OpenSense
Open sensor networks for air quality monitoring

Karl Aberer, EPFL
Joint work with: Boi Faltings, Alcherio Martinoli, Martin Vetterli, EPFL
Lothar Thiele, ETH Zürich
and the members of all teams

opensense.epfl.ch
lsir.epfl.ch
OVERVIEW

MOTIVATION

SENSING SYSTEM

RESEARCH PROBLEMS
FROM DATA TO INFORMATION
CONTROL
THE USERS

CONCLUSION
Air pollution in urban areas is a global concern
  • affects quality of life and health
  • urban population is increasing

Air pollution is highly location-dependent
  • traffic chokepoints
  • urban canyons
  • industrial installations
Deaths from Urban Air Pollution

UAP deaths/million

- 0 - 30
- 30 - 60
- 60 - 100
- 100 - 150
- 150 - 200
- 200 - 230


The boundaries shown on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps

OpenSense
Health studies show that air pollution increases the risk of cardiovascular mortality (heart attacks) by 5% to 20% at least.
**Air Pollution Monitoring**

Precise location-dependent and real-time information on air pollution is needed

**Officials**
- environmental engineers: location of pollution sources
- municipalities: creating incentives to reduce environmental footprint
- public health studies

**Citizens**
- advice for outside activities
- assessment of long-term exposure
- pollution maps

Air pollution levels in the city center of Zürich (micro-scale model)
OpenSense

**User Expectations: Nokia User Study**

They are sure they cannot detect all the harmful air pollutants with their senses.

- Family lives in the city center. They have three small children and a dog. The father and the children suffer from allergies. Parents want to offer their children fresh playtime outside. They are very concerned about the environment.

- They feel that they are breathing other's exhaust fumes when commuting by bike.

- They recognize most of the major air pollutants but do not know how they harm people and environment.

- They are sure they cannot detect all the harmful air pollutants with their senses.

- They are interested in health/wellbeing and try to be as ecological as possible.

- They are concerned that the air quality is the roadworks and traffic.

- The worst kind of weather condition is hot and non-windy summer days.

- Furthermore, the road dust in spring time causes eye irritation.

- Family makes trips mainly inside Finland, they visit their relatives’ summer cabins in summer time.

- They follow the real time radar weather information regularly from their smart phones.

- Usually they make short or all-day all-hoc trips in the city. They often go to the indoor and outdoor playgrounds, dog parks and islands where the air is fresher.
MONITORING TODAY

- Few stationary and expensive stations
- Models that extrapolate from pollution sources
- Data mostly inaccessible to the public

NABEL station Zürich
**Opportunities**

- *Wireless communication*: deploy larger numbers of stations
- *Mobility*: deploy mobile stations and increase coverage
- *Communities*: citizens as data producers and information consumers
Air pollution as exemplary use case for other environmental phenomena: Radiation, noise, energy
Sensing System
**Lausanne Deployment**

12 stationary stations
- NO2, CO, Humidity, Temperature
- Solar panel powered
- Communication: GSM, Wireless multi-hop routing

8 mobile stations
- NO2, CO, CO2, Humidity, Temperature
- Positioning module
- Powered by bus
- Communication: GSM

1 prototype station mounted on bus

Vetterli/Martinoli/Transport Lausanne
Lausanne Coverage
Zürich Mobile

1 prototype station mounted on Tram 14
- O3, CO, fine particles
- Communication: GSM, WLAN

Calibration
- Under lab conditions at EMPA
- In real time next to Nabel station Dübendorf

Plan: 10 stations
HARDWARE
Installation @ Tram 3005
**Sensor Behavior**

**Open sampling**
Sensors directly exposed

**Benefits:**
- simple & “slim” solution

**Drawbacks:**
- no absolute concentration values
- noisy signal

**Typical response:**

**Closed sampling**
Sensors inside controlled chamber

**Benefits:**
- absolute measurements

**Drawbacks:**
- complex & bulky
- non-continuous sampling

**Typical response:**

![Graphs showing typical responses for open and closed sampling methods.](image)
**IDEA:** Combine the two approaches and get the benefits of both

*(windtunnel tests)*
FROM DATA TO INFORMATION
**Value of Dense Measurements**

**Traditional approach**
- Few stations
- Low resolution interpolated estimates of pollutant concentrations across massive regions

**Recent results**
- Massive deployment of stations (150) at street-level (2008/2009 New York City Community Air Quality Survey)
- Pollutants of interest heavily concentrated along roads with high traffic densities
The Role of Models

- More data
  - but also more noise
  - but also more redundancy
- Can we produce better quality data?

- Models to the help
  - Traditionally used for interpolation and prediction
  - But also: data cleaning, sensor placement, sparse and mobile sampling

Data-driven statistical and machine learning

Context-aware machine learning

Physics-based

Types of models
CLASSICAL AIR QUALITY MODELS

EMISSIONS

DISPERSION
(transport and turbulence)

TRANSFORMATION
(Chemistry)

DEPOSITION
(dry and wet)

Primary Pollutants
Hydrocarbons
Sulfur Oxides
Carbon Oxides
Particulates

Secondary Pollutants
Photo-oxidation (ozone)
Particles

Acids
Humidity
Rain

Plant Growth

Human and Animal Health

Biogenic

Sources

Industry
Heaters
Traffic

Effects

OpenSense
**Granularity of Models**

- **Macroscale**: 100 km\(^2\)
- **Mesoscale**: 1 km\(^2\)
- **Microscale**: 5 m\(^2\)
- **Statistical**
**Example 1: Inferring Dynamic Density Metrics**

**Goal:** Inferring future dynamics of a time series
- Outlier detection
- Supporting probabilistic querying

**Problem**
- Given time series up to $t-1$
- Estimate probability distribution of values $r_t$ at time $t$

Purely data-driven method

---

**Dynamic Density Metrics Model**

Measured value $r_t$ is a sample of a probability distribution $N(\hat{r}_t, \hat{\sigma}_t^2)$

Different statistics for estimation

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{r}_t$</th>
<th>$\hat{\sigma}_t^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA-GARCH</td>
<td>ARMA</td>
<td>GARCH</td>
</tr>
<tr>
<td>Uniform Thresholding (UT)</td>
<td>ARMA</td>
<td>$u$ (user-specified)</td>
</tr>
<tr>
<td>Variable Thresholding (VT)</td>
<td>ARMA</td>
<td>sample variance of $S_{t-1}^H$</td>
</tr>
<tr>
<td>Kalman-GARCH</td>
<td>Kalman Filter</td>
<td>GARCH</td>
</tr>
</tbody>
</table>

Problem: the true distribution is not observable
How to determine the quality of the estimation?
**Determining the Quality**

**Indirect Method**

Suppose \( p_1(R_1), \ldots, p_T(R_T) \) are the inferred densities and let \( z_t = P(R_t \leq r_t) \) then \( z_t \) is uniformly distributed between \((0, 1)\) when \( p_t(R_t) = \hat{p}_t(R_t) \) [Deibold et. al.].

\[
d\{U_Z(z), Q_Z(z)\} = \sqrt{\sum_{x=0}^{1} (U_Z(x) - Q_Z(x))^2}, \tag{1}
\]

where \( U_Z(z) \) is the ideal uniform cdf between \((0, 1)\) and \( Q_Z(z) \) is the observed cdf of \( z_t \). We call \( d\{U_Z(z), Q_Z(z)\} \) the density distance.
Experimental Evaluation

- **campus-data**: ambient temperature values for over sixty five hours
- **car-data**: more than one hour of GPS data

![Graphs showing results for campus-data and car-data](image)

(a) campus-data

(b) car-data
MODEL-BASED ANOMALY DETECTION

MODEL-BASED ANOMALY DETECTION

original data stream ↓
approximation using user-selected models ↓
detecting anomalies ↓
user confirmation: anomaly is an actual error?
**Example 2: Optimal Sensing for Moving Sensors**

**Goal:** find an optimal sensing strategy, which provides an appropriate balance between “maximize sensing coverage” of moving sensors and “minimize sensing cost (sampling)”?

**Question:** Can segmentation help?

**Model: SVM-based Regression**

- Feature vector for supervised learning
  - Time, $x = \text{position along the bus line, land-use, temperature, humidity, CO2, NO2, CO}$

![Graph showing CO2 levels over time with interpolate and x markers]
Segmentation Strategies

Optimal segmentation
- Dynamic programming, $O(k \, n^2)$ (n readings, k segments)
- Too expensive

Top-down binary segmentation
- Basic strategy: $O(k \, \log n)$
- Binary$^+$: optimized approach in finding segment boundaries

Error-based heuristic segmentation
- Heuristic: segmentation using absolute errors
- Heuristic$^+$: segmentation using relative errors

Near-optimal segmentation
- Binary$^+$ + Heuristic$^+$: $O(k \, n \, \log n)$
Segmentation Results

One day data as training and test sets
- Heuristic is better than Binary, specially for testing
- Large number of segments ($k > 5$) does not help much
**Example 3: Region-based models**

Region-based model
- Identify appropriate regions
- Emission parameter for regions
- Dispersion parameter between regions

Models
- Bayesian model
- Gaussian process interpolations

---

J. J. Li, B. V. Faltings, Towards a Qualitative, Region-Based Model for Air Pollution Dispersion, IJCAI Workshop on Space, Time and Ambient Intelligence (STAMI), 2011.
**Multi-Model Query Processing**

- **Approach**
  - Middle layer produces a model cover from a set of regression models on an area
  - Continuous sensor updates
  - Continuous and ad-hoc queries

- **Goal**
  - Handling spurious updates to the database
  - Minimizes data storage

---

**Aggregate Queries**
CO$_x$ emitted yesterday in Lausanne center

**Continuous Moving Queries**
Give a (in car) pollution update every 30 mins
CONTROL
Control: What is the problem?

Two mobile nodes: who should measure?

1. Node decides individually depending on its state, e.g. calibration
2. Nodes communicate with WSN and coordinate
3. Base station schedules nodes using mobility model: a third node arrives, don’t measure!
4. Air quality model: don’t need measurement!
5. Privacy model: node 1 should measure!
6. Application model (e.g. health service): no measurement needed!
Data Acquisition and Control

- Application model: Relevance and cost
- User activity model: Mobility and user state
- Trust and privacy model: Reliability and security
- Air quality model: Sampling and correlation
- Mobility model: Prediction
- Wireless sensor network: Local coordination
- Sensors: Individual state

Control: translate high level requirement to low level control
Data: translate low level data to high level information
**On-the-Fly Calibration**

**Challenge:**
- Supplied calibration may not match requirements
- Baseline drift due to sensor aging

**Approach:**
- Initial calibration using stationary, high quality instruments
- When deployed periodic recalibration when passing reference station
- Multi-hop calibration: “transfer” of calibration to other sensors

Original calibration performs with an average error of 30ppb

After recalibration the average error drops below 3ppb
ROUTE SELECTION AND SCHEDULING

Route Selection

Measurement Scheduling

Coverage

Uniformity

OpenSense
Measurement Scheduling

Unrestricted coverage

Void zones

Track sampling

Voronoi Tessellation

Mapping to accessible area

Shortest track first scheduling

60% coverage of the Zürich city center with the tram network

approx. 14 measurements per tram for one way
**Mobility Modeling**

- **Goal**
  - Simulate *realistic trajectories* of vehicles
  - Testing different control strategies before deployment
- **Multi-layer integrated simulation**
- **Mobility: SUMO**
  - For section of Lausanne realized
- **Wireless networking: OMNeT++**
- **Air pollution and other models to be integrated**
The Users
Objective: Automatically annotating trajectories of different types of moving objects (cars, people)

**Stops**
- Hidden Markov Model (HMM)
- Stop behaviors

**Moves**
- Map matching
- Transportation means

**Trajectory**
- Land use coverage

---

**User Privacy vs. Data Reliability**

**Participatory sensing**
- Users reveal location
- Semi-honest aggregation server infers user activity
- Obfuscation affects data quality

**Approach**
- Personalized privacy
- Users estimate potential privacy loss

---

Local Privacy Estimation

Possible moves

Obfuscated trajectory

Distortion metric

\[ D(t) = \sum_{Y \in pos_{true}(t)} dist(pos_{true}(t), Y(t)) P(Y, t) \]

\[ D(t_3) = \frac{7}{16} \cdot 1 + \frac{7}{16} \cdot 0 + \frac{1}{16} \cdot \sqrt{2} + \frac{1}{16} \cdot 1 \]
**Experimental Results**

- Real data for electrosmog sensing by Nokia campaign
- Avg Static: static parameters that meet the threshold on the average
- Max Static: static parameters that always meet the threshold
PARTICIPATION - TRUST - PRIVACY

Users as consumers
- Different concerns, perceptions, user groups
- Can we satisfy them simultaneously?

Users as producers
- Incentives for participation
- Trusted data
- Protecting privacy
- Can we reconcile these?

Incentives
User
Sensing System
Utility-based framework

**OpenSense Architecture**

**Context**
- Map data
- Landuse data

**Data Flow**
- Mobile sensors
  - Sensor model (e.g., sensor wear)
  - Mobility model
- Calibration model
- Cleaning model
- Data aggregation server
  - Environment models (interpolation/segmentation)
  - cleaned calibrated data
- sampling for locations considering error, value
- required samples priority
- schedule (measurements, priority)
- local coordination
- sensor locations
- predictions
- sensor status
- predictions

**Control Flow**
- Applications
  - response
  - queries
  - checks data offers submits requests
- Service market
- Data market
- Scheduling component
  - schedule (measurements, priority)
- Service market
  - charges data cost submits offers

**Data**
- required samples priority
- schedule (measurements, priority)

**Control**
- landuse data
- Mobility model
- sensor locations
- predictions
- sensor status
- predictions

**Applications**
- checks data offers submits requests
- response
- queries
CONCLUSIONS
CONCLUSION

End-to-end system view essential
- Investigate all system layers: sensor – user interfaces
- Utility-based framework as integrative approach
- System modeling as a key requirement

Mobility is crucial and challenging at the same time
- Coverage, maintenance, flexibility, data dissemination

Results applicable beyond air pollution
- Complex, distributed, participatory measurement

For more information: opensense.epfl.ch
TEAM

Karl Aberer, EPFL-LSIR, project leader
  • Thanasis Papaioannou, postdoc
  • Zhixian Yan, postdoc
  • Hoyoung Jeung, postdoc
  • Sofiane Sarni, engineer
  • Rammohan Narendula, PhD
  • Mehdi Riahi, PhD
  • Alex Arion, PhD
  • Saket Sathe, PhD
  • Tian Guo, PhD

Boi Faltings, EPFL-LIA, PI
  • Jason Jingshi Li, postdoc

Martin Vetterli, EPFL-LCAV, PI
  • Guillermo Barrenetxea, postdoc
  • Andrea Ridolfi, postdoc
  • Juri Ranieri, PhD

Alcherio Martinoli, EPFL-DISAL, PI
  • Chris Evans, PhD
  • Emanuel Droz, engineer
  • Adrian Arfire, PhD

Lothar Thiele, ETH Zürich, PI
  • Olga Saukh, postdoc
  • Jan Beutel, postdoc
  • David Hasenfratz, PhD
  • Christoph Walser, engineer