A K-means Cluster-Driven Calibration to Improve the Accuracy of Personal Wearable UV Sensors

Thomas Pumir
Princeton University
tpumir@princeton.edu

Emmanuel Dumont
Shade Inc.
em@shade.io

Peter Kaplan
Shade Inc.
peter@shade.io

Shayak Banerjee
Shade Inc.
shayak@shade.io

Abstract

In recent years, there has been a surge in the commercialization of wearable "personal" ultraviolet (UV) sensors designed to help consumers stay ahead of overexposure to UV radiation. To be accurate, these sensors must exhibit a spectral sensitivity to sunlight similar the skin's sensitivity, also called the "erythema spectrum". Sensors, however, never perfectly exhibit this spectral sensitivity and as a result their measurement are erroneous when they are exposed to "unknown" spectra, *i.e.* spectra against which they were not calibrated. On the other hand, because of the atmosphere, sun exposure at the surface of the earth exhibits an infinity of different spectra. In theory, sensors need to be calibrated for all spectra to measure them accurately outside of the laboratory but, in practice, it is impossible. Here, we use a K-means clustering method on the "UVNet" public database of solar spectra to generate what we call "centroid spectra" designed to be replicated in a laboratory and used for calibrating personal UV sensors. We hypothetize that using these centroid spectra for calibration will greatly improve the accuracy of sensors when they face unknown spectra, which we validated through numerical simulations. To our knowledge, this work brings the first rigorous approach into the selection of calibration spectra in radiometry, with useful applications beyond personal UV dosimetry.

1 The two main issues impeding the accuracy of wearable UV sensors

UV exposure, whether from sunlight or tanning beds, has been demonstrated to have on the skin both short-term impacts, *e.g.* a sunburn, and long-term impacts, *e.g.* skin cancer. [Matsumura and Ananthaswamy, 2004]. In recent years, there has been a surge in the commercialization of wearable UV sensors as a quantitative tool to better guide sun protection. Essentially, these products are UV dosimeters that alert consumers when they reach a UV dose that would be harmful to them. Because UV can lead to life-threatening skin reactions in some cases, ensuring that these personal wearable UV sensors are accurate is of paramount importance. The accuracy of UV sensors in real-world conditions, that is when they measure UV radiation from sunlight in direct sunlight, indirect sunlight, in the shadow, under a cloud cover, *etc.*, is primarily driven by their spectral sensitivity, *i.e.* the relative efficiency of detection of light as a function of the wavelength, and the spectral profile of the light source used to calibrate them.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

1.1 Personal wearable UV sensors must follow the erythema spectrum

To be useful to consumers, UV sensors need to exhibit a spectral sensitivity similar to the skin's sensitivity. McKinlay and Diffey [1987] found that the sensitivity of the human skin to light follows a very specific pattern, which they called the "erythema spectrum". This spectrum weights higher-energy UVB (280-310 nm) exponentially higher than lower-energy UVA (310-400 nm). It was later adopted by the World Health organization and standardized as ISO 17166. The erythema spectrum definition is given in Equation 1.

$$w(\lambda) = \begin{cases} 1 & 250 < \lambda \leq 298 \\ 10^{0.094(298-\lambda)} & 298 < \lambda \leq 328 \\ 10^{0.015(139-\lambda)} & 328 < \lambda \leq 400 \\ 0 & 400 < \lambda \end{cases} \tag{1}$$

Where λ is the wavelength in nanometers. The World Health Organization has defined the "UV index" to be 25 mW/m^2 of erythemaly-weighted irradiance.

There is no wearable UV sensor, today, whose spectral sensitivity matches perfectly the erythema spectrum. As a result, all wearable UV sensors will be prone to errors when they are used to estimate the UV index of a light spectrum different from the spectrum of the light source used for their calibration. However, the error arising from this spectral mismatch can be compensated by a thorough calibration against all spectra that the sensors are supposed to measure.

1.2 A proper calibration would require the replication of sunlight in the laboratory

Historically, UV sensors were commercialized to ensure that the UV lamps used in water treatment are still functioning at the desired level of irradiance. In this situation, the UV spectrum of interest is not changing except in its overall intensity over time. As a result, a calibration of these sensors against a similar UV lamp of known intensity would ensure that the sensors remain accurate when measuring the same UV lamps with declining intensity over time.

As opposed to a UV lamp, sunlight is known to exhibit an infinity of different spectra at the surface of the earth, due to the impact of the atmosphere [Iqbal, 1983]. Since wearable UV sensors are used by consumers to measure sun exposure, they will be used to estimate the UV index of an infinite amount of different spectra. The only way to ensure perfect accuracy would be to calibrate UV sensors against all these spectra.

Of course, it is impossible to replicate all sunlight spectra in a laboratory. To solve this problem, we use a K-means clustering method on public database of solar UV spectra to generate what we call "centroid spectra" designed to be replicated in a laboratory and used for calibrating personal UV sensors. We hypothesize that using these centroid spectra for calibration will greatly improve the accuracy of sensors when they are used to estimate the UV index of spectra against which they were not calibrated. We validated this hypothesis through numerical simulations.

2 Using the UVNet database of solar spectra to inform sensor calibration

To better understand what type of spectra personal UV sensors are supposed to be measuring, we used a public database of solar spectra. From 1998 to 2003, the US Environmental Protection Agency (EPA) used ten ground Brewer spectrophotometers across the United States to collect data on UV exposure. Every 6 minutes, each spectrophotometer measured the irradiance of the sun every 0.5 nm from 286.5 nm to 363 nm, except during night and maintenance times. The two primary goals of this governmental effort were to provide information on the geographical distribution and temporal trends of UV radiation in the United States and provides long-term records of UV irradiance to scientists studying effects of UV on biota and materials. The outcome database of these solar UV spectra is the UVNet dataset [UVN].

Specifically, the UVNet database is composed of 73,557 spectra. Each spectrum is a represented by a vector of dimension 154 with the irradiance in W/nm^2/nm from 286.5 to 363 nm every 0.5 nm. Header information include GPS coordinates, the date, and the time of the day. Some sample spectra are represented in figure 1.

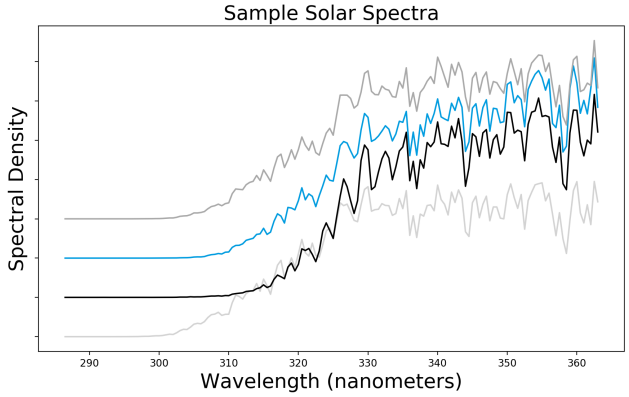

Figure 1: Sunlight spectra measured by spectrometers of the UVNet network. These 4 spectra were sampled randomly among the 73,557 spectra found in the UVNet database.

2.1 Computation of the UV index from the UVNet database

Following the Equation 1, for a given solar spectrum S, we approximated the UV Index of that spectrum by the following trapezoid rule of integration:

$$\text{UVI}(S) = \frac{1}{25 \text{ mW/m2}} \left(\frac{1}{2} \sum_{i=0}^{152} \left[I(\lambda_i) w(\lambda_i) + I(\lambda_{i+1}) w(\lambda_{i+1}) \cdot (\lambda_{i+1} - \lambda_i) \right] \right) \tag{2}$$

where λ_i is the wavelength in nm defined by $286.5 + 0.5i$ for $i = 0$ to 153 and where $I(\lambda)$ is the spectral irradiance in W/m^2/nm.

2.2 Removal of spectra considered to be outliers

Although Brewer spectrophotometers used by the UVNet network are state-of-the-art instruments, due to the external conditions and the limitations inherent to a hardware system, some outlier spectra are inevitably present in the dataset. The maximum UVI ever recorded on earth was 43.3 at an elevation several thousands of meters above sea level in the tropical Andes. We believe such a value would not be found in spectra measured by the UVNet network of Brewer spectrophotometers. Therefore, we remove all spectra whose UVI is above 43.3.

In the rest of this work, we consider the dataset with such outliers removed.

The overall data cleaning process is described below:

1. Compute erythema action spectrum i.e. reweight points
2. Renormalize each measurement by its UVI (computed as in (2)): $I_i = I_i / UVI_i$
3. Compute center of mass: $M = \frac{1}{N} \sum_{i=1}^{N} I_i$
4. Keep the points that are in a range of 3 times the 80th percentile of the center of mass, where the distance to the center of mass of measurement i is defined as $\|M - I_i\|_2$, the euclidean norm of the difference between M and I_i.

3 K-means clustering of the UVNet spectra

3.1 Clustering Algorithms

We now present a methodology to group the spectra with similar patterns. We have no label or no prior information on the spectra besides their vector representation and we want to group these spectra

in a way that would improve the overall accuracy of personal UV sensors if the centroid spectra of these groups were used for calibration.

The partitioning task used herein comprises three parts: (1) identifying the adequate number of clusters (2) a clustering technique that partition the data and finally (3) identifying one relevant spectra per partition.

The idea of clustering goes back to Steinhaus [1957] and clustering techniques are widely used in the data science community. Several algorithms like K-Means [Lloyd, 1982], Mean-Shift [Fukunaga and Hostetler, 1975], Density-Based Clustering [Ester et al., 1996], Expectation-Maximization [Dempster et al., 1977], and Hierarchical Clustering [Defays, 1977], [Sibson, 1973] are available and have been thoroughly studied.

We tried to cluster the UV spectra with these algorithms and found out that the K-means approach yields to a more balanced clustering. In practice, we use the scikit-learn implementation [Pedregosa et al., 2011] of the K-Means algorithm based on Lloyd's algorithm [Lloyd, 1982].

3.2 Using the centroid clusters in calibration

The approach we propose for the calibration process can be divided into two different steps. First, K different spectra are chosen. Then, the response of the sensing device to each spectra is computed using the spectral response of the sensing diode. For each of the selected spectra, this response is compared to the UVI of the spectra, computed as in (2). In order to correct the mismatch between the theoretical prediction and the observed value, the observed value is linearly regressed against the theoretical UVI value. The obtained linear relationship gives the correction of the response. The overall calibration process is described below:

1. pick a number of clusters K
2. Cluster the UVNet dataset using K-Means with K clusters
3. compute the nearest neighbors of the centroids obtained by K-Means in the UVNet dataset.
4. observe the response of the diode to each chosen spectra
5. linearly regress the observed response against the theoreticaly predicted UVI

4 Results and numerical validation

In this section we investigate the performance of the clustering approach to calibration. In particular, for different cluster size, we cluster the UVNet dataset and use the nearest neighbor of each centroid in the dataset for calibration. The quality of the calibration is evaluated as follows: for different threshold values, we display the portion of points such that the relative error between $UVI^{predicted}$, the predicted UVI and the real UVI:

$$\frac{|UVI^{predicted} - UVI|}{UVI}$$

is less than the tolerance. The obtained results are depicted in figure 2. We notice that calibrating with a higher number of points obtained by clustering, yields a greater number points with a relative error than a given tolerance. That is, the calibration accuracy increases with the number of clusters. Besides, we compare the calibration accuracy obtained with the clustering approach with the calibration accuracy obtained by picking random spectra. As explained above, we display the evolution of the proportion of points with relative error less than a certain tolerance. For a number of calibrating sources equal to 3, 5 and 7 we obtain that calibrating with the points obtained with the clustering approach yields a better accuracy. The results are displayed in figures 3, 4 and 5.

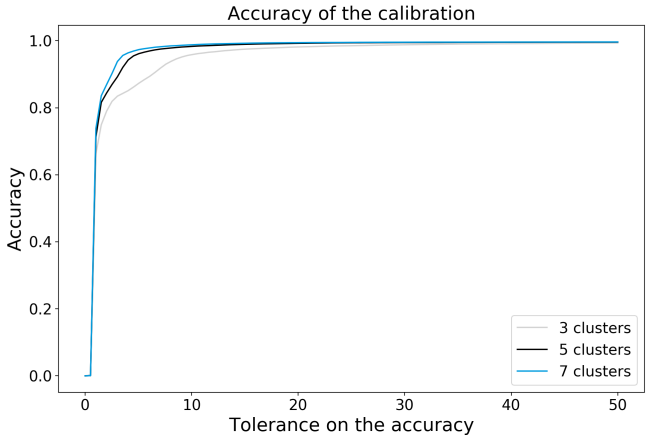

Figure 2: Evolution of the accuracy for different cluster sizes.
X-axis: tolerance values, Y-axis: proportion of points such that the relative error is less than the tolerance.

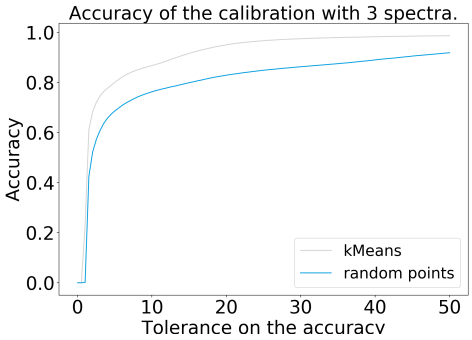

Figure 3: Calibration with 3 Spectra

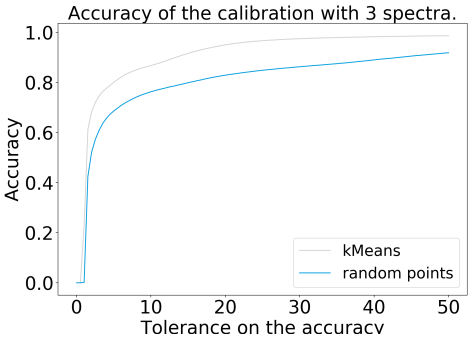

Figure 4: Calibration with 5 Spectra

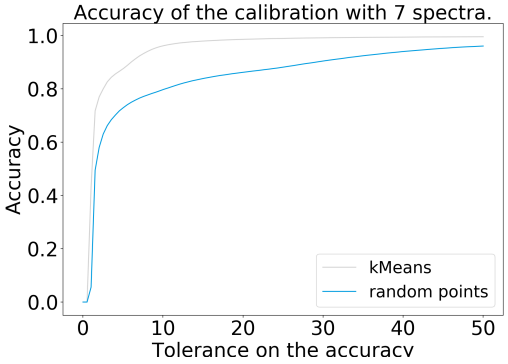

Figure 5: Calibration with 7 Spectra

5 Conclusion

In this work, we proposed to use clustering techniques to identify the most relevant signals among solar irradiance data, in order to improve the calibration process of a UV sensing device. If the calibration process of a UV sensor is well understood, choosing the radiations against which to calibrate can be a delicate question. In particular, we show that it possible to extract a representative spectra of the UV solar radiation as measured on the ground using the UVNet dataset. The methodology discussed here appears to be new in the context of UV sensor calibration and opens the door for other potential applications.

References

Uvnet dataset. URL https://archive.epa.gov/uvnet/web/html/index.html.

D. Defays. An efficient algorithm for a complete-link method. *The Computer Journal. British Computer Society*, 20(4):364–366, 1977.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining*, pages 226–231, 1996.

K. Fukunaga and L. D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

M. Iqbal. *An Introduction to Solar Radiation*, volume XVII. 01 1983. doi:10.1016/B978-0-12-373750-2.50009-4.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, March 1982. ISSN 0018-9448. doi:10.1109/TIT.1982.1056489.

Y. Matsumura and H. Ananthaswamy. Toxic effects of ultraviolet radiation on the skin. *Toxicology and Applied Pharmacology*, 159(3):298–308, 2004.

A. F. McKinlay and B. L. Diffey. A reference action spectrum for ultraviolet induced erythema in human skin. *CIE Research Note*, 1987.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal. British Computer Society*, 16(1):30–34, 1973.

H. Steinhaus. Sur la division des corps matriels en parties. *Bulletin de l'Academie polonaises des Sciences (in French)*, 4(12):801–804, 1957.