to the data
    # mu, std = np.mean(vals), np.std(vals, ddof=0)

    # Fit a student distribution to the data
    mu = np.mean(vals)
    print(f"{mu=:.2f}")
    std = np.std(vals, ddof=1)
    print(f"{std=:.2f}")
    # Plot the normal distribution
    xs = np.linspace(0, 500, 1000)
    ys = stats.t(loc=mu, scale=std, df=len(vals) - 1).pdf(xs)
    plt.plot(xs, ys)
    # Mark the mean
    plt.axvline(mu, color='r', linestyle='--', label = 'Mean')
    # Mark the values via rag plot
    plt.plot(vals, [0]*len(vals), 'k|', label = 'Values')

    standard_error_of_mean = std / np.sqrt(len(vals))
    print(f"{standard_error_of_mean=:.2f}")
    if city == 'Delhi':
        margin_of_error = standard_error_of_mean * 2.0
    elif city == 'Kolhapur':
        margin_of_error = standard_error_of_mean * 12.7
    elif city == 'Hyderabad':
        margin_of_error = standard_error_of_mean * 2.26
    elif city == 'Jabalpur':
        margin_of_error = standard_error_of_mean * 3.18
    else:
        raise ValueError("City not listed in the code")

    print(f"{margin_of_error=:.2f}")
    plt.fill_between(xs, 0, ys, where = (xs > mu - margin_of_error) & (xs < mu + margin_of_error), color =
'r', alpha = 0.5, label = 'Margin of Error')
    plt.legend()
    plt.title(f"Student's T Distribution of values in {city}\n Mean: {mu:.2f}, Sample STDEV: {std:.2f} Margin
of Error: {margin_of_error:.2f}")
    plt.savefig(f'MOE_{city}.png', dpi=300)

#example command
plot_distribution (vals_Delhi, city='Delhi')
plot_distribution (vals_Kolhapur, city='Kolhapur')
plot_distribution (vals_Jabalpur, city='Jabalpur')
plot_distribution (vals_Hyderabad, city='Hyderabad')
```

8. Recommended Minimum No. of Monitoring Locations for Indian Airsheds Under NCAP

Based on the guidelines issued by the Central Pollution Control Board for ambient monitoring in 2003, the following minimums were calculated.

Full publication on the methods are published here
Plugging the ambient air monitoring gaps in India's national clean air programme (NCAP) airsheds (Atmospheric Environment, 2023)
https://www.sciencedirect.com/science/article/pii/S1352231023001383

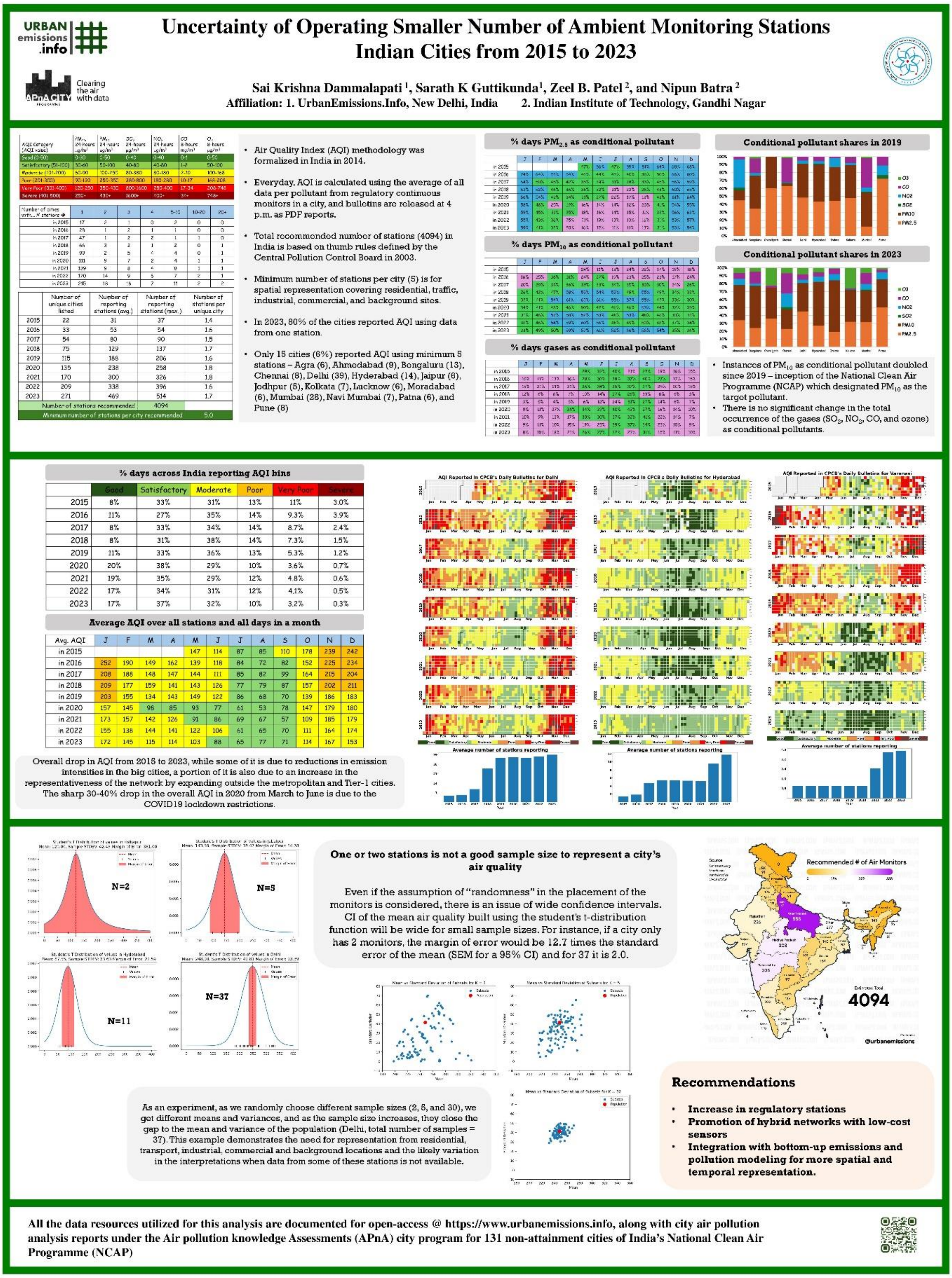
And the associated databases are published here
https://urbanemissions.info

Table: Characteristics of airsheds designated for NCAP non-attainment cities. B = cities included in the airshed from the NCAP list; C = cities included in the airshed, but not on the NCAP list; D = airshed size in grids of equal size (0.01°); E = total airshed population (in million); F = fraction of grids designated as urban using built-up area; G = fraction of population in the urban grids; H, I, J, K = number of continuous monitoring stations recommended for tracking PM, SO_2, NO_2, and Others respectively.

	State/UT	Airshed	B	C	D	E	F	G	H	I	J	K
1	Andhra Pradesh	Anantapur			30 x 30	0.6	8%	60%	10	6	8	2
2	Andhra Pradesh	Chitoor			30 x 30	0.5	8%	50%	9	5	7	2
3	Andhra Pradesh	Eluru		Hanuman Junction	30 x 30	0.7	8%	50%	10	6	8	2
4	Andhra Pradesh	Kadapa			30 x 30	0.5	6%	62%	9	6	8	2
5	Andhra Pradesh	Kurnool			30 x 30	0.7	10%	65%	10	6	9	3
6	Andhra Pradesh	Nellore			30 x 30	0.8	15%	66%	12	7	9	3
7	Andhra Pradesh	Ongole			30 x 30	0.5	9%	54%	9	5	7	2
8	Andhra Pradesh	Rajahmundry			30 x 30	1.4	25%	55%	17	9	10	4
9	Andhra Pradesh	Srikakulam			30 x 30	0.7	8%	41%	10	6	8	2
10	Andhra Pradesh	Vijayawada	Guntur	Tenali	50 x 50	3.1	23%	65%	22	11	10	6
11	Andhra Pradesh	Vishakhapatnam		Anakapalle	50 x 50	2.9	18%	68%	20	11	10	6
12	Andhra Pradesh	Vizianagaram			30 x 30	0.9	9%	47%	12	8	10	3
13	Assam	Guwahati	Byrnahati	Dispur	40 x 30	1.7	36%	73%	18	9	10	4
14	Assam	Nagaon			30 x 30	1.2	47%	20%	36	8	10	3
15	Assam	Nalbari			30 x 30	0.9	31%	56%	11	8	10	3
16	Assam	Sibsagar			30 x 30	0.5	19%	32%	12	5	7	2
17	Assam	Silchar			30 x 30	1.1	14%	18%	19	8	10	3
18	Bihar	Gaya			30 x 30	1.6	18%	30%	19	9	10	4
19	Bihar	Muzaffarpur			30 x 30	2.7	42%	30%	35	11	10	6
20	Bihar	Patna			60 x 40	7.0	38%	46%	43	17	10	10
21	Chandigarh	Chandigarh	Dera Bassi, Parwanoo	Panchkula, Kalka	50 x 40	2.9	40%	76%	23	11	10	6
22	Chhattisgarh	Korba			40 x 40	0.9	11%	58%	12	7	10	3
23	Chhattisgarh	Raipur	Bhillai	Durg	60 x 30	3.2	29%	76%	22	11	10	6

24	Delhi	Delhi	Faridabad, Ghaziabad, Noida	Greater Noida, Gurugram, Palwal, Manesar, Sonipat	100 x 100	32.8	43%	79%	101	20	10	23
25	Gujarat	Ahmedabad		Gandhi Nagar	50 x 50	7.9	40%	79%	38	18	10	10
26	Gujarat	Rajkot			30 x 30	1.5	24%	80%	16	9	10	4
27	Gujarat	Surat		Hazira	50 x 50	5.8	23%	61%	30	15	10	9
28	Gujarat	Vadodara			30 x 30	2.6	34%	82%	21	10	10	5
29	Himachal Pradesh	Kala Amb			30 x 30	0.4	7%	29%	9	5	7	2
30	Himachal Pradesh	Nalagarh	Baddi		30 x 30	0.3	20%	62%	9	5	7	2
31	Himachal Pradesh	Paonta Sahib			20 x 20	0.2	12%	53%	7	4	5	2
32	Himachal Pradesh	Sunder Nagar			20 x 20	0.2	22%	63%	8	4	6	2
33	Jammu & Kashmir	Jammu			30 x 30	1.3	47%	65%	19	8	10	3
34	Jammu & Kashmir	Srinagar			30 x 30	2.1	56%	77%	23	10	10	5
35	Jharkhand	Dhanbad			60 x 40	3.8	23%	39%	28	12	10	7
36	Jharkhand	Jamshedpur		Bokaro, Jaropokhar	40 x 40	2.2	12%	61%	16	10	10	5
37	Jharkhand	Ranchi			40 x 40	1.9	20%	58%	17	9	10	4
38	Karnataka	Bangalore			60 x 60	11.7	50%	81%	50	20	10	12
39	Karnataka	Devanagere			30 x 30	0.9	12%	65%	12	7	10	3
40	Karnataka	Gulburga			30 x 30	0.8	10%	71%	11	7	9	3
41	Karnataka	Hubli-Dharwad			30 x 30	1.3	18%	77%	14	8	10	3
42	Madhya Pradesh	Bhopal			40 x 40	2.6	23%	86%	19	10	10	5
43	Madhya Pradesh	Gwalior			30 x 30	1.4	17%	71%	15	9	10	4
44	Madhya Pradesh	Indore	Dewas, Ujjain	Mhow, Pitampura	80 x 80	5.5	11%	51%	26	15	10	9
45	Madhya Pradesh	Jabalpur			40 x 40	1.9	15%	75%	16	9	10	4
46	Madhya Pradesh	Sagar			30 x 30	0.5	8%	61%	9	6	8	2
47	Maharashtra	Akola			30 x 30	0.8	10%	64%	11	7	9	3
48	Maharashtra	Amravati			30 x 30	0.9	10%	74%	12	8	10	3
49	Maharashtra	Aurangabad			40 x 40	1.9	16%	73%	16	9	10	4
50	Maharashtra	Chandrapur			30 x 30	0.7	12%	73%	11	7	9	3
51	Maharashtra	Jalgaon			30 x 30	0.8	10%	66%	11	7	9	3
52	Maharashtra	Jalna			30 x 30	0.6	7%	51%	9	6	8	2
53	Maharashtra	Kolhapur	Sangli		60 x 40	3.9	23%	47%	26	12	10	7
54	Maharashtra	Latur			30 x 30	0.8	10%	60%	11	7	9	3
55	Maharashtra	Mumbai	Badlapur, Navi Mumbai, Thane, Ulhasnagar, Vasai Virar	Kalyan, Karjat	80 x 80	25.1	21%	78%	67	20	10	19
56	Maharashtra	Nagpur			40 x 40	3.6	28%	88%	23	12	10	7
57	Maharashtra	Nashik			40 x 40	2.6	29%	75%	20	10	10	5
58	Maharashtra	Pune		Pimpri-Chinchwad, Hinjewadi	40 x 40	6.8	60%	86%	40	17	10	10
59	Maharashtra	Solapur			30 x 30	1.1	16%	79%	13	8	10	3
60	Nagaland	Dimapur			30 x 30	0.5	22%	80%	10	5	7	2
61	Nagaland	Kohima			30 x 30	0.2	5%	54%	7	4	6	2
62	Orissa	Angul	Talcher		40 x 40	0.7	11%	39%	12	7	9	3
63	Orissa	Balasore			30 x 30	0.8	8%	36%	12	7	9	3
64	Orissa	Bhubaneswar	Cuttack, Kalinga Nagar		40 x 40	3.2	21%	60%	22	11	10	6
65	Orissa	Rourkela			30 x 30	1.2	16%	56%	15	8	10	3
66	Punjab	Amritsar		Tarn Taran	40 x 40	2.2	38%	69%	21	10	10	5
67	Punjab	Jalandhar		Phagwara	40 x 40	1.9	44%	65%	22	9	10	4
68	Punjab	Khanna	Gobindgarh		30 x 30	0.7	37%	69%	14	7	9	3
69	Punjab	Ludhiana		Philaur	40 x 40	2.7	45%	78%	23	11	10	6
70	Punjab	Naya Nangal		Una	30 x 30	0.5	29%	65%	11	5	7	2
71	Punjab	Pathankot/Dera Baba	Damtal		30 x 30	0.7	30%	70%	13	7	9	3
72	Punjab	Patiala			60 x 40	1.8	22%	48%	19	9	10	4
73	Rajasthan	Alwar			30 x 30	0.9	18%	67%	13	7	10	3
74	Rajasthan	Jaipur			40 x 40	4.5	54%	90%	31	13	10	8

75	Rajasthan	Jodhpur			40 x 40	1.9	26%	83%	17	9	10	4
76	Rajasthan	Kota			30 x 30	1.1	25%	83%	14	8	10	3
77	Rajasthan	Udaipur			30 x 30	1.4	27%	71%	16	9	10	4
78	Tamil Nadu	Chennai			50 x 50	10.9	44%	83%	46	20	10	12
79	Tamil Nadu	Madurai		Singrauli	30 x 30	2.1	27%	86%	18	10	10	5
80	Tamil Nadu	Thoothukudi			40 x 40	0.9	11%	66%	12	7	10	3
81	Tamil Nadu	Trichy			30 x 30	1.8	31%	78%	18	9	10	4
82	Telangana	Hyderabad	Patancheru, Sangareddy		60 x 60	9.0	36%	85%	39	20	10	11
83	Telangana	Nalgonda			30 x 30	0.4	6%	44%	8	5	7	2
84	Uttar Pradesh	Agra			40 x 40	3.7	22%	66%	23	12	10	7
85	Uttar Pradesh	Allahabad			40 x 40	3.7	31%	49%	28	12	10	7
86	Uttar Pradesh	Anpara			40 x 40	0.8	15%	65%	12	7	9	3
87	Uttar Pradesh	Bareily			30 x 30	2.4	25%	63%	20	10	10	5
88	Uttar Pradesh	Firozabad			30 x 30	1.5	11%	43%	15	9	10	4
89	Uttar Pradesh	Gajraula			30 x 30	0.8	16%	43%	13	7	9	3
90	Uttar Pradesh	Gorakhpur			30 x 30	2.3	44%	60%	24	10	10	5
91	Uttar Pradesh	Jhansi			30 x 30	0.9	17%	72%	13	8	10	3
92	Uttar Pradesh	Kanpur		Unnao	40 x 40	4.0	23%	70%	24	13	10	8
93	Uttar Pradesh	Khurja		Bulandshahr	30 x 30	1.2	14%	32%	16	8	10	3
94	Uttar Pradesh	Lucknow		Barabanki	60 x 60	6.4	22%	54%	32	16	10	10
95	Uttar Pradesh	Meerut			30 x 30	2.5	42%	73%	23	10	10	5
96	Uttar Pradesh	Moradabad			30 x 30	2.0	29%	51%	21	10	10	5
97	Uttar Pradesh	Raebareli			30 x 30	1.1	7%	27%	14	8	10	3
98	Uttar Pradesh	Varanasi			40 x 40	4.6	52%	57%	37	13	10	8
99	Uttarakhand	Dehradun			30 x 30	1.1	31%	82%	15	8	10	3
100	Uttarakhand	Kashipur			30 x 30	1.0	22%	46%	16	8	10	3
101	Uttarakhand	Rishikesh		Haridwar	30 x 30	0.8	20%	75%	12	7	9	3
102	West Bengal	Asansol	Durgapur	Ranigunj	60 x 40	3.6	26%	43%	27	12	10	7
103	West Bengal	Haldia			40 x 40	2.2	11%	7%	34	10	10	5
104	West Bengal	Kolkata	Barrackpore, Howrah		60 x 60	20.4	50%	61%	82	20	10	17



Poster presented at the IGAC conference in Kuala Lumpur, Malaysia, September 2024. Download full resolution copy @ https://urbanemissions.info www.urbanemissions.info