 http://dx.doi.org/10.1145/2802083.2808406



# Robust In-situ Data Reconstruction from Poisson Noise for Low-cost, Mobile, Non-expert Environmental Sensing

Matthias Budde, Marcel Köpke, Michael Beigl
*TECO / Pervasive Computing Systems Group*
*Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*
budde@teco.edu, koepke@teco.edu, michael@teco.edu

***Abstract*—Personal and participatory environmental sensing, especially of air quality, is a topic of increasing importance. However, as the employed sensors are often cheap, they are prone to erroneous readings, e.g. due to sensor aging or low selectivity. Additionally, non-expert users make mistakes when handling equipment. We present an elegant approach that deals with such problems on the sensor level. Instead of characterizing systematic errors to remove them from the noisy signal, we reconstruct the true signal solely from its Poisson noise. Our approach can be applied to data from any phenomenon that can be modeled as particles and is robust against both offset and drift, as well to a certain extent against cross-sensitivity. We show its validity on two real-world datasets.**

***Author Keywords:*** Mobile Sensing; Signal Reconstruction; Environmental Sensing; Noise Analysis; Untrained Users; Feature;

***ACM Classification Keywords:*** *C.3 – Special-Purpose and Application-Based Systems – Signal Processing Systems*

## I. Introduction

Air quality sensing is undergoing a paradigm shift as low-cost pollution sensors enable individuals to monitor their surrounding environment [17]. The *U.S. Environmental Protection Agency (EPA)* recognizes growing support for passive monitoring with wearable sensors [2] that are attached to the body or worn in the hand or head. Example technology ranges from multi-sensor devices (e.g. *CommonSense* [8], *Envboard* [5]) over mobile phone accessories and body-worn sensors (e.g. *Breathe* [15]) to sensor-enabled clothing (e.g. *WearAir* [10], *Conscious Clothing* [1]). When compared to standardized procedures, most handling requirements to ensure results of high validity are typically not fulfilled, as:

- Correct and fixed placement of sensors -– in wearables sensors are often placed opportunistically,
- Periodic or constant calibration of sensor — not feasible in a wearable or mobile device
- Standardized measurement process — difficult to reach due to unexperienced user
- Controlled environment conditions — very difficult due to mobility of the user or movement of worn sensor

Thus, the significance of wearable measurements is low and errors restrict the credibility and therewith the use of the gathered information. *Statistical error* refers to a deviation between multiple measurements of the same phenomenon, e.g. due to sensor noise and/or the statistical nature of the sensing process. In contrast, *systematic error* means that any measurement differs from the actual value in the same way, or in other words: The measurement system is de-calibrated. Systematic errors can stem from a number of sources:

- *Low-cost sensors* may be susceptible to systematic cross-sensitivity, e.g. caused by temperature dependencies of electro-chemical sensors, cameras or photodiodes [6].
- *Sensor aging* can introduce both drifts and sudden offsets. Drift may e.g. be caused by dirt deposition, abrupt changes by degradation, e.g. pixel defects in image sensors.
- *Limited parameter control* is an issue if existing smart-phone or smart watch sensors are re-purposed as environmental sensors. Limited hardware control may e.g. result in unwanted or even unnoticed automatic gain or sensitivity adjustments, potentially causing de-calibration.
- *Novice/untrained users*: If sensing requires proper device handling and/or involves assembly, as e.g. with clip-on approaches [4], the changing integrity of the sensing system is another source of errors. In camera-based sensing, a user could e.g. inadvertently put a smudge on the camera lens, which would then create an offset in subsequent readings.

In principle, all of these systematic errors can be quantified and removed from the data. However, it is often not feasible to do this in-situ without recalibration and/or a reference device. We present a simple signal reconstruction scheme for the monitoring of certain environmental phenomena that is robust against the presented errors by reconstructing the true signal solely from the Poisson noise of the erroneous signal. While the signal itself may be skewed or distorted, its noise is a relative property. Our approach is robust against both static and dynamic baseline shifts (offset and drift), as well to a certain extent against cross-sensitivity. We demonstrate the feasibility on data from clip-on particulate matter sensors for camera-phones [4] and from low-cost ozone sensing [13].

## II. Related Work

There is a number of different purposes for which noise is analyzed: One is fingerprinting sensors to uniquely identify them, which has been demonstrated for cameras [7] and accelerometers. In image processing, Poisson noise is usually seen as something undesirable that should be to be removed from an image [3]. However, in that context, the noise in question is the pixel noise of a camera sensor element. In contrast to the *sensor noise* which these approaches deal with, we analyze *sensor data noise*, i.e. the Poisson noise of the signal variation. Kalman filters can be used for state estimation of noisy signals, e.g. for sensor drift correction. Yet, they typically require a priori knowledge about the noise characteristics, i.e. the models used in the filter need to match the physical situation. While Adaptive Kalman Filters can to a certain extent learn these models, they still require readings from other sensors and/or periodic access to ground truth to somehow rate the quality of the current estimation. An approach to reconstruct signals from noisy data or incomplete information is *Compressed Sensing*. Haupt et al. [9] showed that signal information can be obtained from several random Fourier projections. Furthermore, Model et al. [14] demonstrated the use of multi-sensor arrays to reduce data noise of spatially separated signal sources. A more general approach addresses data that is corrupted by some sort of Poisson noise, in particular image noise [12]. However, all these attempts have the assumption in common, that the noisy data does actually represent the signal and is not altered by some sort of systematic error or drift over time. As motivated above, systematic error can also often be attributed to improper handling by untrained users. Interestingly, while research on mobile, low-cost and participatory sensing recognizes the need to ensure credible readings from cheap sensors [11], the focus is seldomly placed on the effects that are caused by novice users. An intuitive approach is to either train participants or try and determine their skill level or reputation beforehand [18] and/or select them accordingly [16]. However, this again requires some kind of ground truth determined by expert users or a series of campaigns, making it an intricate option. So, while some of the related work can deal with certain types of errors, none of the approaches are able to compensate for all of the presented sources of systematic errors without additional knowledge or sensor data. Our approach uses only raw data and is applicable for environmental sensing of any phenomena that can be modeled as Poisson processes.

## III. Approach

The basic idea of our approach is very simple: Measuring an environmental phenomenon that can be modeled as particles is observing its current concentration in a certain measurement volume. Thus, we are basically looking at a series of counting experiments that are conducted in parallel. Because we are only observing a small measurement chamber we have a certain chance that e.g. a particle is present during our measurement or that it is not. This is – generally speaking – the process of counting independent, uniformly distributed events in a spatial volume. This is by definition a spatial *Poisson process*. A vivid example for a Poisson process is observing raindrops on the tiles of a rooftop.

**Data**: raw environmental time series data
$X = \{X_t : t \in T\}$, window size $w$
**Result**: reconstructed signal $Y = \{Y_t : t \in T\}$
`/* 1: moving avg. to smooth */`
$X_S = \mathrm{avg}(X)$;
`/* 2: spline interpolation */`
$s(t) = \mathrm{spline}(X_S)$;
`/* 3: extract noise */`
$N = \{N_t : N_t = X_t - s(t)\}$;
**forall the** $I_t = [t - \frac{w}{2}, t + \frac{w}{2}] \subset T$ **do**
  `/* 4: std. dev. of` $N$ `on interval` $I_t$ `*/`
  $Y_t = \mathrm{stdev}(N|_{I_t})$;
**end**

**Algorithm 1:** Signal reconstruction from Poisson noise.

The number of observed occurrences in a Poisson process fluctuates with a standard deviation of $\sigma_p = \sqrt{n}$ around its mean $n$, or in other words, there is signal dependent noise. From this noise, we can directly calculate the mean concentration of the signal. This of course only works reasonably well if the signal fluctuations $\sigma_p$ are greater than the sensor background noise, since the magnitude of the signal is always higher than that of the noise. Still, this approach has some huge advantages: As noise is a relative property, the inferred concentration is unaffected by drift or de-calibration offset shifts. This allows to derive the actual signal values if the *noise-to-signal-dependency* is known. This statement is generally true, even for non-Poisson processes. In case of a Poisson process however, the dependency is known: The noise will behave like the square-root of the signal.

The algorithm to reconstruct the signal has four steps, as shown in Algorithm 1. We first apply a simple moving average to the noisy data to reduce it to its mean values. Then, a spline interpolation is constructed on the smoothed data in order to determine the mean value on any point within the measurement series. After that, the actual noise can be extracted by subtracting the mean from the raw data. Finally, to obtain a measure of fluctuation, we calculate the standard deviation of that noise on several time intervals $I_t$, which is then linearly correlated to the square-root of the signal mean values corrected for drifts. Whenever a summation is done on parts of the data (e.g. averaging) we weigh this sum

**(a)** **(b)** **(c)**

**Figure 1:** *MobileDust* dataset: (a) sensor design (passive version), (b) active prototype of removable dust sensor attached to *Nexus 4* phone and (c) example light scattering image taken from the recorded data.

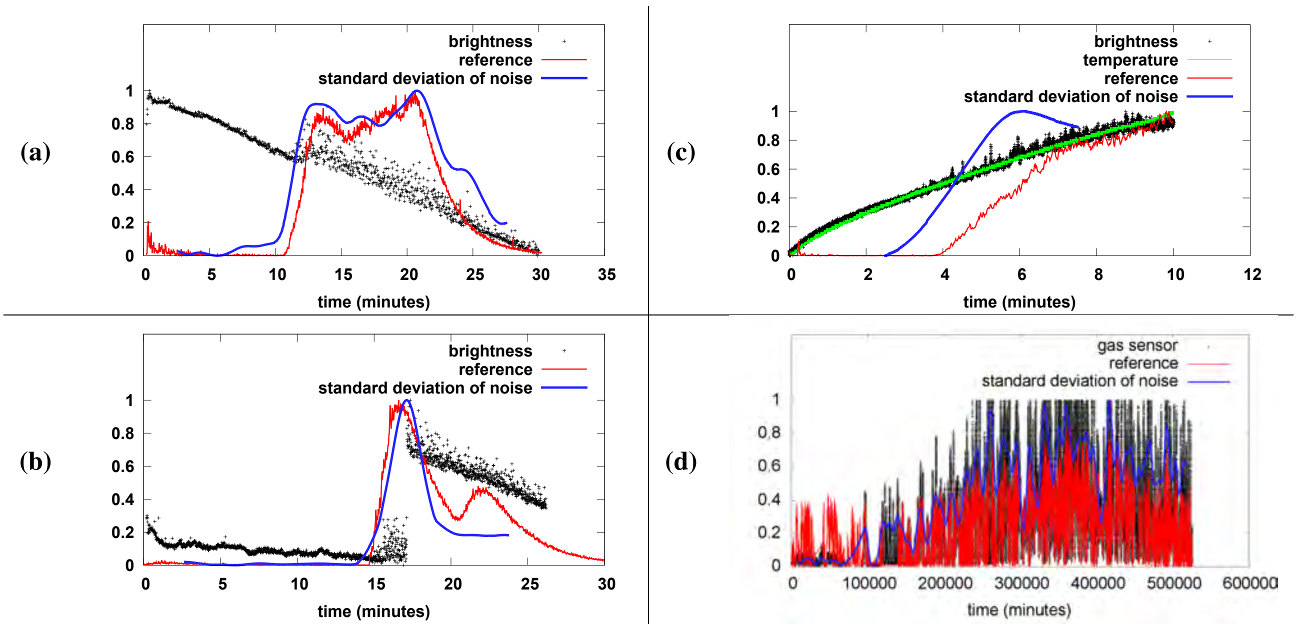


**Figure 2:** Brightness signal vs. particle concentration (all normalized) for different measurements from the *MobileDust* dataset. Our approach is robust against (a) dynamic baseline shifts, (b) sudden offsets and (c) systematic cross-sensitivity, in this example a strong temperature dependency. Applicability on a different phenomenon is shown in (d): raw data and derived signal from the *OpenSense* dataset.

with a Gaussian window to minimize boundary effects of convolution.

## IV. Evaluation

We selected data from two separate real-world projects to test the feasibility for different phenomena: Particulate matter sensing with camera-phone dust sensors (*MobileDust* [4]) and low-cost gas sensing (*OpenSense* [13]). The *MobileDust* dataset was recorded in a lab setting with a *Google Nexus 4* smartphone and a clip-on dust sensor that is integrated in a removable back-cover (see Figure 1). Image resolution was $3264 \times 2448$ and sampling frequencies ranged from $0.5$ Hz to $1.5$ Hz, depending on the settings for the individual measurement run. The phone was placed in a container so that the reference device was exposed to the same air flow. From the images, we calculated the accumulated brightness per picture as feature. As window size for the algorithm, we empirically selected $130 - 150$ samples. Three illustrative examples from the dataset are shown in Figure 2. The plots show that even though the original brightness signal is affected by **(a)** a dynamic baseline shift **(b)** sudden offsets and **(c)** systematic cross-sensitivity (in this example a strong temperature dependency), the true signal can be reconstructed without additional information. The *OpenSense* dataset contains one full year of ozone ($O_3$) measurements, recorded with the low-cost *MiCS-OZ-47* sensor (one reading every 60 seconds) as well as by a fixed station from the national air pollution monitoring network (*NABEL*) in Zürich [13] (10-minute mean values). Figure 2 **(d)** shows an envelope of the original reference signal as signal reconstruction for this dataset.

To evaluate calibration stability, i.e. if it is possible to find fixed parameters to map the noise to absolute concentration values, we analyzed twelve measurement series from the *MobileDust* dataset. All data was recorded using the same smartphone, but in independent sensing runs over the course of one year with long pauses between them. Thus, data was subject to possible systematic errors due to varying temperature, slightly different experimental setup and even changed methods of data collection: The first few datasets were recorded with an older version of the sensing app, which heavily used automatic white balance and exposure compensation. Between sessions, the experimental setup was disassembled and the clip-on sensor fully detached. Measurements were taken at diverse daytimes and seasons, with possible influence of ambient light and temperature on the mean brightness of the pictures. Still, our analysis yielded that the correlation behaves like the square-root of reference, as expected for a Poisson process. The exact model we used is $y = A \cdot \sqrt{x} + B$ with $B = 800$ due to background noise of the sensor. A least mean square (LMS) fit of this model yields the parameters as shown in Table I, $\Delta A$ being the statistical error of the fit. The mean relative statistical error is only $\Delta \bar{A} = 0.38\%$. This is a measure of correctness of the model. Therefore the claim of Poisson distributed data values holds with high probability. Furthermore the standard deviation of the fit parameter $A$ from its mean $\bar{A} = 16414.8$ is $\sigma_A = 1874.4$. So we have relative error of about $11.4\%$

| | data01 | data02 | data03 | data04 | data05 | data06 |
|---|---|---|---|---|---|---|
| $A$ | 19214.2 | 16515.3 | 19424.1 | 16026.0 | 15847.0 | 18162.2 |
| $\Delta A$ | 115.7 | 57.4 | 62.5 | 49.4 | 48.2 | 84.5 |
| | data07 | data08 | data09 | data10 | data11 | data12 |
| $A$ | 13451.2 | 12934.0 | 15961.0 | 16424.4 | 15958.4 | 17059.5 |
| $\Delta A$ | 71.3 | 67.4 | 40.6 | 37.9 | 55.5 | 57.9 |

**Table I:** Calibration parameters derived by LMS fit for the individual measurements. The relative error between parameters is only $11.4\%$.

in adjusting the fit parameters between different measurement series, which is a measure for stability of the calibration approach.

## V. Discussion

While the proposed method is clearly able to compensate for systematic errors, there are some constrains regarding its application. If the dataset has already been smoothed, e.g. due to some commercial and/or inaccessible sensor setup, the time resolution will decrease. Also, there is a restriction to drift compensation. If systematic errors change the signal too rapidly, that is with high frequency, they cannot be distinguished from signal noise. In a sense, it is then possible to regard the systematic error approximately as pure statistical fluctuations. If these fluctuations overcome the signal noise by means of magnitude, it's difficult to apply the noise extraction, since the noise-to-signal-dependency will be blurred. This is not a practical limitation though, as poor data with huge background noise will always be problematic, no matter which method of analysis is used. Applicability regarding drift from cross-sensitivity may be limited in case that both measured phenomena are Poisson distributed. If a gas sensor is e.g. sensitive to two different gases, our approach may not be able to reconstruct the signal, depending on the individual magnitudes. By measuring the noise of the signal, we deliberately sacrifice part of the signal-to-noise ratio (SNR), because signal noise is by magnitude always smaller than the signal itself. Our method thus is a trade-off between SNR and stability. On the one hand, it will never yield as good results as proper characterization and removal of the systematic error, but on the other, no additional information is needed to account for de-calibration and drift of (almost) any kind.

## VI. Conclusion and Future Work

In this work, we presented an elegant signal reconstruction method to compensate dynamic and systematic errors in environmental sensing that can be modeled as Poisson process. Our technique is robust to a variety of error sources that would usually require an advanced measurement setup to control. This includes classes of errors caused by low-cost sensors, limited parameter control and untrained, non-expert users. We have confirmed the principle of operation of our simple approach and shown that it works well on two separate real-world datasets. Overall, we proposed and validated a novel way to account for the natural instability of mobile and wearable measurement setups for end-user environmental sensing. In future work, we will investigate if this method can be extended from stand-alone measurement to combinations with other approaches, e.g. to stabilize readings and compose more sophisticated signal processing techniques.

## VII. Acknowledgments

Partially funded by the German Federal Ministry of Education and Research (BMBF) as part of *Software Campus* (grant 01IS12051). We thank David Hasenfratz for the *OpenSense* gas dataset (http://www.opensense.ethz.ch/).

## References

[1] 'Conscious Clothing' measures the junk you're breathing. http://www.cnet.com/news/conscious-clothing-measures-the-junk-youre-breathing/, June 2013. Accessed April 17th, 2014.

[2] DRAFT Roadmap for Next Generation Air Monitoring. Tech.R., U.S. Environmental Protection Agency, 2013.

[3] Boie, R., and Cox, I. An analysis of camera noise. *Pattern Analysis and Machine Intelligence 14*, 6 (1992).

[4] Budde, M., Barbera, P., El Masri, R., Riedel, T., and Beigl, M. Retrofitting smartphones to be used as particulate matter dosimeters. In *ISWC'13* (2013).

[5] Budde, M., Berning, M., Busse, M., Miyaki, T., and Beigl, M. The TECO Envboard: A mobile sensor platform for accurate urban sensing – and more. In *INSS'12*, IEEE (2012), 1–2.

[6] Budde, M., El Masri, R., Riedel, T., and Beigl, M. Enabling low-cost particulate matter measurement for participatory sensing scenarios. In *MUM'13* (2013).

[7] Chen, M., Fridrich, J., Goljan, M., and Lukáš, J. Determining image origin and integrity using sensor noise. *Information Forensics and Security 3*, 1 (2008).

[8] Dutta, P., Aoki, P. M., Kumar, N., Mainwaring, A., Myers, C., Willett, W., and Woodruff, A. Common sense: participatory urban sensing using a network of handheld air quality monitors. In *SenSys'09* (2009).

[9] Haupt, J., and Nowak, R. Signal reconstruction from noisy random projections. *Information Theory* (2006).

[10] Kim, S., Paulos, E., and Gross, M. D. WearAir: Expressive T-shirts for Air Quality Sensing. In *Tangible, Embedded, and Embodied Interaction*, TEI '10 (2010).

[11] Lane, N., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. A survey of mobile phone sensing. *IEEE Communications Magazine 48*, 9 (2010).

[12] Le, T., Chartrand, R., and Asaki, T. J. A variational approach to reconstructing images corrupted by poisson noise. *Mathematical Imaging and Vision 27*, 3 (2007).

[13] Li, J. J., Faltings, B., Saukh, O., Hasenfratz, D., and Beutel, J. Sensing the air we breathe-the opensense zurich dataset. In *AAAI: Artificial Intelligence* (2012).

[14] Model, D., and Zibulevsky, M. Signal reconstruction in sensor arrays using sparse representations. *Signal Processing 86*, 3 (2006), 624 – 638.

[15] Nieuwenhuijsen, M. J., Donaire-Gonzalez, D., Rivas, I., de Castro, M., Cirach, M., Hoek, G., Seto, E., Jerrett, M., and Sunyer, J. Variability in and agreement between modeled and personal continuously measured black carbon levels using novel smartphone and sensor technologies. *Environ. Science & Technology 49* (2015).

[16] Reddy, S., Estrin, D., and Srivastava, M. Recruitment framework for participatory sensing data collections. In *Pervasive Computing*, vol. 6030 of *LNCS*. 2010.

[17] Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W., Shelow, D., Hindin, D. A., Kilaru, V. J., and Preuss, P. W. The changing paradigm of air pollution monitoring. *Environmental Science & Technology 47*, 20 (2013).

[18] Truskinger, A., Yang, H., Wimmer, J., Zhang, J., Williamson, I., and Roe, P. Large scale participatory acoustic sensor data analysis: Tools and reputation models to enhance effectiveness. In *E-Science* (2011).