CEG5103/EE5024 Case Study

Machine Learning (ML)-based Air Quality Monitoring using Vehicular Sensor Networks

by
Duc Van Le and **Chen-Khong Tham**
National University of Singapore

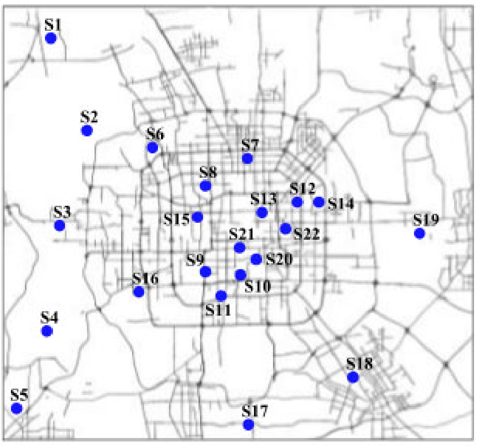
IEEE ICPADS 2017, Shenzhen, China

Dec. 17, 2017

Presentation Outline

1. Introduction to Urban Air Quality Monitoring
2. Our Proposed Algorithm
3. Conclusions

Conventional Stationary Monitoring Network



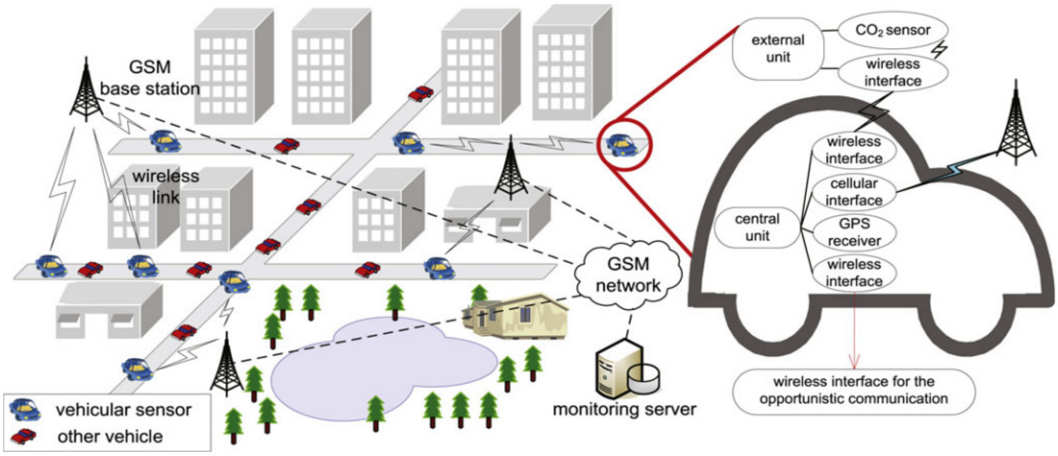
A) Configuration of a station B) Air quality measurement stations in Beijing

- Air quality is a major concern in modern cities
- To monitor the air quality in a city, important air quality parameters such as CO, $PM_{2.5}$ need to be collected

Static Monitoring Stations:

- Can accurately measure a wide range of air quality parameters
- Require a big land area, high cost (about USD 200,000 for construction and USD 30,000 per year for maintenance)
- Non-scalability for changes in urban arrangement, activities or regulation, which require relocating stations or adding new stations

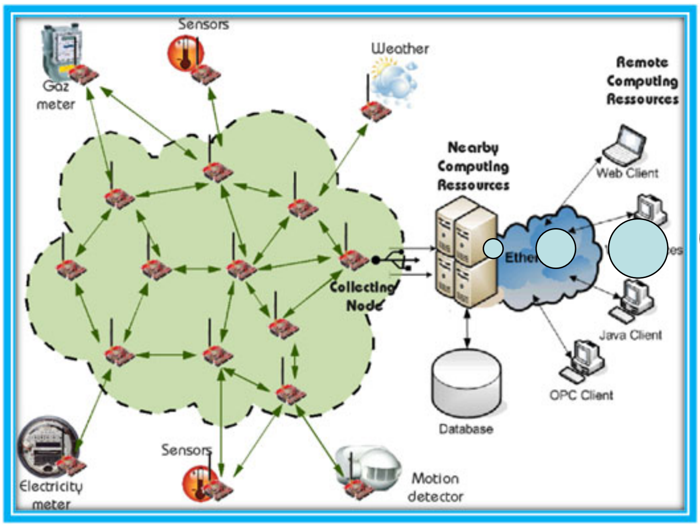
Air Quality Monitoring using Vehicular Sensor Network



Source: S. C. Hu *et al.* "A vehicular wireless sensor network for CO2 monitoring," in Proc. 2009 IEEE Sensors

- Vehicular sensor network (VSN) is an efficient solution for the urban air quality monitoring
- Formed by group of vehicles (e.g., cars, buses) which are equipped with computing units and sensing devices
- Vehicles move around the city area and measure air quality parameters

System Model



Build a global sensing map over a monitoring area based on the collected sensing data

Existing Centralized Method:

- Sensing vehicle sends all its collected data to the monitoring center
- The center builds the sensing map by utilizing the collected sensing data
- Main disadvantage is the high communication cost for transmitting the data from the vehicles to the center

Spatial Interpolation Methods

- Normally, spatial interpolation methods are used to predict the air quality at unsampled location
- In general, the value z at unsampled location x_0 are interpolated based on values of sampled ones [1]

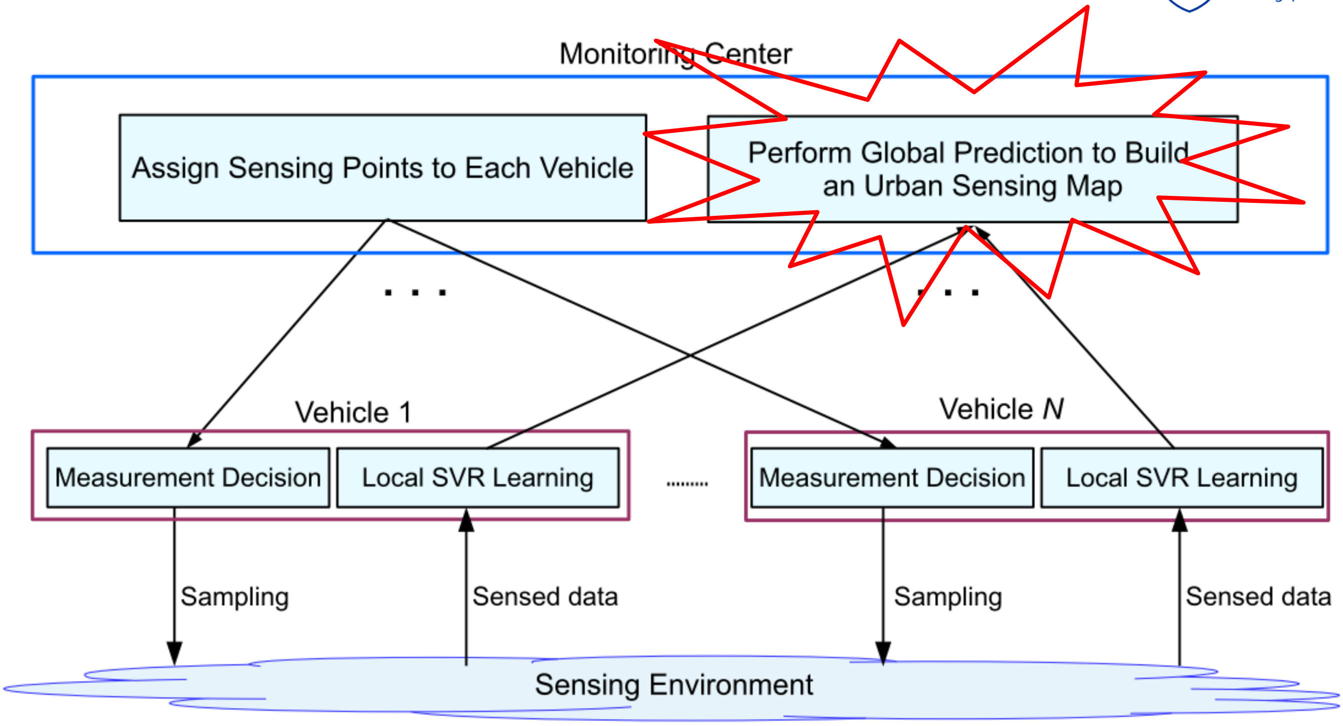
$$z(x_0) = \sum_{i=1}^{n} \omega_i z(x_i) \ and \sum_{i=1}^{n} \omega_i = 1$$

where ω_i represents the weights assigned to each of sampled locations.

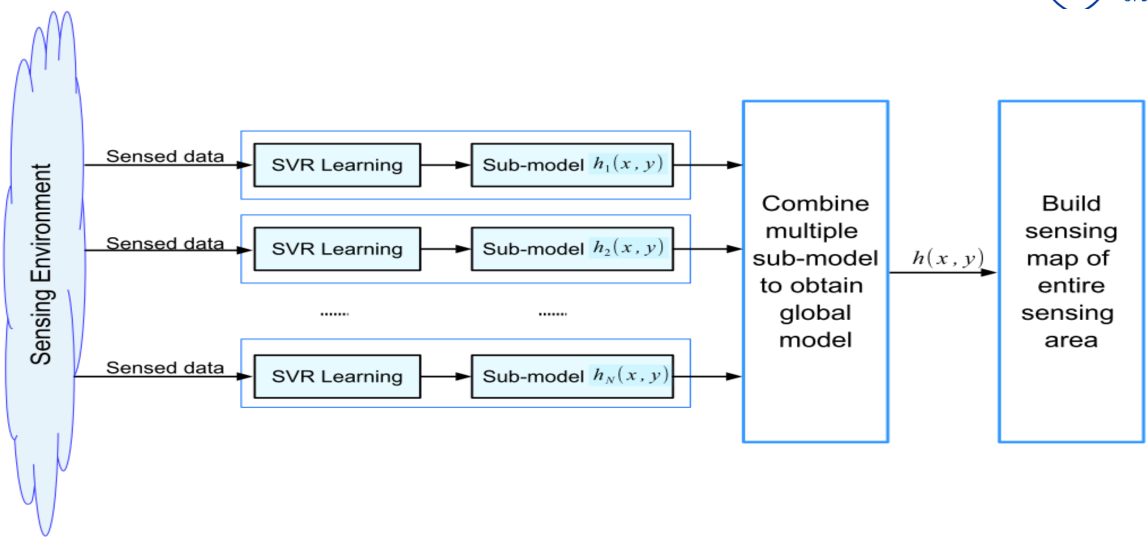
- Depending on the way of determining the weight value ω_i, three interpolation methods are widely used:
  - Nearest Neighbor (NN): take the value of the sampled point which is the nearest to x_0
  - Inverse Distance Weighting (IDW): assign a higher weight for the closer sampled point.
  - Kriging: use a variogram to compute the weight which minimizes the variance of the estimated value

[1] Wong DW, Yuan L, Perlin SA, "Comparison of spatial interpolation methods for the estimation of air quality data", J Expo Anal Environ Epidemiol, 2004;14(5):404–415.

MLAirM: Machine Learning (ML)-based Distributed Air Quality Monitoring Scheme



Building a Global Sensing Map based on Distributed Machine Learning



- Each vehicle is assigned to take measurements on a set of points then build a local sensing map using support vector regression (SVR) model
- After a sensing period, all vehicles will sends local model parameters to the center.
- Then, the center uses the received models to build the entire sensing map of the interest area.
- Communication cost is reduced since vehicles do not need to send raw sensing data to the center

Distributed Support Vector Regression (SVR)

- Each vehicle constructs its SVR sub-model using its collected data:

$$f(x) = \langle w, x \rangle + b \text{ with } w \in \mathcal{X}, b \in \mathbb{R}$$

mapping function $\varphi(x)$

$$\text{minimize} \quad \frac{1}{2}\|w\|^2$$

$$\text{subject to} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \le \varepsilon \\ \langle w, x_i \rangle + b - y_i \le \varepsilon \end{cases}$$

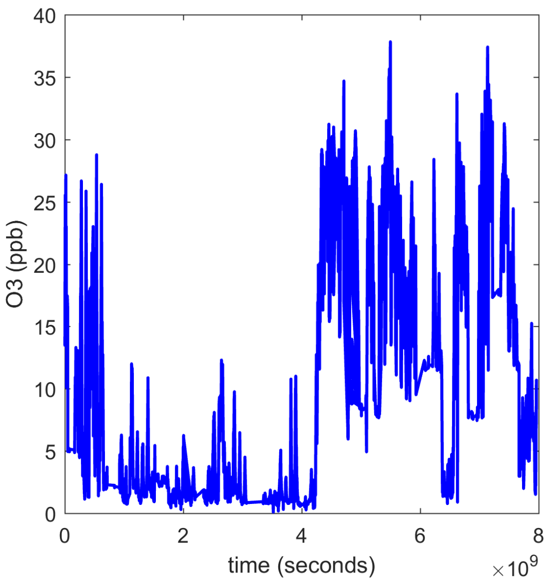
- The center uses a fuzzy synthesis[2] method to predict the measurement y at location x, based on the predicted value y_i of N vehicles' sub-models

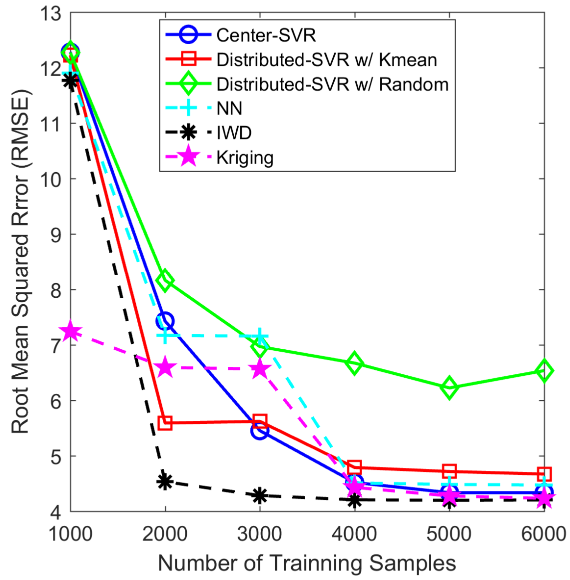
$$y = \sum_{i=1}^{N} \eta_i \, y_i \text{ where } \begin{cases} \eta_i = 1, \eta_{j \neq i} = 0, & \text{if } d_i = 0 \\ \eta_i = \dfrac{\frac{1}{d_i}}{\sum_{i=1}^{N} \frac{1}{d_i}}, & otherwise \end{cases}$$

where $d_i = \|x - c_i\|$, c_i is the center of i[th] subset of samples of sub-model i

[2] J. Cheng, J. Qian, Y. Guo, "A distributed support vector machines architecture for chaotic time series prediction", Proc. Neural Inform. Process, 2006.

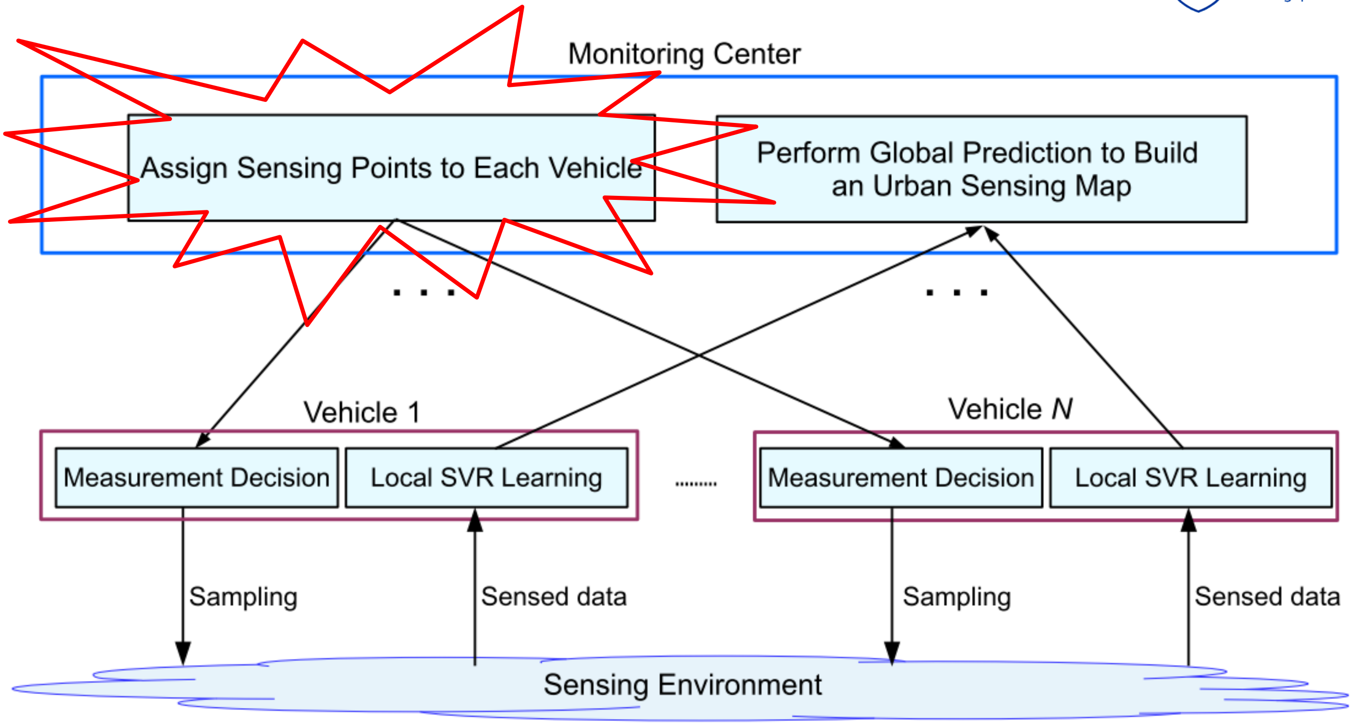
Motivating Example



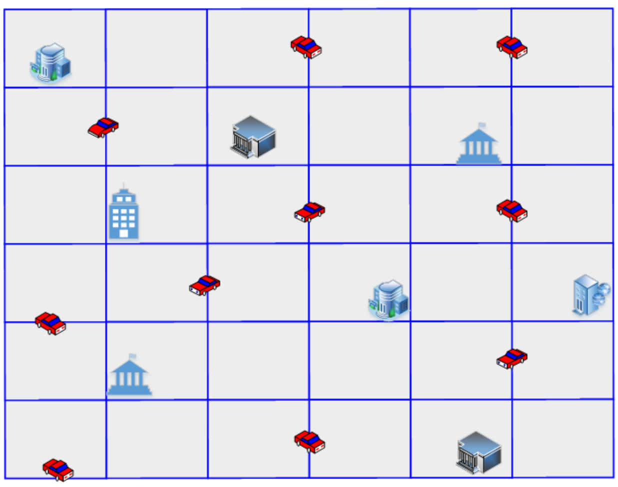


- Use the air quality data, which is collected by a tram in the city of Zurich at different locations over time in **OpenSense** project conducted by researchers at ETH Zurich

Our Proposal: Machine Learning (ML)-based Distributed Air Quality Monitoring Scheme



How to Assign Sensing Locations to Vehicles



An Urban Area of Interest

- Denote $\mathrm{V} = \{1, 2, \ldots, N\}$ as set of vehicles.
- The area is divided into grid of L square sub-areas as $\mathbb{L} = \{1, 2, \ldots, L\}$.
- The measurement which is taken at any location inside the sub-area indicates the sensing value of the cell.
- Each vehicle has a different probability to visit a sub-area during the sensing duration

Main objective is to assign L sub-areas to N vehicles such that the probability that every sub-areas is sensed with the required number times while the prediction error is minimized

Quality of Information Requirements

- Denote m_l as number of different measurements required to take in a sub-areas l to capture the sensing value variation.
- The value of m_l can be determined as the required number of samples to capture 95% of the observed variability with standard deviation (σ) and an accepted error (e) as

$$m_l = \left(1.96\frac{\sigma_l}{e}\right)^2$$

- Denote $\mu_{il}\ (i = 1, \ldots, N; l = 1, \ldots, L)$ as the expected number of measurements that the vehicle i can take at the cell l during the sensing period.
- The value of can μ_{ij} be estimated using the vehicle's trajectory history.

Successful Measurement Probability Aware Location Assignment (SMP-LA)

- Denote x_{il} as a decision variable which is equal to 1 if the sub-area l is assigned to the vehicle i. Otherwise, it is zero.
- Main objective is to maximize value of function F_i on each set of points assigned to vehicle i

$$F_i = \underbrace{\beta_1\left(\sum_{l=1}^{L} P_{il}x_{il}\right)}_{\text{probability of successfully taking a number of required measurements}} - \underbrace{\beta_2\sum_{l=1}^{L-1}\sum_{k=l+1}^{L} x_{il}x_{ik}d_{lk}^2}_{\text{sum of squared inter-cluster distances of assigned points}}$$

where

- d_{lk} is distance between sub-areas l and k
- $\frac{\mu_{il}}{m_l}$ is probability that vehicle i can take m_l number of measurements required in sub-area l
- $P_{il} = \min(1, \frac{\mu_{il}}{m_l})$

SMP-LA: Optimization Formulation

- Main objective is to maximize value of function F_i on each set of points assigned to vehicle i

$$F_i = \beta_1 \left(\sum_{l=1}^{L} P_{il} x_{il} \right) - \beta_1 \sum_{l=1}^{L-1} \sum_{k=l+1}^{L} x_{il} x_{ik} d_{lk}^2$$

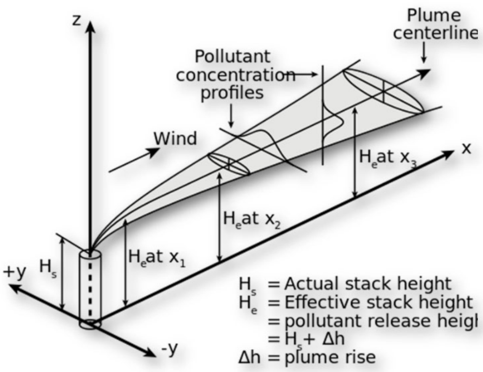
- The optimization problem is formulated as

$$\begin{aligned} & \max \min(F_i, \ldots, F_N) \\ \text{s.t.} \quad & \sum_{i=1}^{N} x_{il} = 1 \qquad && l = 1, \ldots, L \\ & x_{il} \in \{0, 1\} && i = 1, \ldots, N,\ l = 1, \ldots, L \end{aligned}$$

- To linearize the product of two binary variables, we introduce $z_{lk}^i = x_{il} x_{ik}$ and some additional constraints as $z_{lk}^i \leq x_{il} + x_{ik}$

Performance Evaluation

- Use the T-Drive dataset which contains trajectories of 10,357 taxis travelling in the city of Beijing over one week
- Third Ring Road of Beijing is selected as a sensing area which is divided into L = 40,000 (200×200) sub-areas
- To simulate air pollution in the sensing area, we adopt the Gaussian plume equation to calculate the pollutant concentration of downwind position (x, y, z) as

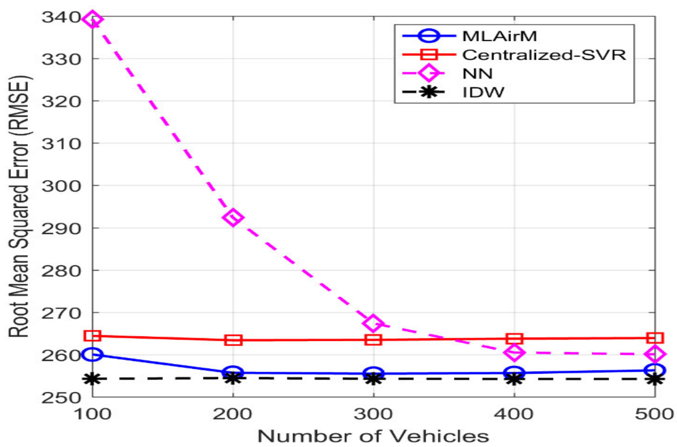


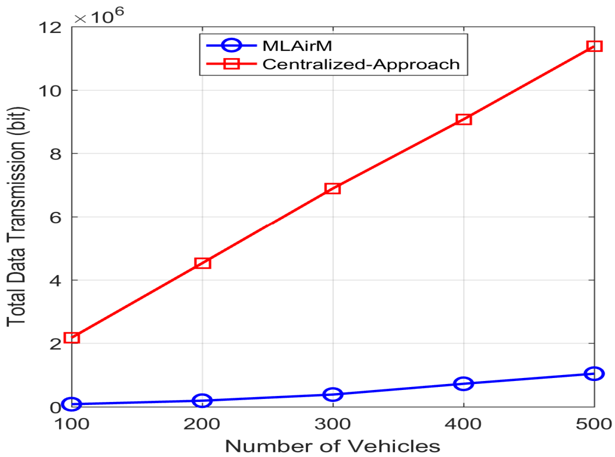
$$C(x, y, z) = \frac{Q}{2\pi u \sigma_y \sigma_z} \exp\left(-\frac{y^2}{2\sigma_y^2}\right)\left[\exp\left(-\frac{(z-H)^2}{2\sigma_z^2}\right) + \exp\left(-\frac{(z+H)^2}{2\sigma_z^2}\right)\right]$$

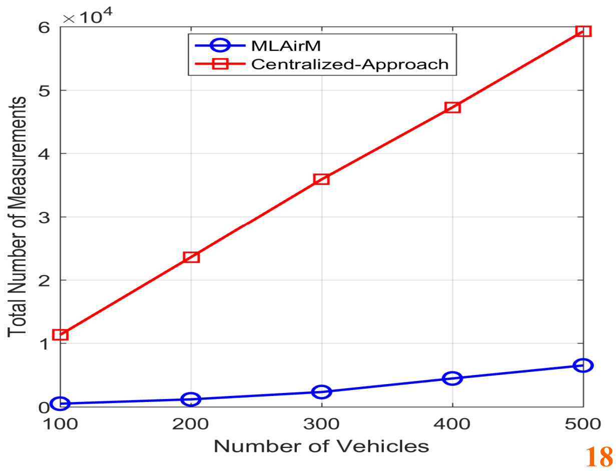
Performance Evaluation

- Evaluate the performance of following algorithms:
  - MLAirM: our proposed scheme = distributed SVR + optimal sensing location assignment
  - Centralized approaches:
    - Centralized-SVR
    - Nearest Neighbor (NN)
    - Inverse Distance Weighting (IDW)
- Performance metrics:
  - Root mean squared error (RMSE): between predicted and accuracy measurements at all sub-areas in the monitoring area.
  - Total data transmission: total number of bits transmitted from all vehicles to the center
  - Total number of measurements: total number of measurements taken by all vehicles

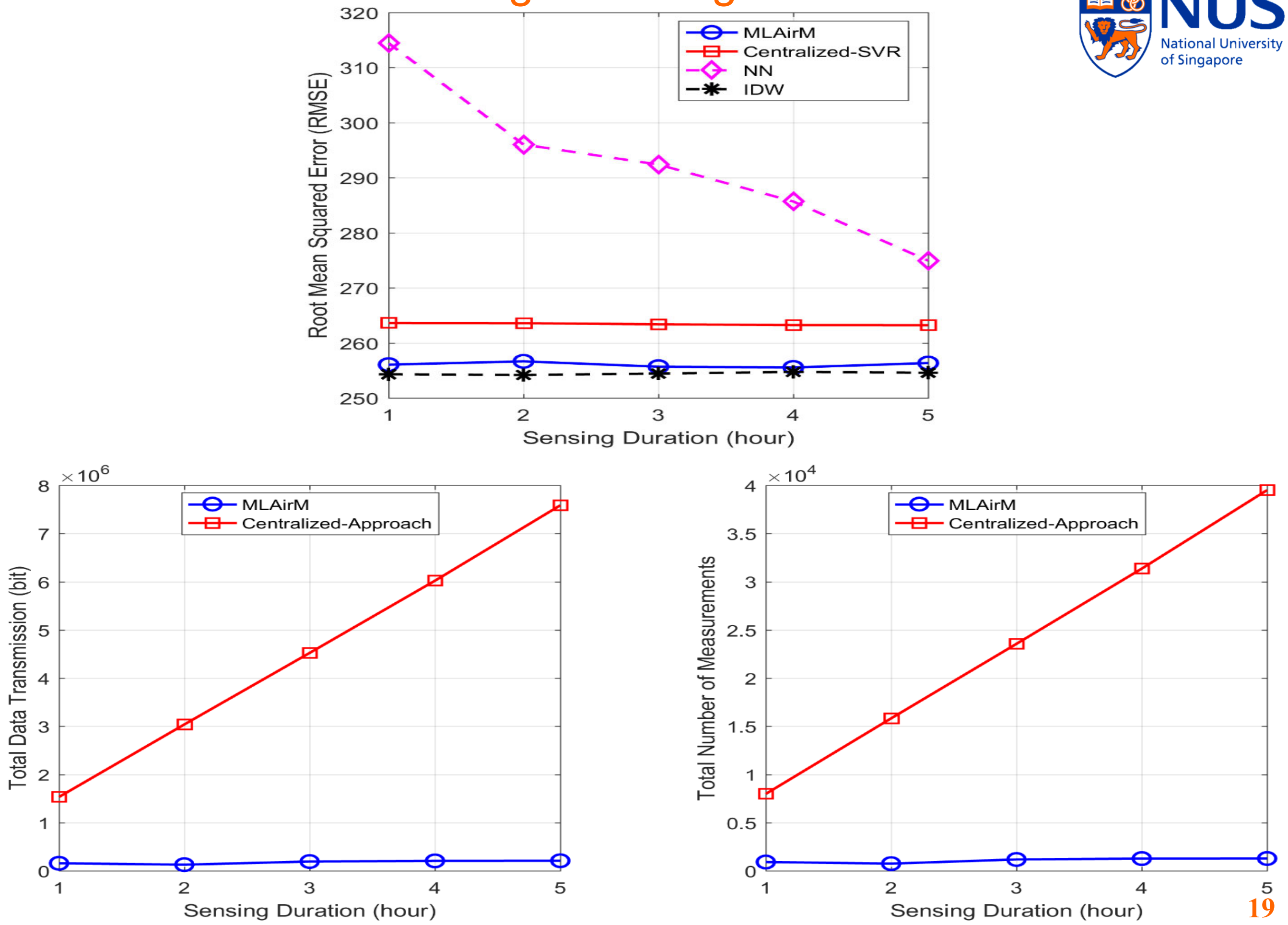
Performance Results: Change Number of Vehicles







Performance Results: Change Sensing Duration



Conclusions

- We propose a machine learning (ML)-based Air quality Monitoring (MLAirM) system using the vehicular sensor networks (VSNs).
- Each vehicle utilizes the support vector regression (SVR) algorithm to learn a local model of the air quality.
- We focused on the problem of optimally assigning the sensing locations to vehicles
- An optimization problem is formulated and a greedy algorithm is proposed to find the assignment solution.
- The simulations results based on realistic vehicular traces show that the proposed MLAirM system can achieve a similar accuracy with a significant reduction in communication and sensing costs compared to other approaches.

Thank You