Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference of SSCA, 24-28 February 2021; pp 87-92

Augmented Reality: A Computational Framework Applied to Modeling the Dynamics of Air Pollution

Saumyadipta Pyne[1,2], **Ryan Stauffer**[3] **and Benjamin Kedem**[4]

[1] *Health Analytics Network, Pittsburgh, PA, USA.*
[2] *Public Health Dynamics Lab, and Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA.*
[3] *Atmospheric Chemistry and Dynamics Lab, NASA Goddard Space Flight Center, Greenbelt, MD, USA.*
[4] *Department of Mathematics and Institute for Systems Research, University of Maryland, College Park, MD, USA.*

Received: 10 June 2021; Revised: 18 June 2021; Accepted: 21 June 2021

Abstract

In recent years, we have developed a new Augmented Reality (AR) framework to combine real data with computer-generated synthetic samples to “look under the hood”, as it were, for gaining insights into rare, dynamic phenomena. Using data fusion and density ratio model, AR allows us to estimate the tail probabilities of exceeding large thresholds that are far beyond the limited range of observations in moderately sized data. Such thresholds represent extreme events such as the drastic change in air pollution levels in Washington DC caused by lockdown due to the COVID-19 pandemic in 2020, as modeled in this study.

Key words: Data fusion; Tail probabilities; Density ratio model; Synthetic data; Air pollution.

1. Introduction

In its February 4, 2017, edition, *The Economist* claimed that “Replacing the real world with a virtual one is a neat trick. Combining the two could be more useful.” Combining real data with synthetic data produces augmented reality (AR), which, we believe, opens up new perspectives regarding statistical inference. Indeed, augmentation of observations of the real world with virtual information is transforming engineering, healthcare and AI with emerging powerful technologies such as robotics, Internet of Things, and more recently, Digital Twins (Tao and Qi, 2019).

In a recent article, we advanced the notion of repeated AR in the estimation of very small tail probabilities even from moderately sized samples (Kedem and Pyne, 2021). Our approach, much like the bootstrap, is computationally intensive. However, unlike bootstrap, we look repeatedly outside the sample. Synthesis of a given sample of real world observations repeatedly with computer-generated data is based on repeated out of sample fusion (ROSF, Kedem *et al.* 2019; Zhang, Pyne and Kedem, 2020). This strategy proves to be useful for inference in various surveillance applications in which the available datasets usually have a limited range of observations and a moderate size due to limited storage capacity.

In particular, we are interested to estimate the tail probability p of observations exceeding a given high threshold T. Our repeated AR approach is based on numerous data fusions. We use an iterative method that can generate a large number of upper bounds B_i for

Corresponding Author: Saumyadipta Pyne
Email: spyne@pitt.edu

p. Say, such a method is fast and probabilistic, and the upper bounds exceed p with a 95% chance. Thus, many of these exceed p but many do not. Therefore, there are subsequences of ordered upper bounds which approach p from above and from below. We showed how upper bounds can be produced by repeated fusion of real data with computer-generated samples, where the number of fusions is arbitrarily large, and where the support of the generated data is large enough so that it ranges beyond T. Hence, using the connection between the real and generated data, we have a computational approach to "peek" into the realm above T.

Notably, the repeated AR approach allows us to model many phenomena of sudden yet significant change that are of great interest to researchers, *e.g.*, for predicting stock market crashes, disease outbreaks, and extreme climatic events. On January 12, 2021, it was reported in the *New York Times* that "America's greenhouse gas emissions from energy and industry plummeted more than 10 percent in 2020, reaching their lowest levels in at least three decades as the coronavirus pandemic slammed the brakes on the nation's economy". It pointed out that "transportation, the nation's largest source of greenhouse gases, saw a 14.7 percent decline in emissions in 2020 as millions of people stopped driving to work" due to lockdowns that were implemented in many states of the U.S. over the course of the COVID-19 pandemic.

Nitrogen dioxide (NO_2) is a gaseous pollutant emitted from the burning of fossil fuels at high temperatures primarily by vehicles, and thus, its level is a good indicator of traffic volume at a given area over a given interval of time. According to the American Lung Association, NO_2 causes a range of harmful effects on the lungs, including increased inflammation of the airways, worsened cough and wheezing, reduced lung function, increased asthma attacks, and a greater likelihood of emergency department and hospital admissions. The U.S. Environmental Protection Agency's (EPA) National Ambient Air Quality Standard (NAAQS), therefore, measures NO_2 as an indicator for the NO_X family of air pollutants.

Given the sharp reduction in traffic after stay-at-home orders were enforced in many areas of the U.S., in this study, we are interested to model the resulting dynamics of air pollution at a given area. At the capital Washington DC, the stay-at-home order came into effect on April 1, 2020. To analyze the differences between the two periods, pre- and post-order, of 3 months on each side, we resort to two methods. First, we test for similarity in the levels of NO_2 in the morning air in the two periods by using their respective probability distributions. This is done by fusion of data from the two periods as described in a previous study (Kedem *et al.*, 2017). Second, we estimate the tail probability of NO_2 level exceeding $T = 100$ parts per billion (ppb) in each of the two periods. This is done by repeated fusion of the data with computer generated samples (Kedem *et al.*, 2019, Kedem and Pyne, 2020).

2. Data and Methods

2.1. Air Pollution Data

The NO_2 emissions data were collected at four monitoring stations of the U.S. Environmental Protection Agency (EPA) in Washington DC area, for the pre- and post-lockdown periods of January-March and April-June, 2020. In this study, we focused on the morning readings, *i.e.*, the hourly surface levels of NO_2 between 6 am and 9 am. For each period, a random sample of size 200 was selected from the data collected at the locations with the (latitude, longitude) coordinates of (38.895572, −76.958072), (38.921847,−77.013178), (38.970092, −77.016715), and (38.89477, −76.953426). Thus, we obtained a NO_2 sample of size 200 from the first period (January 1–March 31), and another sample of size 200 from the second period (April 1–June 30).

2.2. Testing for equidistribution

Let X_0 be a sample of NO_2 observations of size 200 from January-March, following an unknown probability density (pdf) $g(x)$, $x \in (0, \infty)$, and let $G(x)$ denote the corresponding unknown distribution function (CDF). Similarly, let $X_1 \sim g_1, G_1$ be a sample of size 200 from the second period of April-June, following unknown pdf $g_1(x)$ and CDF $G_1(x)$, $x \in (0, \infty)$. We assume the *density ratio model* (Qin and Zhang 1997, Lu 2007)

$$\frac{g_1(x)}{g(x)} = \exp(\alpha_1 + \beta_1' h(x)) \tag{1}$$

where α_1 is a scalar parameter, β_1 is an 2×1 vector parameter, and $h(x) = (x \log x)$. We now combine or fuse the two samples and estimate the parameters in (1) from the combined sample of size of 400. Kernel density estimates of g, g_1 are shown in Figure 1. From the fits in Figure 1 (bottom panel), we see that the estimated g, g_1 are very close to the corresponding histograms, indicating that the choice of "gamma tilt" $h(x) = (x \log x)$ is sensible.

2.3. Estimation of tail probabilities

The basic idea here is to fuse each NO_2 sample with numerous computer-generated "synthetic" samples. This strategy is referred to as *repeated out of sample fusion* (ROSF in Kedem *et al.*, 2019) or *repeated augmented reality* (repeated AR in Kedem and Pyne, 2020).

If $p = P(X > T)$ is a tail probability to be estimated, we generate numerous upper bounds B's for p where most are above p but many are below p. If $B_{(j)}$ are the corresponding order statistics, then there are $B_{(j)}$ which bound p from above and there are $B_{(j)}$ which bound p from below, yet some $B_{(j)}$ fall in a small neighborhood of p. In this paper, we generated 10,000 such B's. The problem is to find $B_{(j)}$ in a small neighborhood of p. This is addressed by an iterative algorithm which produces subsequences of the $B_{(j)}$ sequence which converge to a small neighborhood of p from above and from below as follows:

$$B_{(j^1)} < B_{(j^2)} < B_{(j^3)} < \cdots < B_{(j^n)} < p < B_{(j_m)} < \cdots < B_{(j_3)} < B_{(j_2)} < B_{(j_1)}$$

where $B_{(j^n)}$ and $B_{(j_m)}$ are very close to p.

We now have two relationships. For a sufficiently large number of fusions, say 10,000, there are $B_{(j)}$ which approach p from above and from below, so that there is a $B_{(j)}$ closest to p. This establishes a relationship between $B_{(j)}$ and p.

With $N = 1000$ (see the remark below), another relationship between $B_{(j)}$ and p is obtained from the well known distribution of order statistics,

$$P(B_{(j)} > p) = \sum_{k=0}^{j-1} \binom{N}{k} [F_B(p)]^k [1 - F_B(p)]^{N-k} \tag{2}$$

where F_B is the distribution of B_i (not $B_{(j)}$), which can be computed since F_B practically coincides with the empirical distribution of $B_1, ..., B_{10,000}$.

The iterative algorithm consists of the following steps, starting from some j.
Step 1:
From (2) we can get the smallest p_j such that

$$P(B_{(j)} > p_j) = \sum_{k=0}^{j-1} \binom{N}{k} [F_B(p_j)]^k [1 - F_B(p_j)]^{N-k} \leq 0.95, \tag{3}$$

Step 2:
From Step 1 we get a j corresponding to the smallest p_j. Use this j and go back to Step 1.

Convergence is reached when for some k. we get the same probability values p_{j_k}. For further details of the algorithm, see Kedem and Pyne (2020).

Remark: We get 10,000 upper bounds $B_1, ..., B_{10,000}$ from which F_B is obtained. However, due to computational limitations, in (2) we use $N = 1000$.

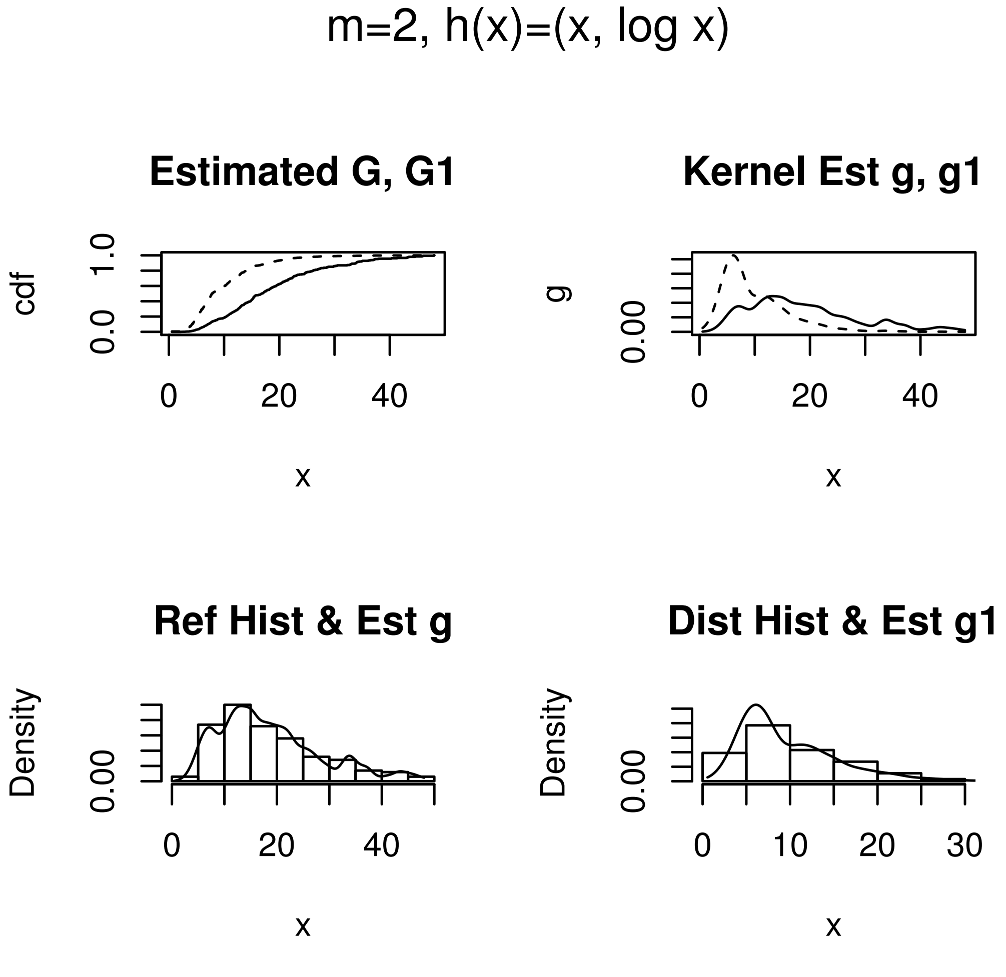

Figure 1: January-March (solid) vs. April-March (dashed) NO_2 distributions

3. Results

The likelihood ratio test of equidistribution $H_0 : \beta_1 = 0$ gives a p-value of 0, indicating that the behavior in the two periods is completely different. This is also seen graphically from the plots of the two CDF's in in Figure 1 (top panel). We see that $\hat{G}$ (solid line) is shifted much to the right relative to $\hat{G}_1$ (dashed line), indicating a great reduction in NO_2 levels in the second period of April-June relative to the first period of January-March.

Indeed a 95% confidence interval for the NO_2 mean in January-March is approximately (14.14,17.24) whereas the same for April-June is approximately (9.11,10.69), again indicating a great reduction in NO_2 levels in latter period.

By setting the threshold $T = 100$, we estimated $p = P(X > 100)$, the probability that NO_2 exceeds a level of $T = 100$, for the two different (pre- and post-shutdown) periods in 2020. Recall that X_0 is a sample of size 200 from the period of January-March. Fusing X_0 10,000 times with generated Uniform(0,180) samples, the algorithm after one iteration gave:

$$B_{(810)} \to 0.00016 \to B_{(808)} \to 0.00016 = \hat{p} = 0.00016 \leftarrow B_{(809)} \leftarrow 0.00016 \leftarrow B_{(813)}$$

Thus, for the period of January-March 2020, we get $\hat{p} = 0.00016$.

Recall that X_1 is a sample of size 200 from the period of April-June. Again, fusing X_1 10,000 times with generated Uniform(0,180) samples, the algorithm gave after eight iterations going down, and a single iteration going up:

$$B_{(500)} \to 6.2e-07 \to B_{(680)} \to 6.2e-07 = \hat{p} = 6.2e-07 \leftarrow B_{(690)} \cdots \leftarrow 2.5e-05 \leftarrow B_{(900)}$$

Thus for the period April-June 2020, we get $\hat{p} = 0.00000062$, which is much smaller than $\hat{p} = 0.00016$ from January-March 2020, echoing the previous results that the NO_2 levels had, in comparison, decreased significantly during April-June 2020 in Washington DC.

4. Discussion

The COVID-19 pandemic has highlighted the need for systematic monitoring and rigorous modeling of dynamic phenomena that can exact a high toll in the form of human suffering and rapid losses in various sectors such as breakdown of supply chains and reduced mobility. Similar lessons are learnt from other areas including extreme climatic events and sudden crashes in the markets. In public health, surveillance is conducted routinely to guard against disease outbreaks and environmental exposures. In this study, we demonstrated how the repeated AR approach could provide a computational framework for modeling the dynamics of air pollution due to traffic emissions during a period of sudden, sharp change.

While estimation of small tail probabilities has long been a topic of research, many of the commonly used methods rely on large number of observations, which makes them less practical for modeling of dynamic phenomena. Methods such as peaks-over-threshold (POT) require observations beyond a threshold, whereas block maxima (BM) require sufficient data such that maxima from each block can be used for estimation. In comparison, both the availability as well as the reliability of computer-generated samples that are representative of real data are increasing with the development of new, powerful computational platforms, e.g., generative adversarial networks (GANs), thus allowing for easier data augmentation.

In this study, we built our AR approach on a density ratio model that starts with a common reference distribution for all sources of information, and then models the individual distributions as *distortions* (*e.g.*, gamma tilt) of that "baseline". Further, we estimate very small tail probabilities even from moderately sized samples. Fusing these repeatedly with computer-generated synthetic data is particularly insightful when the data at hand falls short of the high threshold of interest. Our iterative algorithm constructs a "B-sequence" of bounds that contains a point whose ordinate is very close to the target tail probability, as ensured by the Glivenko-Cantelli theorem. For further details as well as the strengths and the limitations of the repeated AR approach, the reader is referred to Kedem and Pyne (2020) and Kedem *et al.* (2021).

A limitation of the present study is that it does not explicitly account for the fact that the level of NO_2 typically shows a decrease with the advent of spring and summer as it dissociates in sunlight, and tends to collect less near the surface during that period. While the shift in NO_2 was much larger in 2020 compared to previous years, a multi-year extension of our model would be more insightful. Since the primary aim of this study is to introduce our AR computational framework, we plan to address this in our future work.

References

Kedem, B., De Oliveira, V. and Sverchkov, M. (2017). *Statistical Data Fusion.* World Scientific, Singapore.

Kedem, B., Pan, L., Smith, P. and Wang, C. (2019). Estimation of small tail probabilities by repeated fusion. *Mathematics and Statistics*, **7**, 172–181.

Kedem, B. and Pyne, S. (2020). Estimation of tail probabilities by repeated augmented reality. *Journal of Statistical Theory and Practice*, **15**, 25.

Kedem, B., Stauffer, R., Zhang, X. and Pyne, S. (2021). On the probabilities of environmental extremes. *International Journal of Statistics in Medical Research*, **10**, 72-84.

Lu, G. (2007). *Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting.* Ph.D. Dissertation, University of Maryland, College Park.

Qin, J. and Zhang, B. (1997). A Goodness of fit test for logistic regression models based on case-control data. *Biometrika*, **84**, 609–618.

Tao, F. and Qi, Q. (2019). Make more digital twins. *Nature*, **573**, 490–491.