The LearnAir Network

Leveraging Sensor Heterogeneity to Improve Air Quality Estimation for the Semantic Web

David B. Ramsay and Joseph A. Paradiso
MIT Media Laboratory

***Abstract*—LearnAir is a semantic web infrastructure that improves the reliability of air quality data. It unifies sensors—from citizen science to government reference—to automatically characterize device error under various meteorological conditions and device states. LearnAir thus leverages known reference sensors to extract the useful measurements from otherwise untrustworthy devices.**

ONE IN EIGHT deaths worldwide is attributed to air pollution exposure, making it the world's largest environmental health risk. Moreover, this estimate has more than doubled in recent years, hinting at a complex and poorly understood epidemiology.[1] The issue requires careful analysis, and standard reference-grade methodology is not portable or affordable enough to scale. Fortunately, as the research community grapples with a balkanized history, shifting federal priorities, and volatility in funding, consumer sensors offer an open and reliable civil sensing alternative as they become more ubiquitous. For individuals in polluted cities, trustworthy data could lengthen their lives and empower their activism. For academics, higher spatial resolution and personal exposure data represent a promising frontier for modeling and research.

Digital Object Identifier 10.1109/MPRV.2019.2919828
Date of current version 20 November 2019.

The growing need for affordable air quality monitoring has resulted in a steady stream of consumer-grade devices on the market over the last several years. Unfortunately, these devices rarely perform well under real-world conditions.[2] As such, the academic and research communities have become increasingly skeptical of affordable sensors and have fallen admittedly "behind the curve" in engaging with and educating the consumer technology and citizen science communities.[3] There is a need for (1) trustworthy and well-characterized data that can affordably scale, (2) rigorous and painless consumer device evaluation, and (3) simple infrastructure to connect the academic environmental science community with consumer device manufacturers and technologists.

We believe these problems can be addressed with a new approach—instead of attempting to redesign affordable devices to provide reference quality data, our goal is to design techniques and

Published by the IEEE Computer Society

systems that rigorously characterize the reliability of devices that are already in use. We expect that—because the core sensing physics are sound—consumer devices will provide reliable data under a limited set of operating conditions. Based on this assumption, we evaluate sensor trustworthiness over a range of conditions and measurement tolerances to quantify both the noise floor of a device (constant, condition-independent error due to physical sensor limitations) as well as the range of conditions that result in systematic mismeasurement (condition-dependent error based on environmental factors like humidity and high wind). For every sensor reading, the model estimates the confidence that the value falls within a user-specified tolerance based on current local environmental conditions. Users can thus trade the volume of data they keep from affordable devices for stronger guarantees about its quality. Furthermore, combining this analysis with a semantic web infrastructure can address many of the problems facing today's air quality ecosystem.

BACKGROUND AND RELATED WORK

Standard practice for measuring air quality is a sparse network of expensive (>$100k) monitoring stations that require regular calibration, and which lack strong correlation with or chronically under-report the pollution exposure of citizens living nearby.[4] These stations use Federal Reference Method (FRM) devices, or equivalent (Federal Equivalence Method, FEM) devices, which are certified by the U.S. Environmental Protection Agency (EPA).

Affordable, portable sensing techniques are based on sound sensor physics, though error-correction strategies typically employed in higher quality devices are sacrificed for cost, size, or power. Combining these core techniques with a basic insight into systematic sources of error provides a unique opportunity for algorithmic analysis of sensor accuracy.

Pollutants and Sensors

Particulate Matter (PM) that is less than 2.5 microns (PM2.5) is small enough to diffuse directly into the bloodstream and has especially serious health implications. Black Carbon (BC) is a complementary measurement—it is a major constituent of PM2.5 and a result of incomplete combustion. It is the most strongly light absorbing component of ultrafine particulate, and makes up an average 5%–10% of the mass concentration of PM2.5 in the U.S. BC has been shown to account for 21%–45% of PM2.5 elsewhere in the world.[5]

Optical sensing is the dominant methodology for low-cost particulate sensing. This modality is susceptible to airflow variation, fog, humidity, accumulated grime, and relative changes in light absorptive constituents like BC. Pollen and dust (which have negatively skewed log-normal distributions centered in the 10–100 micron range) also confound sensors that attempt to extrapolate concentrations of particulate below a few hundred nanometers in diameter, where most of the mass concentration for PM2.5 lies. In outdoor, realistic conditions, the EPA reported the Dylos DC1100-PRO (~$300) had an R^2 score of 0.27–0.86, while the AlphaSense OPC (~$550) showed scores from 0.11–0.68 (and both showed sensitivity to high humidity) compared to FEM references.[2] These values are typical for well-designed sensors at this price point. Cheaper sensors include more significant tradeoffs, but have been shown to track PM2.5 reasonably well in highly controlled, static indoor environments.[6] While colocation tests reveal clear flaws with these sensors in situ, the core modality has a proven record in more expensive devices.

We also analyze three EPA criteria pollutants—O_3, NO_2, and CO. Gas sensor technologies include spectroscopy, chemiluminescence, and chromatography. The AlphaSense (AS) sensors (~$100/each) we chose are based on an electrochemical oxidation/reduction technique and have a good reputation within the air quality community. Despite this, AS sensors are subject to drift, pollutant cross-sensitivity, temperature and humidity dependence, reaction time constraints, and a limited lifespan. Preliminary tests show they might have a slight dependence on airflow.[7] Although they are provided with a piecewise linear calibration based on temperature, these calibrations continue to settle after shipment, and best practice is to calibrate them with FRM colocation data before and after use.[8]

Cheaper versions of chemical sensors exist, two of which we include in this experiment.

Sensors at this price point typically are not calibrated and are not sensitive to small perturbations. They may be useful as a gross indicator for extremely hazardous conditions, but their capability for accurate, continuous measurement in-situ has not been demonstrated.

State of Industry

The air quality research community faces challenges in defining and sharing FRM/FEM data internally, with groups like the Environmental Defense Fund's (EDF) Air Sensor Workgroup actively working to solve this problem. There is also a recognized need to create an open ecosystem that engages with and leverages consumer manufacturers and citizen scientists, but the academic community is lagging behind the entrepreneurial one in engagement and outreach.[3]

Of the many consumer devices released in the past few years, few have been tested rigorously in situ, and none have faired well in co-location studies conducted by organizations like the EPA and the South Coast Air Quality Management District.[2] Despite their importance, the data is not widely cited. Furthermore, colocation studies require significant human intervention, suffer from a lack of standardization, and result in a single R^2 characterization. Environmental groups typically test 1–3 samples of a product for anywhere between two and nine months, carry out analysis by hand and publish results without comparison to similar tests from other geographies.[2]

These studies are an important first step of engagement with citizen scientists and technologists, but they fail to provide a collaborative solution that allows researchers to actually leverage consumer data unless it is of FRM quality (a high bar for a $100 sensor). Collaboration requires a more nuanced, holistic evaluation of consumer sensors and a complete characterization of their sources of error.

RELATED APPROACHES

Air quality sensor networks are a common topic in the research literature. Some of the best work in sensor network algorithm design comes from the ETH Zurich OpenSense project, which has pioneered useful multihop calibration algorithms in heterogeneous networks.[9,10] Matrix factorization strategies have also demonstrated success in handling network sparsity and redundancy, assigning source-specific reliability measurement to each device at the system level without a priori information. This works well to characterize time/condition-invariant sources of error (one clearly better sensor than another, or one user that is chronically misusing their device).

Machine learning has been applied to predict and improve networked sensor data in various fields.[11] Some of the most advanced air quality implementations can be found in Cheng *et al.*'s work, where they implement a state-of-the-art model for PM2.5 measurement and achieve 64% accuracy. While this is the one of the most refined air quality systems to-date, sensor error is again modeled as time- and condition-invariant.

COLOCATION STUDY

We expect affordable sensors to have predictable, systematic errors (like sensitivity to airflow variation) and to have predictable relationships with other sensors (like O_3 from NO_2 and sunlight). We test our ability to leverage these relationships to improve sensor reliability using a two-month colocation study at a Massachusetts Department of Environmental Protection (MassDEP) monitoring site in the Roxbury neighborhood of Boston, MA.

Hardware

The test device (see Figure 1A) includes O_3, NO_2, and CO AlphaSense (AS) gas sensors, a Sharp GP2Y10-10AUoF Optical Dust Sensor, and an off-the-shelf 'SmartCitizen' kit (SCK) featuring a $10 MiCS-4514 MEMS Reduction/Oxidation CO and NO_2 Sensor. These six sensors are logged alongside readings for temperature, humidity, light level, and audible noise. The face of the device is protected but slotted to promote air exposure. A differential pressure sensor is attached with one side exposed to the open face of the device to serve as an inexpensive measure of airflow inside the cover.

The test device was placed 3 feet from a flow-controlled, size-selective MassDEP inlet. This inlet feeds four FEM quality Teledyne machines that measure BC, O_3, NO/NO_2, and CO. A high quality Met One vane-style anemometer was used to capture wind speed and direction.

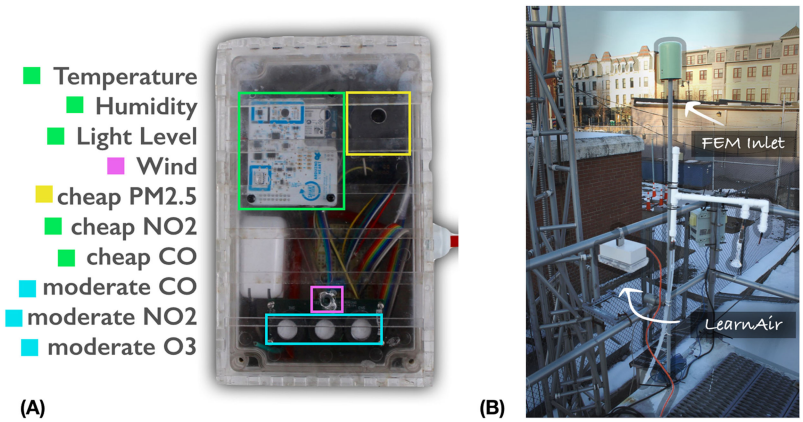

Figure 1. (A) Hardware sensor used for colocation study, with sensor types denoted. (B) The LearnAir test device installed next to MassDEP FEM Reference in Boston, MA.

The colocation study captured a wide range of meteorological conditions as the seasons changed (see Figure 2). Our final dataset included 1.4k samples of hourly BC data, 75k samples of minute-resolved SmartCitizen CO and NO_2 data, and 86k samples of minute-resolved Alphasense data (CO, NO_2, and O_3). Sharp BC estimates—compared to an hourly reference metric that measures accumulated BC through a filter—was calculated as the average of minute-resolved readings over the hour prior to the reference reading.

Raw Data

For our test sensors, we took a simple Least Mean Squared Error (LMSE) calibration approach against the EPA reference to find a proper offset and scale factor for the SCK and Sharp data after applying any suggested mapping from the datasheet. Our technique included slight modifications to ignore certain regions of divergence we expect under nonideal conditions. AS sensors include a piecewise linear temperature-dependent calibration from the factory based on their electrode quiescent currents and sensitivities (which continue to settle after shipment)—we similarly tune the parameters driving this model using an LMSE technique against the EPA reference. The AS O_3 sensor has the calibrated/scaled NO_2 values to which it is cross sensitive incorporated into its calibration.

After calibration, our data capture the types of complications we expect in consumer-level devices—passive airflow and long sensing time constants result in missed transients; sun exposure severely increases the operating temperature and humidity inside the device; even quantization artifacts for pollutants whose concentrations typically fall at the bottom of the sensor's usable range. The cheapest sensors have poor correlation with their FRM counterparts. Fortunately, we also see evidence of condition-variant error in the raw data—the Sharp sensor in Figure 3 clearly diverges during the high humidity, rainy week of

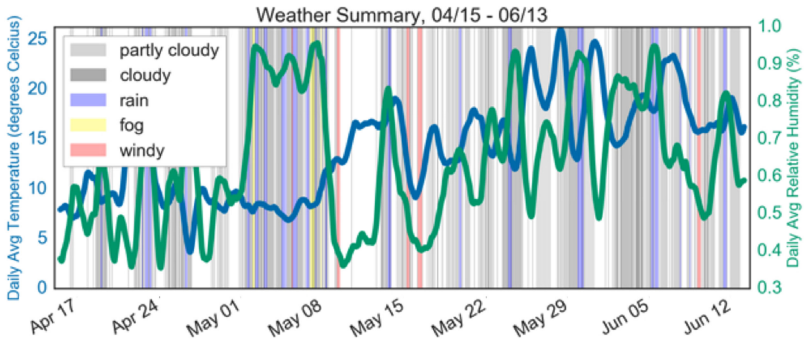

Figure 2. Summary of weather throughout the two month colocation study (24-hr average humidity in green, 24-hr average temperature in blue, and hourly weather type in the background).

Figure 3. Raw data from various sensors (green) compared against the MassDEP reference (orange) over 3–5 representative days. The gas sensors are shown as raw minute-resolved readings, while the Sharp BC data (bottom right) is recorded every hour and averaged over 48 hours. Note the optical Sharp sensor's large divergence corresponds to the week of heavy rain. Our model will evaluate this data against the reference over a range of tolerances, providing insight into the best tradeoff between tight tolerances, data quality (% within spec), and data volume (% removed). Data is selectively eliminated based on patterns in external conditions that lead to error.

our test. This is a realistic dataset to use for testing our efficacy in extracting useful information from off-the-shelf consumer devices and supports the idea that condition-dependent error can be identified.

Modeling

Logistic regression—a frequent choice for failure analysis—provides a desirable probabilistic result. In order to train this binary classification technique using our data, we first assign each reading from the affordable sensor a classification as "accurate" (1) or "inaccurate" (0) based on given tolerance around the ground-truth FRM data. We repeat this over ten tolerances (i.e., accurate within 10, 35, 60, . . . ppb), and observe how the model performs as we tighten and relax the constraint.

Extreme tolerances introduce significant class imbalance, which we correct for in training with a combination of the Synthetic Minority Oversampling Technique and Random Oversampling. We test feature reduction strategies, particularly favoring those that eliminate features biased towards a False Positive Rate in the univariate case and apply a grid search over model parameters using a nested 5-outer/3-inner fold cross-validation technique. These folds are chosen in two ways—randomly/uniformly sampled from the entire data pool or sequentially dividing them into chunks. Shuffled sampling provides the upper bound model performance (by artificially removing the effect of seasonal trends), while sequential/chunked sampling mimics our real-world task (predicting future performance based on previous, contiguous co-location data). We also test Scalable Variational Gaussian Processes (SVGPs)[13] as an alternative method to Logistic Regression.

The 135 features used to predict the accuracy of each sensor reading include:

- The sensor reading itself, which could indicate its usable range,
- The temperature, humidity, pressure, light level, cloud cover, dew point, fog, precipitation level, etc. measured both inside the box and outside from the ForecastIO API,
- Several synthesized features based on the above set, including derivatives of these values to capture rapid changes and long-term averages to capture general trends,
- Features indicating commute times, day of the week, and noise level at the microphone

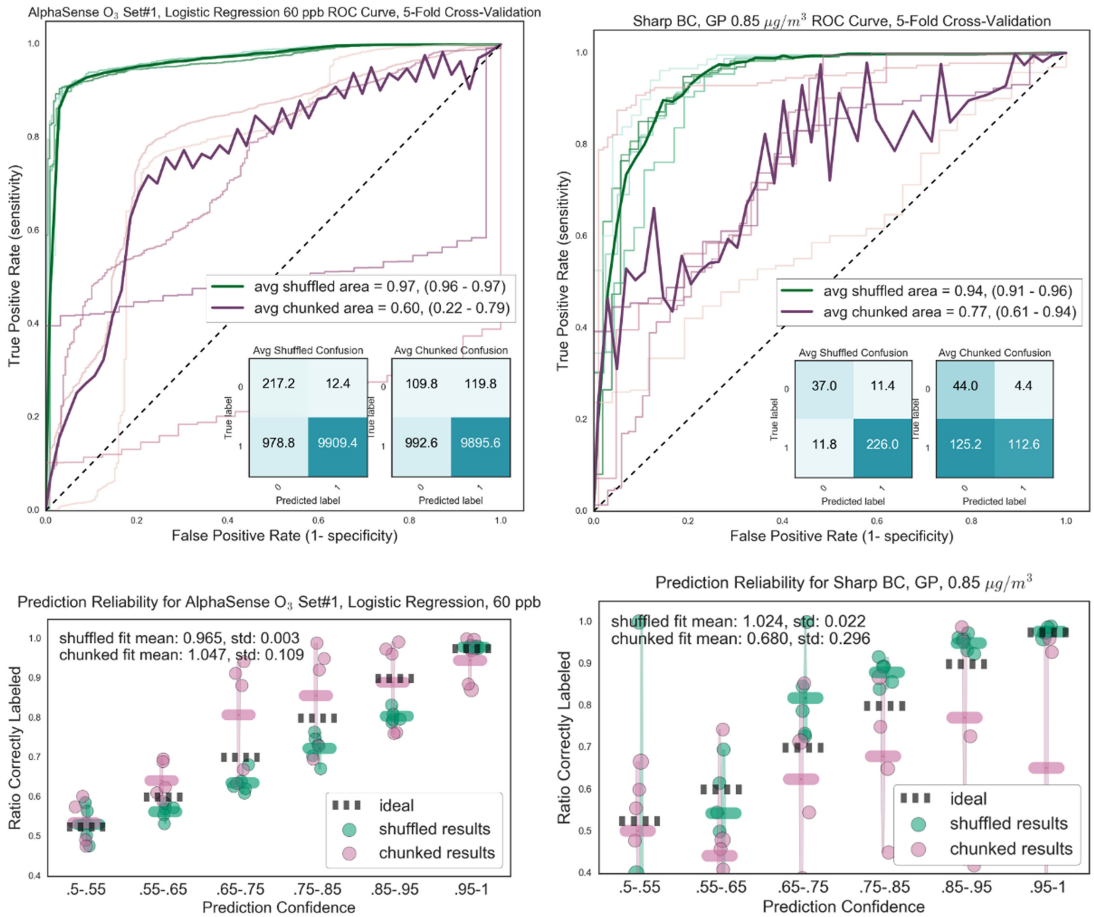

Figure 4. Example ROC Curve and Best-Fit plots; the left side represents a logistic regression model of the AS O_3 Sensor at 60 ppb tolerance, while the right is an SVGP model of the Sharp BC Sensor at 0.85 μg/m^3 tolerance. Results from all five folds are shown, with shuffled results in green and sequential results in purple (dark color represents the average). Corresponding best-fit plots are on the bottom; average best-fit values closer to one are better. We see good predictive power from the chunked model with exception of a couple of folds that have very poor quality (due to predicting one winter week using data trained from other seasons or vice versa), though model fit leaves room for improvement. When we eliminate seasonal effects by training using randomized training/test splits, we see very strong predictions and much tighter fits.

to serve as a simple proxy for traffic or construction,

- Measurements from other affordable gas sensors for which the current sensor may be cross-sensitive (only $<=$ in cost sensors),
- Wind direction, wind speed, airflow at the sensor surface, temperature differential outside to inside the box, and other factors affecting airflow.

RESULTS

In Figure 4, the area under the curve of the receiver operating characteristic (AUC-ROC) scores are complimented with a modified Hosmer–Lemeshow goodness of fit test, broken out into a scatter plot as described in Hosmer *et al.* (bottom of the figure). These plots compare the actual percentage of correct classifications to their labeled confidence across folds (i.e. are ~60% of the values correctly classified when the model reports 55%–65% certainty). Confusion matrices are included as a basic check; a high number of false negatives are acceptable, while high false positives are not. As we move through different tolerances, we shift from nearly all test values having a true label of "inaccurate" to the opposite—the confusion matrix gives us insight

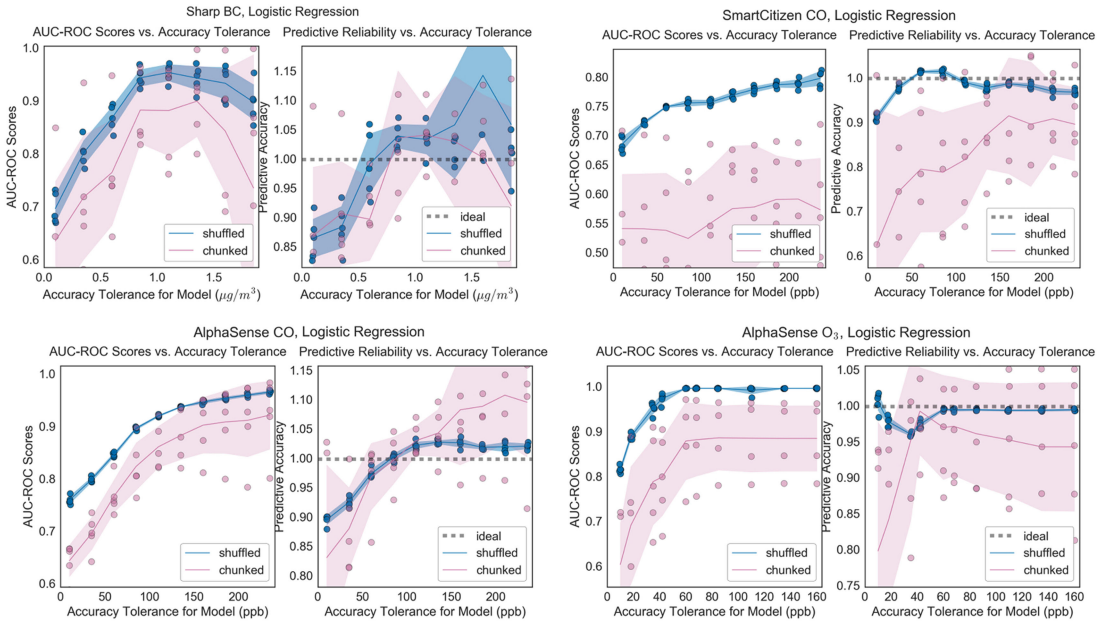

Figure 5. For select models, the change in performance as the tolerance around the reference measurement of what is classified as "accurate" is increased. On the left of each pair is the model accuracy (AUC score, each point represents a single value from the top graphs in Figure 4); on the right is the goodness of fit (closer to 1 is better, each point represents a value from the bottom graphs in Figure 4). Inflection points (indicated with an vertical orange line) in the AUC score suggest a shift from a regime dominated by a device's unpredictable noise behavior to a regime dominated by systemic, predictable errors that can be modeled.

into this imbalance and confidence that the AUC score is meaningful. AUC scores from graphs like Figure 4 are aggregated and summarized over several tolerances for a subset of sensors in the left of each pair of graphs in Figure 5; the corresponding Hosmer–Lemeshow summary (from the bottom of Figure 4) appears on the right of each Figure 5 pair.

We observe different regimes in the predictive quality of the model as tolerance increases—an inflection occurs in the AUC scores as the model moves from a regime dominated by the device noise floor to one characterized by predictable, systemic errors that the model can address. After inflection, we observe that model performance at larger tolerances continues to improve in some cases (as the tolerance is further relaxed, further reducing the effect of unpredictable noise) and degrade in others (as the prediction task becomes more imbalanced and dominated by rare transient events). We select the tolerance at the knee in the AUC summary plots as the tightest accuracy for which we can still meaningfully separate accurate readings from those resulting from systemic device errors. We use these summary graphs to choose a best model (based on the tradeoff between tolerance, the goodness of fit, and ROC-AUC score) for each sensor and present results in Table 1.

We see no increase in performance over standard Logistic Regression models using feature reduction techniques or SVGPs. As we can see in Figure 5, Sequential models do not have strong predictive fits at inflection, and include a large spread in the quality of model performance over different folds, due to a relatively short test during high seasonal variation. In this case, we observe that splits where the test fold shares similar meteorology to the training split perform well, while splits, where the test week is the meteorologically distinct, perform poorly. Despite this, our sequential model still shows improvements of 4%–49% in PPV and ROC-AUC scores of 0.7–0.9. Seasonally agnostic, shuffled models give strong predictive models with tight fits excepting the cheapest SCK sensors.

Discussion

Our test was not long enough to capture seasonal variation sequentially, as demonstrated

Table 1. Summary of best results for logistic regression with shuffled folds.

Tolrence		ROC-AUC (mean ± std)	fit (mean ± std)	ΔPPV
AlphaSense CO Logistic	85ppb	0.86 ± 0.00	0.997 ± 0.003	0.26
AlphaSense O_3 Logistic	60ppb	0.95 ± 0.00	0.961 ± 0.004	0.01
AlphaSense NO_2 Logistic	10ppb	0.89 ± 0.00	0.995 ± 0.008	0.21
SmartCitizen NO_2 Logistic	10ppb	0.81 ± 0.00	1.022 ± 0.003	0.17
SmartCitizen CO Logistic	60ppb	0.75 ± 0.00	1.016 ± 0.002	0.18
Sharp Logistic	0.85μg/m^3	0.94 ± 0.02	1.040 ± 0.020	0.13
Sharp 48 hr Avg Logistic	0.35 μg/m^3	0.98 ± 0.00	1.012 ± 0.060	0.32

from the high variability in the sequential results. Focusing on the shuffled data—a reasonable proxy under the assumption that a model will be used in conjunction with colocation data that has previously sampled that season—we evaluate their utility in two specific use cases: (1) to improve data accuracy and (2) to drive future probabilistic models.

In the first case, our goal is to maximize our Positive Predictive Value (PPV)–the proportion of readings that we label "accurate" that are in fact so. We thus measure our increase in PPV against the fraction of incoming data we must remove to achieve it, as shown in Table 2. As we apply more stringent criteria, we throw away a larger percentage of the data but achieve a corresponding increase in accuracy. For example, we see that our original AS CO readings are accurate 59% of the time at a tolerance of 110 ppb. Our data accuracy increases to 84% and 93%, at the cost of eliminating 45% or 86% of the data, respectively. This is a powerful tool for mining the trustworthy information from generally untrustworthy devices, and useful for status quo analyses that do not tolerate inaccuracy.

While this is a useful technique for data cleaning, we expect a new paradigm for modeling and research replaces it—instead throwing out data because it lacks credibility, probabilistic models can take advantage of *all of the data* as long as its uncertainty is correctly modeled—i.e., when conditions arise for which our sensor is accurate 60% of the time, our model reports 60% confidence and classifies 60% of the readings correctly. Our goodness of fit plots show that it is possible to model this uncertainty well; we believe that models that include this uncertainty will become a dominant paradigm in the near future.

Reading-specific models allow researchers, scientists, and individuals to engage with systems that were previously unreliable and unusable. This approach can also be used to drive condition- and time-invariant network models like the GP model of Cheng *et al*. with more precise inference.

Seasonal Variability and Design Insight

Besides our main findings, useful insights that emerged from our analysis include (1) a method for measuring seasonal variability, and (2) analysis and design tools for consumer devices.

State-of-the-art colocation studies remain a fragmented and ill-defined characterization process; one outstanding issue is a standardized duration across climates. In our work, we see the strong model performance with shuffled data; sequentially trained models, however, sometimes perform poorly, as they predict device behavior in warm summer months with training data from winter (or vice versa). The predictive power of shuffled and sequential models should converge when external conditions have been sampled sufficiently—i.e., quickly in geographies with a stable climate and longer in areas with large seasonal variability. Moreover, this method factors in the device robustness across seasons—devices whose accuracy does not vary with external conditions would not need to be tested as long.

This objective metric ensures models will perform to the standard of the shuffled models

Table 2. Summary of PPV results for representative models at various tolerances/thresholds with shuffled folds compared to the base PPV with no filtering.

			Thresh = 0.5		Thresh = 0.9	
		Base PPV	PPV	% Removed	PPV	% Removed
CO AS Logistic	10 ppb	0.05	0.12	67	0.17	99
	35 ppb	0.17	0.38	62	0.56	97
	60 ppb	0.30	0.57	57	0.79	93
	85 ppb	0.45	0.73	51	0.88	90
	110 ppb	0.59	0.84	45	0.93	86
	135 ppb	0.74	0.91	39	0.98	81
CO SCK Logistic	10 ppb	0.10	0.15	56	nan	100
	35 ppb	0.32	0.47	52	0.73	100
	60 ppb	0.51	0.69	48	0.86	99
	85 ppb	0.66	0.82	44	0.93	97
	110 ppb	0.78	0.90	41	0.96	96
	135 ppb	0.87	0.94	37	0.98	93
SHARP Logistic	0.1 $\mu g/m^3$	0.14	0.21	55	0.15	98
	0.35 $\mu g/m^3$	0.47	0.67	49	0.88	88
	0.6 $\mu g/m^3$	0.70	0.86	32	0.97	79
	0.85 $\mu g/m^3$	0.83	0.96	22	1.00	56
	1.1 $\mu g/m^3$	0.87	0.97	18	1.00	48
	1.35 $\mu g/m^3$	0.90	0.98	17	0.99	51

we present for errors that are systemic and condition-dependent, which we expect to be the majority. Errors arising from sensor degradation over time (from wear, grime, or chemical depletion) can be captured with a linear mapping using this type of modeling—even in sequential tasks—by including sensor age as a feature. To model more complex and dynamic age-related behavior, it would be necessary to capture data across the device's lifespan—effectively recreating the shuffled model using one sensor across season and age as training data for others.

Additionally, these models provide useful information about the sensors and devices themselves. We find some devices and tolerances for which a model fails to identify useful information in a device's output—a clear indication of its quality. To identify these models we compare them to a model trained to identify transients from the mean in the reference. For example, SCK sensors provide no useful information in our study; AS O_3 and CO sensors converge to predictability around 60–85 ppb—much larger than the stated tolerances of the sensors themselves (likely due to enclosure airflow). This value gives us a lower bound on useful data from the device.

We also find that predictive models where ground-truth MassDEP reference data for other pollutants are included as predictive features are useful for (1) capturing codependence or cross-sensitivity of an affordable sensor to other pollutants, and (2) evaluating design possibilities that exploit pairings between high quality and affordable sensors. We found, for instance, that a high-quality BC measurement is strongly predictive of errors correlated with transients.

Finally, we rank predictive features to gain insight into the causes of a device error. For example, feature analysis of the Sharp sensor model shows a distrust of readings when both

the reading itself and the ambient humidity are unusually high. These insights can drive a targeted, iterative design process based on weaknesses identified in situ.

As a two-month study, future in situ testing would strengthen our conclusions around seasonality. With the data we have collected, we find this modeling strategy works well for identifying and characterizing complex condition-dependent, time-invariant sensor errors. It can be applied to data after calibrations for sensor-drift or cross-sensitivity, accurately capturing inherent errors in the calibration process. In the future, it may also be possible to seed new models of novel sensors with existing models of sensors that follow the same patterns of systematic error.

BUILDING IT INTO A NETWORK

After validating our approach, we built and released a set of tools to automatically apply and scale these techniques in a semantic web network. We hope to improve upon and scale the model of current successful crowd-sensing platforms—like Safecast's radiation network—extended with relaxed hardware/software constraints to enable scaling in an already fractured market.[15] The following sections give a brief overview of this new platform and its associated tooling, which is available at *https://github.com/mitmedialab/learnair*.

The Semantic Web

Semantic web technologies aim to decentralize data storage and administration and generalize the rules for creating data ontologies, allowing schemas to develop and compete organically (much like the World Wide Web). Data is connected through hypermedia—an approach that is scalable and robust and leverages existing web technologies for access control and security.

Semantic web technology is being adopted for social linked data with projects like Solid (from Tim Berners Lee), as well as the rise of linked semantic data in web structures like DBpedia or Wikidata. However, centralized approaches continue to dominate the Internet of Things (IoT) landscape. With the consolidation of IoTivity and AllJoyn—the two largest competing IoT communication protocols—there is evidence that semantic principles and distributed, linked, RESTful JSON APIs may be the bedrock of the future.

We have built a LearnAir network typology using ChainAPI, a JSON standard based on Hypermedia Application Language.[16] ChainAPI is a true semantic web infrastructure that has been adopted for large-scale ecological sensor installations. LearnAir extends the ChainAPI framework with an air quality specific data ontology and several new tools for interacting with ChainAPI data structures, designed with industry needs in mind. Our ontology—published at *https://github.com/mitmedialab/learnair*—is a novel semantic web specification designed specifically for air quality network deployments, including (1) a simple schema for managing location metadata, calibration records, and fixed or mobile sensors; (2) tooling to allow reference material and calibration procedures to be centrally updated, shared, or automatically applied—providing manufacturers a window into how their devices are performing across geographies; and (3) "virtual sensor streams" allow for multiple raw, calibrated, and processed data to coexist seamlessly for a sensor or device. It incorporates the best practices from Semantic Web standards with the practical needs of air quality community.

A New Toolkit

Web crawlers that traverse hypermedia relations and understand relational graphs are important for discovery and interaction in hypermedia frameworks. As part of LearnAir, we have also extended ChainAPI with an open source suite of crawlers for traversal, discovery, and data manipulation.

Alongside practical tools for Excel/CSV upload and similar interactions, the most notable addition is called *ChainProcessor,* which provides developers simple hooks to (1) apply arbitrary functions to data in the network and upload it (i.e., apply a calibration to raw sensor data based on local temperature and humidity), and (2) run supervised machine learning tasks based on colocation events. In the second case, a local database (indexed by location and timestamp) is created by the crawler for each denoted resource type. Developers only need to specify the data they are interested in, a spatiotemporal colocation tolerance, and train/apply functions for their algorithm

to automatically discover, retrain, and apply itself to all current and future data in the ecosystem.

This platform and toolkit were implemented and used as part of our LearnAir study—we uploaded our GPS tagged data to an active LearnAir system from CSV files and deployed our calibration algorithms and our data quality models as crawlers in the network. The crawlers find sensors, calibrate data, and model systematic errors for every sensor by type (using colocated events with reference devices that are mined from the web.) Our implementation shows new data triggers retraining of each crawling sensor model automatically. The crawlers then continue to reprocess and replace exposed old data with the latest, most accurate estimates.

Example Hardware

The final part of the LearnAir semantic network infrastructure is a reference hardware design that mates with the AS front-end board, timestamps, and multiplexes sensor data, and pushes it over Bluetooth Low Energy to a phone for GPS tagging/upload. It has a 5V SPI header for additional sensing, and mates to a daughter board that we designed to track contextual information (including light level, UV level, 3-axis pressure/airflow, temperature, humidity, and vibration/orientation). We hope these circuits alleviate the barrier to entry for designers looking to include contextual on-device measurement as a supplement to their core sensing modalities, completing the toolkit for intelligent and scalable air quality semantic web infrastructure.

CONCLUSION

A survey of the consumer air quality landscape reveals that affordable sensors are susceptible to many types of condition-dependent errors that render their data untrustworthy. It is costly to build systems that can mitigate these errors physically, and the increasingly important consumer marketplace is already saturated with error-prone devices. Prior work has focused on improving the quality of data from these devices using new calibration techniques, modeling corrections, and sensor designs; the best of these, however, still cannot produce reference quality data from an affordable device. (High fog in a cheap optical system, for instance, may introduce enough noise to make a correct reading impossible, even with a state-of-the-art model).

In contrast to models that attempt to *correct* unreliable data, we design and validate models that *predict reliability* for each reading based on local meteorology and device state. These models can drive probabilistic frameworks and improve device accuracy by filtering out the subset of data that does not meet probabilistic quality standards. This methodology is useful for quantitative insight into problems other than data quality as well, including (1) the proper duration of colocation field characterizations across climates, and (2) diagnostics of core sensor quality, device design, and failure modes.

We build all these tools into a scalable, open infrastructure that applies these principles automatically, along with a hardware platform that incorporates the most important contextual measurements. This network addresses the problems facing the air quality community—it allows sensors of various quality to coexist in an open and scalable ecosystem, simplifies and improves existing field validation techniques and cross-organization data sharing, and appends reliability estimates to every reading in the database so that useful information can be extracted from otherwise unreliable devices.

In the future, these techniques naturally extend to condition-dependent spatial relationships as we move away from intentional colocation testing toward ad-hoc network-level approaches. While this study assesses the viability of these approaches in a limited context—six sensors closely mounted to an FEM inlet—these techniques naturally extend to spatial error as well. The radius for which sensors are close enough to meaningfully compare will vary across sensor types, pollutants, geographies, climate, and other changing ambient conditions. This variability can and should be modeled in a similarly probabilistic manner within a framework like LearnAir.

LearnAir creates the backbone for an extensible ecosystem that is scalable and automatic, trustworthy and well characterized—a necessary step for researchers to leverage affordable sensor data and for consumers to have transparency and trust in their devices. Furthermore, these

methods are useful as a design and analysis tool for device manufacturers. LearnAir serves as an open platform that can easily be extended and adapted by developers. Its implications extend beyond the air quality space into the future of connected devices more generally.

ACKNOWLEDGMENTS

The Environmental Defense Fund has funded this work. The authors would like to thank Dr. S. Hamburg (EDF), Millie Chu Baird (EDF), and Professor E. Zuckerman (MIT) for their continued involvement. The authors are beholden to MassDEP for allowing them to use their data and site (especially J. Lane and T. McGrath). Finally, we are very appreciative of conversations with Safecast's S. Bonner and P. Franken, as well as MIT Environmental Engineering Department's Dr. J. Kroll and D. Hagan.

REFERENCES

1. WHO News Release. '7 million premature deaths annually linked to air pollution.' March 2014. [Online]. Available: http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/
2. R. Williams *et al.*, "Sensor evaluation report," EPA Evaluation Report EPA 600/R-14/143, 2014.
3. A. Lewis and P. Edwards, "Validate personal air-pollution sensors," *Nature*, vol. 535, no. 7610, pp. 29–31, 2016.
4. S. Steinle, S. Reis, and C. E. Sabel, "Quantifying human exposure to air pollution–moving from static monitoring to spatio-temporally resolved personal exposure assessment," *Sci. Total Environ.*, vol. 443, pp. 184–193, 2013.
5. B. A. Begum, A. Hossain, N. Nahar, A. Markwitz, and P. K. Hopke, "Organic and black carbon in pm2.5 at an urban site at Dhaka, Bangladesh," *Aerosol Air Quality Res.*, vol. 12, pp. 1062–1072, 2012.
6. J. Prabakar, V. Mohan, and K. Ravisankar, "Evaluation of low cost particulate matter sensor for indoor air quality measurement," *Int. J. Innov. Res. Sci., Eng. Technol.*, vol. 4, no. 2, pp. 366–369, 2015.
7. D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory air pollution monitoring using smartphones," *Mobile Sensing*, vol. 1, pp. 1–5, 2012.
8. J. Kroll and D. Hagan, "Personal communication," MIT Dept. Civil Environ. Eng., Cambridge, MA, USA, 2016.
9. O. Saukh, D. Hasenfratz, and L. Thiele, "Reducing multi-hop calibration errors in large-scale mobile sensor networks," in *Proc. 14th Int. Conf. Inf. Process. Sensor Netw.*, 2015, pp. 274–285.
10. D. Hasenfratz, O. Saukh, and L. Thiele, "On-the-fly calibration of low-cost gas sensors," *Wireless Sensor Netw.*, vol. 7158, pp. 228–244, 2012.
11. M. Smith, C. Castello, and J. New, "Machine learning techniques applied to sensor data correction in building technologies," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, 2013, pp. 305–308.
12. Y. Cheng *et al.*, "Aircloud: A cloud-based air-quality monitoring system for everyone," in *Proc. 12th ACM Conf. Embedded Netw. Sensor Syst.*, 2014, pp. 251–265.
13. J. Hensman, A. G. Matthews, and Z. Ghahramani, "Scalable variational Gaussian process classification," *J. Mach. Learn. Res.*, vol. 38, pp. 351–360, 2015.
14. D. W. Hosmer *et al.*, "A comparison of goodness-of-fit tests for the logistic regression model," *Statist. Med.*, vol. 16, no. 9, pp. 965–980, 1997.
15. D. Ramsay, J. Paradiso, and S. Hamburg, "Making air (Quality) visible: Exploiting new technology to dramatically improve atmospheric monitoring," *IEEE Pervasive Comput.*, vol. 17, no. 3, pp. 90–94, Jul.–Sep. 2018.
16. S. Russell and J. Paradiso, "Hypermedia apis for sensor data: A pragmatic approach to the web of things," in *Proc. 11th Int. Conf. Mobile Ubiquitous Syst., Comput., Netw.. Serv.*, 2014, pp. 30–39.

David B. Ramsay is currently working toward the Ph.D. degree at the Responsive Environments Group, the MIT Media Laboratory, Cambridge, MA, USA. Contact him at dramsay@media.mit.edu.

Joseph A. Paradiso is an Alexander W. Dreyfoos (1954) Professor with Media Arts and Sciences and the Director of the Responsive Environments Group, the MIT Media Laboratory, Cambridge, MA, USA. Contact him at joep@media.mit.edu.