On the Probabilities of Environmental Extremes

Benjamin Kedem[1,*], Ryan M. Stauffer[2], Xuze Zhang[1] and Saumyadipta Pyne[3,4]

[1]Department of Mathematics and Institute for Systems Research, University of Maryland, College Park, MD 20742, USA

[2]Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD 20742, USA

[3]Health Analytics Network, Pittsburgh, PA 15237, USA

[4]Public Health Dynamics Laboratory, and Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15261, USA

Abstract: Environmental researchers, as well as epidemiologists, often encounter the problem of determining the probability of exceeding a high threshold of a variable of interest based on observations that are much smaller than the threshold. Moreover, the data available for that task may only be of moderate size. This generic problem is addressed by repeatedly fusing the real data numerous times with synthetic computer-generated samples. The threshold probability of interest is approximated by certain subsequences created by an iterative algorithm that gives precise estimates. The method is illustrated using environmental data including monitoring data of nitrogen dioxide levels in the air.

Keywords: Tail probabilities, repeated fusion, nitrogen dioxide, epidemiological, order statistics.

INTRODUCTION

Environmentalists, as well as epidemiologists, often encounter the following basic problem. Suppose the value of a high threshold is T, but the data values at hand regarding a certain environmental variable are much smaller than T. Moreover, the data size could be moderately large at best. Based on the data, what is the probability p of exceeding T? For example, if the data values are all less than $T/2$, what is the chance that a future value exceeds T? This generic problem is dealt with in this paper in terms of nitrogen dioxide emission.

According to the American Lung Association (www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/nitrogen-dioxide), nitrogen dioxide (NO_2) is a gaseous pollutant emitted from the burning of fossil fuels at high temperatures primarily by vehicles, followed by power plants, diesel-powered heavy construction equipment and other movable engines. However, most of the NO_2 in ambient air is formed in the atmosphere through photochemical reactions between nitric oxide and other air pollutants. Moreover, 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. When exposed to NO_2, infants and children, due to their greater breathing rate for their body weight, have a higher risk of respiratory failure.

Epidemiological studies have demonstrated associations between NO_2 exposure and premature death, cardiopulmonary effects, intensified allergic responses, and lower birth weight in newborns. It was found that exposure to NO_2 and other outdoor air pollutants shortened the survival of lung cancer patients [1]. The International Agency for Research on Cancer (IARC) classified outdoor air pollution and particulate matter (PM) as carcinogenic (Group 1); and the evidence of a long-term effect of NO_2 on mortality has been found to be as great as that of PM [2]. Consistent evidence of a relationship between NO_2 with lung cancer was noted by a systematic meta-analysis [3]. The U.S. EPA's National Ambient Air Quality Standard (NAAQS) (dep.wv.gov/daq/planning/NAAQS/Pages/default.aspx) measures NO_2 as an indicator for the NO_x family. To protect the public's health from outdoor air pollution, NAAQS set the 1-hour NO_2 standard as 100 parts per billion (ppb) and an annual (arithmetic average) standard of 53 ppb.

Repeated Fusion

Regarding the posed exceedance problem, as they are, the data do not give much information about such questions, and particularly so when the samples are not large. However, the situation differs dramatically if somehow we could only have a "peek" into the domain above the threshold. And if we can do that once, then we can repeat that numerous times henceforth. A way of doing that is by repeated fusion of real and synthetic or artificial data. As we shall see, synthetic data can enhance patterns in real data, a statistical idea highlighted by augmented reality (AR) explored in Kedem, De Oliveira, and Sverchkov (2017, Ch. 5),

*Address correspondence to this author at the Department of Mathematics and Institute for Systems Research, University of Maryland, College Park, MD 20742, USA; E-mail: bnk@umd.edu

E-ISSN: 1929-6029/21

Kedem *et al*. (2019), and in Kedem and Pyne (2021) [4-6].

Our approach to the estimation of the small exceedance or tail probability p is based on *numerous fusions*. It runs as follows. Given any method which produces numerous upper bounds B_i for p. Say, upper bounds which exceed p with a 95% chance. Then many upper bounds exceed p but many do not. Therefore, there are subsequences of ordered upper bounds which approach p from above and from below. In this paper, the numerous upper bounds are produced by *repeated fusion* of the 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, if the given and generated data are somehow connected, we then have a way to "peek" into the realm above T.

Repeated fusion of the data with external computer-generated data is referred to as *repeated out of sample fusion* or ROSF. Unlike the bootstrap, we seek information repeatedly outside the sample.

To describe our approach, we must first review some ideas and then illustrate an iterative method for the estimation of small threshold probabilities. Thus, as such, part of this paper is a review.

The following items are new. Lemmas 0.2 and 0.3 which support Proposition 0.1 are new. If $B_{(j)}$ are ordered upper bounds, then Lemma 0.3 predicts when there is a shift in $B_{(j_i)}$ subsequences converging to p from above and from below. Also new is the fact, not emphasized hitherto, that already by themselves the quartiles and mean of numerous upper bounds for the true threshold probability p provide useful approximations for p, as illustrated in Tables **1**,**3**,**5**, and **7**. This is a fast, albeit crude, way to assess the magnitude of tail probabilities. We will show how to improve these crude assessments. The NO_2 data analysis is new as well. Certain technical details are described in the APPENDIX.

METHODS

Upper Bounds for p by Data Fusion

Let X be a random variable. The problem is to estimate a small tail probability $p = P(X > T)$ for a given threshold T from a *reference sample* $X_0 = (X_1, ..., X_{n_0})$, where $\max(X_0) < T$. This section follows for the most part Kedem *et al*. (2019) and Kedem and Pyne (2021) [5, 6].

Assume that X_0 is from some unknown reference probability density (pdf) $g(x)$, $x \in (0, \infty)$, and let $G(x)$ denote the corresponding distribution function (CDF). Since $\max(X_0) < T$ and since $g(x)$ is unknown, there is not much we can say about p. However things are very different when an independent random sample X_1 exists from a distribution with pdf $g_1(x)$ and CDF $G_1(x)$ supported over a region stretched beyond T. We shall assume that X_0 and X_1 "talk" to each other via a relationship between their distributions, whence useful information is gained about p.

Let $X_1 \sim g_1, G_1$ be a computer-generated random sample of size n_1 and consider the fusion of X_0 and X_1,

$$t = (t_1, \ldots, t_{n_0+n_1}) = (X_0, X_1), \tag{1}$$

of size $n_0 + n_1$. We shall assume the *density ratio model* [7]

$$\frac{g_1(x)}{g(x)} = \exp(\alpha_1 + \beta_1' h(x)) \tag{2}$$

where α_1 is a scalar parameter, β_j is an $r \times 1$ vector parameter, and $h(x)$ is an $r \times 1$ vector-valued function. Clearly, to generate X_1 we must know the corresponding g_1. However, beyond the generating process, we do not make use of this knowledge. Thus, by our estimation procedure, none of the probability densities g, g_1 and the corresponding CDF's G, G_1, and none of the parameters α_1 and β_1 are assumed known, but, strictly speaking, the so called tilt function h must be a known function. However, in the present application the requirement of a known h is weakened considerably by the mild assumption (3) below, which may hold even for misspecified h, as numerous examples with many different tail types show. Accordingly, based on numerous experiments, some of which discussed in Kedem *et al*. (2019) and Kedem and Pyne (2021) [5, 6], for non-negative data we shall assume the "gamma tilt" $h(x) = (x, \log x)$. Further justification for choosing the gamma tilt is provided by the rather precise p-estimates in the first eight entries in Table **10** below. Notice that the "normal tilt" $h(x) = (x, x^2)$ is used in the last entry in the table.

Under the density ratio model (2), the maximum likelihood estimate of $G(x)$ based on the fused data $t = (X_0, X_1)$ is given in (15) in the APPENDIX along with its asymptotic distribution described in Theorem 0.2.

From the theorem we obtain confidence intervals for $p = 1 - G(T)$ for any threshold T using (18). In particular we get an upper bound B_1 for p. In the same way, from additional independent computer-generated samples $X_2, X_3, ..., X_N$ we get additional upper bounds for p. Thus, from the repeated fusions

$$(X_0, X_1), (X_0, X_2), (X_0, X_3), ..., (X_0, X_N)$$

the density ratio (2) assumption produces the sequence of upper bounds

$$B_1, B_2, ..., B_N$$

which, conditional on X_0, is then a sequence of independent and identically distributed random variables from some distribution F_B.

It is assumed that

$$0 < F_B(p) < 1 \tag{3}$$

so that

$$P(B_1 > p) = 1 - F_B(p) > 0.$$

Let $B_{(1)}, B_{(2)}, ..., B_{(N)}$ be the corresponding sequence of order statistics from smallest to largest. Then, as $N \to \infty$, $B_{(1)}$ decreases and $B_{(N)}$ increases.

Theorem 0.1 As N increases,

a. With probability approaching one,

$$B_{(1)} < p < B_{(N)} \tag{4}$$

b. F_B can be approximated arbitrarily closely.

c. There are $B_{(j_i)}$ subsequences that take values in a neighborhood of p.

d. For all $N > N_0$, for some sufficiently N_0,

$$0 < p < F_B^{-1}(0.05^{1/N})$$

Proof. Assumption (3) implies **a**. Part **b** follows from the Glivenko-Cantelli Theorem, and the fact that the number of fusions of X_0 with computer-generated samples is arbitrarily large. More precisely, let $\hat{F}_B$ be the empirical distribution obtained from the sequence of upper bounds $B_1, B_2, ..., B_N$. Then as the number of fusions of X_0 with computer generated data grows, $\hat{F}_B$ converges to F_B almost surely uniformly. To prove **c**, we construct a particular subsequence. For some j_1, there is $B_{(j_1)}$ such that $P(B_{(j_1)} > p) = Q_1$. Hence, the smallest p^* which satisfies $P(B_{(j_1)} > p^*) \le Q_1$ is not far from p. It follows that the closest $B_{(k_1)}$ to p^* falls in a neighborhood of p. Repeating this with $j_2, j_3, \cdots$ we obtain a subsequence $B_{(k_1)}, B_{(k_2)}, B_{(k_3)}, \cdots$ with values in a neighborhood of p. Part **d** follows from the distribution of $B_{(N)}$.

Since the number of fusions can be as large as we wish, *our key idea*, F_B is known for all practical purposes. Hence, from **d**, we see that for sufficiently large N, F_B provides information about p [8]. Clearly, already by themselves, *the quartiles and mean of* F_B *provide useful approximations for* p. We shall illustrate this fact below.

"Down-up" Subsequences

For a sufficiently large number of fusions N, we show how to produce subsequences $\{B_{(j_i)}\}$ which approach p from above ("down") and from below ("up").

A relationship between j and p is obtained from the well known distribution of order statistics,

$$P(B_{(j)} > p) = \sum_{k=0}^{j-1} \binom{N}{k} [F_B(p)]^k [1 - F_B(p)]^{N-k}. \tag{5}$$

This probability is readily available since N is arbitrarily large, and hence, F_B is known for all practical purposes by Theorem 0.1- **b**.

Consider now only $B_{(j_i)}$'s in a neighborhood of the true p, all satisfying the inequality,

$$P(B_{(j_i)} > p) \le 0.95. \tag{6}$$

Observe that (6) is satisfied for small $P(B_{(j_i)} > p)$ when $B_{(j_i)}$ lies to the left of the true p. Suppose now we solve (6) with $B_{(j_i)}$ along some p-increments, and find $B_{(j_2)}$ nearest the smallest p which satisfies (6). Thus $p \approx B_{(j_2)}$. Replacing $B_{(j_1)}$ with $B_{(j_2)}$ in (6) we obtain another approximation for p. We keep doing that to obtain a subsequence $B_{(j_1)}, B_{(j_2)}, B_{(j_3)}, ...$ Clearly, depending on the p-increment, some of the

$B_{(j_i)}$ fall to the left of the true p and some to the right. This process stops when the next $B_{(j_i)}$ is equal to the previous one.

The preceding argument suggests the following Iterative Algorithm for the estimation of tail probabilities proposed in Wang (2018) and Kedem *et al*. (2019) [5, 9]. Recall the order statistics formula (5).

Iterative Algorithm

1. Let $B_{(j)}$, $j=1,\cdots,N$, be the order statistics obtained from $B_1,...,B_N$.

2. Choose a starting point $j=j_1$ and find the smallest p that satisfies the inequality

$$\sum_{k=0}^{j-1}\binom{N}{k}[\hat{F}_B(p)]^k[1-\hat{F}_B(p)]^{N-k} \le 0.95 \tag{7}$$

where $\hat{F}_B$ is the empirical distribution function of the B_i's. Evaluate (7) along p-increments of size α: $\alpha, 2\alpha, \cdots$. Find the smallest integer $k=k_1$ such that $p=k_1\alpha$ satisfies (7).

3. Find j_2 such that $B_{(j_2)}$ is the smallest $B_{(j)} \ge p_1$.

4. Repeat steps 2 and 3 for j_2.

In general, starting with any j, convergence occurs when for the first time $B_{(j_k)} = B_{(j_{k+1})}$ for some k and we keep getting the same probability p_{j_k}. That is, the process stops when p_{j_k} keeps giving the same $B_{(j_{k+1})}$. Thus, the algorithm produces the iterative steps,

$$j_1 \to p_{j_1} \to j_2 \to p_{j_2} \to \cdots j_k \to p_{j_k} \to j_{k+1} \to p_{j_k} \to j_{k+1} \to p_{j_k} \cdots$$

For each starting point j, the algorithm produces a sequence $j_1, j_2, \cdots$. The first thing we wish to show is that such a sequence is either non-decreasing or non-increasing. To show this, we note that the left-hand side of (7) is non-increasing in p. Thus, solving (7) is equivalent to solving

$$\sum_{k=0}^{j-1}\binom{N}{k}[\hat{F}_B(p)]^k[1-\hat{F}_B(p)]^{n-k} = 0.95 \tag{8}$$

To solve (8), we need the following lemma.

Lemma 0.1 Suppose $X \sim b(n,\theta)$ and $Y \sim beta(j+1, n-j)$, then

$$P(X \le j) = P(Y \ge \theta).$$

Proof. See Casella and Berger (2002) [10], Problem 9.21, p. 454.

By Lemma 0.1, conditional on $\hat{F}_B(p)$, we have $P(Y \ge \hat{F}_B(p)) = P(X \le j-1)$ where $Y \sim beta(j, n-j+1)$ and $X \sim b(n, \hat{F}_B(p))$. Then we can rewrite (8) as

$$P(Y \ge \hat{F}_B(p)) = 0.95 \tag{9}$$

and the solution of p is $\hat{p}_j = q^{\hat{F}_B}_{q^{beta(j,n-j+1)}_{0.05}}$ where $q^{\hat{F}_B}$ and $q^{beta(j,n-j+1)}$ are the quantiles of $\hat{F}_B$ and $beta(j, n-j+1)$, respectively.

We wish to show $P(B_{(j)} \ge \hat{p}_j) = 1$.

Lemma 0.2 For every j, $B_{(j)} \ge \hat{p}_j$ almost surely.

Proof. Since $\frac{j}{n} \ge q^{beta(j,n-j+1)}_{0.05}$ for every j, then $\hat{F}_B^{-1}(\frac{j}{n}) \ge \hat{F}_B^{-1}(q^{beta(j,n-j+1)}_{0.05})$ almost surely by the monotonicity of $\hat{F}_B$. Then by the definition of order statistics and the quantile function, we have $B_{(j)} = \hat{F}_B^{-1}(\frac{j}{n})$ and $\hat{p}_j = q^{\hat{F}_B}_{q^{beta(j,n-j+1)}_{0.05}} = \hat{F}_B^{-1}(q^{beta(j,n-j+1)}_{0.05})$.

By Lemma 0.2, we know that if we solve (7) analytically, the sequence obtained should satisfy $j_1 \ge j_2 \ge \cdots$. However, since the algorithm solves (7) approximately by using a p-increment, then we could also have $j_1 \le j_2 \le \cdots$.

Lemma 0.3 For each starting point j_1, the algorithm produces one of the three types of sequences: 1. $j_1 > j_2 \ge \cdots$ i.e. a sequence that goes down; 2. $j_1 = j_2 = \cdots$ i.e. a sequence that stays at j_1; 3. $j_1 < j_2 = \cdots$ i.e. a sequence that goes up for one step and then stays there.

Proof. The proof is technical and is given in Appendix B.

Lemma 0.3 points to the existence of a “down" and “up" sequence produced by the Iterative Algorithm. This and the argument at the beginning of this section

lead to the following proposition, supported by the examples in the next section.

Proposition 0.1 Assume that the samples size n_0 of X_0 is large enough, and that the number of fusions N is sufficiently large so that $B_{(1)} < p < B_{(N)}$. Consider the smallest $p_j \in (0,1)$ which satisfies the inequality

$$\sum_{k=0}^{j-1}\binom{N}{k}[\hat{F}_B(p_j)]^k[1-\hat{F}_B(p_j)]^{n-k} \le 0.95 \qquad (10)$$

where the p_j are evaluated along appropriate numerical increments. Then, iterating between (10) and the ordered sequence $\{B_{(j)}\}$ produces "down" and "up" sequences depending on the $B_{(j)}$ relative to p_j. In particular, in a neighborhood of the true tail probability p, with a high probability, there are "down" subsequences $\{B_{(j_i)}\}$ which converge from above and "up" subsequences $\{B_{(j_k)}\}$ which converge from below to points close to p.

A problem arises as to the choice of the p-increment. The quartiles of F_B from a very large number of fusions are known, and their order of magnitude shed light on the true p. In fact, as we shall see, the sample mean and quartiles of $B_1,...,B_N$ already give us a pretty good idea as to the value of p. Experience tells us that $mean(B)/10$ and $median(B)/10$ or similar orders of magnitude are useful choices for p-increment as we shall demonstrate in the next section.

We further note that:

The "down" and "up" sequences in the proposition are indeed the type 1 and type 3 sequences in Lemma 0.3, respectively. Furthermore, by Lemma 0.3, the Iterative Algorithm produces "down" and "up" sequences depending on the relative size of the p-increment α and the difference $B_{(j)} - \hat{p}_j$.

The proposition states that there is a shift between "down" and "up" patterns around the true p. As we shall see in the next section, the "down" and "up" patterns around the true p essentially converge to what looks like a "fixed point" close to p. Such fixed points could occur elsewhere as well.

For a very small p-increment such that $\alpha < \arg\min_j(B_{(j)} - \hat{p}_j)$, meaning, there is at least one p-increment between $\hat{p}_j$ and $B_{(j)}$ for every j, the algorithm will not produce any "up" sequence.

Algorithm Illustrations

In practice, computational constraints limit the size of the number of fusions N. Hence in (10), F_B is approximated very closely by $\hat{F}_B$ obtained with $N = 10,000$ fusions, while in the binomial coefficients we use $N = 1000$, near the maximum allowed by R. In all cases we use the misspecified tilt function $h(x) = (x, \log x)$, appropriate for gamma data, and the computer-generated data are uniform where the upper limit exceeds T: $X_1 \sim Unif(0, L)$ where $L > T$. The precision of the estimates $\hat{p}$ of p obtained at the "down-up" transition supports these choices.

The only exception is the normal case in Section 8 where $h(x) = (x, x^2)$ and $N = 5,000$. In that case $X_1 \sim Unif(-L, L)$ where $L > T$.

To observe the "down-up" pattern near the true p, each entry in the following tables is obtained from a different sample of 1000 independent B's, at times with the same j. This is to show that different samples lead to the same $\hat{p}$ observed at the shift from "down" to "up". From the following tables we see that as the "down-up" sequences approach p with any j, the number of iterations from the Iterative Algorithm decreases, a telltale sign we approach the true p.

Illustration in Terms of Lognormal Data

We start with a simulated lognormal example, with parameters $\mu = \sigma^2 = 1$, denoted by LN(1,1), where we know for sure that the tail of the distribution is far from that of a gamma tail, meaning that $h(x) = (x, \log x)$ is misspecified. We have $X_0 \sim LN(1,1)$, $\max(X_0) = 44.82807$, $T = 112.058$, $p = 0.0001$, and $X_1 \sim Unif(0,130)$. Hence $\max(X_0) < T < 130$.

The descriptive statistics of the upper bounds $B_1,...,B_{10,000}$ obtained from $N = 10,000$ fusions of X_0 with independent computer-generated samples $X_1 \sim Unif(0,130)$, where $n_0 = n_1 = 100$, are given in Table **1**. Conspicuously, the median and mean are of the same order of magnitude as that of the true p.

We chose a p-increment of 0.000015 which is of the same order of magnitude of both the $median(B)/10 = 0.00003828$ and $mean(B)/10 = 0.00005504$.

Table 1: Descriptive Statistics of $B_1,...,B_{10,000}$ from Fusions of LN(1,1) with Unif(0,130) Samples

1st Qu.	Median	Mean	3rd Qu.	Max.
.624e-04	3.828e-04	5.504e-04	7.752e-04	3.729e-03

The pattern in Table **2** points to a shift from "down" to "up" at $\hat{p} = 0.0001045544$ close to the true $p = 0.0001$ with an error of $4.5544e-06$.

Table 2: $p = 0.0001$, $X_0 \sim LN(1,1)$. $X_1 \sim Unif(0,130)$, $\max(X_0) = 44.82807$, $T = 112.058$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p -Increment 0.000015. Down-Up Shift at $\hat{p} = 0.0001045544$

Starting j	Convergence to	Iterations	
800	0.0001945544	22	Down
500	0.0001795544	9	Down
300	0.0001345544	4	Down
200	0.0001195544	1	Down
170	0.0001045544	1	Down
160	0.0001045544	1	Down
155	0.0001045544	1	Up
152	0.0001045544	1	Up
150	0.0001045544	1	Up

The next lognormal example concerns $X_0 \sim LN(0,1)$ and misspecified $h(x) = (x, \log x)$. We have $\max(X_0) = 13.77121$, $T = 41.22383$, $p = 0.0001$, and $X_1 \sim Unif(0,70)$. Hence $\max(X_0) < T < 70$.

The descriptive statistics of the upper bounds $B_1,...,B_{10,000}$ obtained from $N = 10,000$ fusions of X_0 with independent computer-generated samples $X_1 \sim Unif(0,70)$, where $n_0 = n_1 = 100$, are given in Table **3**. Conspicuously, the median and mean are of the same order of magnitude as that of the true p.

Table 3: Descriptive Statistics of $B_1,...,B_{10,000}$ from Fusions of LN(0,1) with Unif(0,70) Samples

1st Qu.	Median	Mean	3rd Qu.	Max.
.972e-05	2.012e-04	4.287e-04	5.759e-04	4.288e-03

We chose a p-increment of 0.000015 which is of the same order of magnitude of both the $median(B)/10 = 0.00002012$ and $mean(B)/10 = 0.00004287$.

The pattern in Table **4** points to a shift from "down" to "up" at $\hat{p} = 0.0001042241$ close to the true $p = 0.0001$ with an error of $4.2241e-06$.

Table 4: $p = 0.0001$, $X_0 \sim LN(0,1)$. $X_1 \sim Unif(0,70)$, $\max(X_0) = 13.77121$, $T = 41.22383$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p -Increment 0.000015. Down-Up Shift at $\hat{p} = 0.0001042241$

Starting j	Convergence to	Iterations	
900	0.0002392241	27	Down
800	0.0001042241	24	Down
700	0.0001042241	17	Down
420	0.0001042241	3	Down
370	0.0001042241	1	Down
360	0.0001042241	1	Down
355	0.0001042241	1	Up
350	0.0001042241	1	Up
350	0.0001042241	1	Up

Illustration in Terms of Mercury Data

The present illustration concerns levels of mercury data measured in marine life in mg/kg. The data source is NOAA's National Status and Trends Data

https://products.coastalscience.noaa.gov/nsandt_data/data.aspx.

The mercury data consists of 8,266 observations of which 9 observations exceed $T = 22.41$ giving the proportion of $p = 0.001088797$. We treat the mercury data as a population and draw a reference random sample without replacement X_0 of size $n_0 = 200$ from it.

The results of 10,000 fusions of the mercury reference sample X_0 with $X_1 \sim Unif(0,50)$ samples of size $n_1 = 200$ gave 10,000 B upper bounds. Their descriptive statistics are summarized in Table **5**.

Table 5: Descriptive Statistics from 10,000 Mercury B Upper Bounds

1st Qu.	Median	Mean	3rd Qu.	Max.
.098e-03	3.319e-03	3.475e-03	4.715e-03	8.929e-03

Observe that the 1st quartile, median, mean, and 3rd quartile are all of the same order of magnitude as that of the true p, and hence they give us an idea as to the value of p. Here $\max(X_0) = 13.8 < T = 22.41 < 50$.

One-tenth of both the median and the mean suggest a p-increment of 0.0001. That is, an increment of the order of $mean(B)/10$ and $median(B)/10$.

The pattern in Table **6** points to a shift from "down" to "up" at $\hat{p} = 0.001092137$ close to the true $p = 0.001088797$ with an error of $3.34e-06$.

Table 6: $p = 0.001088797$, X_0 a Mercury Sample. $X_1 \sim Unif(0,50)$, $\max(X_0) = 13.8$, $T = 22.41$, $n_0 = n_1 = 200$, $h = (x, \log x)$, p -Increment 0.0001. Down-Up Shift at $\hat{p} = 0.001092137$

Starting j	Convergence to	Iterations	
775	0.002792137	18	Down
600	0.000999351	16	Down
300	0.001492137	9	Down
200	0.001192137	7	Down
100	0.001192137	1	Down
90	0.001192137	1	Up
85	0.001092137	1	Down
84	0.001092137	1	Up
83	0.001092137	1	Up
81	0.001092137	1	Up
80	0.001092137	1	Up

Illustration in Terms of Radon Data

Residential radon is a tasteless, colorless and odorless radioactive gas naturally abundant in the soil. Approximately 40 percent of Pennsylvania homes have radon levels above the EPA action guideline of 4 pCi/L.

The iterative algorithm is applied here to Beaver County radon data from 1989 to 2017. The data consist of 7,425 radon observations, taken as a population, of which only 2 exceed 200. Hence, with $T = 200$ we wish to estimate the small probability $p = 2/7425 = 0.0002693603$. The reference sample X_0 was chosen without replacement from the 7,425 radon observations. The generated X_1 samples are from $Unif(0,300)$ and $n_0 = n_1 = 500$. We observe that $\max(X_0) = 107 < T = 200$, so that the largest data point is close to $T/2$.

The results of 10,000 fusions of a radon reference sample X_0 of size $n_0 = 500$ with $X_1 \sim Unif(0,300)$ samples of size $n_1 = 500$ gave 10,000 B upper bounds. Their descriptive statistics are summarized in Table **7**.

Table 7: Descriptive Statistics from 10,000 Mercury B Upper Bounds

1st Qu.	Median	Mean	3rd Qu.	Max.
.175e-04	1.806e-04	2.077e-04	2.686e-04	1.077e-03

Observe that the 3rd quartile of 0.0002686 is very close to true $p = 0.0002694$. The p-increment was chosen as 0.00003, which is of the same order of magnitude as one tenth of either the 1st quartile, mean, median, or 3rd quartile of the 10,000 B's.

From Table **8**, the down-up shift occurs at $\hat{p} = 0.0002689389$ very close to the true $p = 0.0002693603$ giving an error of $4.611e-07$.

Changing to a p-increment of 0.000018, close to $median(B)/10 = 0.00001806$, we get the exact same $\hat{p} = 0.0002689389$, whereas a p-increment of 0.00002686 which is equal to one-tenth of the 3rd quartile gives $\hat{p} = 0.0002675389$ with an error of $1.8611e-06$.

Table 8: $p = 0.0002693603$, X_0 a Radon Sample. $X_1 \sim Unif(0,300)$, $\max(X_0) = 107$, $T = 200$, $n_0 = n_1 = 500$, $h = (x, \log x)$, p-Increment 0.00003. Down-Up Shift at $\hat{p} = 0.0002689389$

Starting j	Convergence to	Iterations	
1000	0.0005389389	6	Down
800	0.0002989389	1	Down
773	0.0002689389	1	Down
762	0.0002689389	1	Down
750	0.0002689389	1	Down
749	0.0002689389	1	Down
748	0.0002689389	1	Up
745	0.0002689389	1	Up
743	0.0002689389	1	Up
740	0.0002689389	1	Up
739	0.0002689389	1	Up
730	0.0002689389	1	Up

A different approach to residential radon exceedances using fused county data is studied in Zhang *et al*. (2020a,b) [11, 12], employing a density ratio model with variable tilt functions.

Illustration in Terms of Normal Data

Throughout the paper we deal with tail probabilities of non-negative continuous data where we use $h(x)=(x,\log x)$. However, the Iterative Algorithm is applicable in more general situations, including the situation when the data are normal, provided the tilt function is chosen judicially. To underscore this, we bring next an example with normal sample X_0, in which case the "normal tilt" $h(x)=(x,x^2)$ is specified when X_1 is a normal sample as well, but nearly specified when X_1 is uniform over a sufficiently large support.

Accordingly, consider $X_0 \sim N(0,4)$, $T=6.180465$, $p=0.001$, $\max(X_0)=4.168643$, $X_1 \sim Unif(-10,10)$, and $n_0=n_1=100$. From 5,000 fusions we have $median(B)=0.0037089$ and $mean(B)=0.0039251$ (same order of magnitude as that of p), one tenth of which is on the order of the used p-increment=0.0001. The down-up results are given in Table **9**, showing a shift at $\hat{p}=0.001008915$, giving an error of 8.915e-06, on par with the previous results.

Table 9: p **=0.001.** $X_0 \sim N(0,4)$, $X_1 \sim Unif(-10,10)$, $\max(X_0)=4.168643$, $T=6.180465$, $n_0=n_1=100$, $h=(x,x2)$, **p-Increment** 0.0001**. Down-Up Shift at** $\hat{p}=0.001008915$

Starting j	Convergence to	Iterations	
77	0.001008915	9	Down
50	0.001008915	7	Down
22	0.001108915	5	Down
14	0.001008915	3	Down
10	0.001008915	2	Down
9	0.001008915	1	Down
8	0.001008915	1	Up
7	0.001008915	1	Up
6	0.001008915	2	Up
	0.001008915	2	Up

Results Summary

To get a picture of of the repeated fusion method in tail estimation, the top four entries in Table **10** depict the pairs $(p,\hat{p})$ obtained from the previous examples with a misspecified $h(x)=(x,\log x)$. The next four entries are from nearly specified cases where $h(x)=(x,\log x)$ is sensible, and where the increment was chosen as before. There is no apparent difference from the four misspecified cases. In the Weibul(1.1,1) case the down-up shift alternated between 0.0001051111 and 0.0001201111 and we report the average. Similar results can be found in Kedem *et al*. (2019) and Kedem and Pyne (2021) [5, 6].

Again, already the median and mean of $B_1,...,B_{10,000}$ ($B_1,...,B_{5,000}$ in the last entry) give a good idea as to the value of p, however, the Iterative Algorithm improves greatly on the mean and median as we see from Table **10** where in all cases $\hat{p}$ is close to p. We have seen that in many other cases.

Comparison with POT

Possibly, an alternative method to repeated fusion is the extreme value theory method of peaks over threshold (POT), where only the values above a sufficiently high threshold are used in the estimation of small tail probabilities [13, 14]. This results in a reduced sample which could prove to be problematic when the original sample is too small to begin with. By contrast, with the repeated fusion method the total number of observations, albeit some of which are artificial, increases. A brief comparison between these two methods is given in Wang (2018) and Kedem *et al*. (2019) [5, 9]. It shows that, across quite a few distributions, for sample sizes on the order of 100 and $p=0.001$, the repeated fusion method tends to give higher coverage and smaller mean absolute error. However, it must be emphasized that this could be reversed for much smaller probabilities and much larger samples. We shall look into this problem elsewhere.

RESULTS

Nitrogen Dioxide Data Analysis

Finally, we apply next the Iterative Algorithm to nitrogen dioxide (NO_2) data from Washington DC, where, as before we use 10,000 fusions giving upper bounds $B_1,...,B_{10,1000}$ for p, the chance that NO_2 exceeds T. The algorithm was applied for $T=100$, where $\max(X_0)=48.06<T/2$. As in the previous rather precise computational results, the down-up shift point is the point estimate $\hat{p}$.

Our repeated fusion analysis was applied to a random sample X_0 of hourly measurements of size $n_0=400$. The generated X_1 samples were from $Unif(0,150)$, $h(x)=(x,\log x)$, and $n_0=n_1=400$. We have $\max(X_0)=48.06<T=100<150$, so that the largest data point is close to $T/2$. The results of

Table 10: Comparison between p and $\hat{p}$ Obtained from 10,000 Fusions. $X_1 \sim Unif(0,L)$. The Last Entry was Obtained from 5,000 Fusions and $X_1 \sim Unif(-10,10)$. Some Figures are Rounded

X_0	T	max(X_0)	L	$n_0 = n_1$	Median(B)	p	$\hat{p}$	Error
LN(1,1)	112.06	44.83	130	100	0.0003828	0.0001000	0.0001046	4.6e-06
LN(0,1)	41.22	13.77	70	100	0.0002012	0.0001000	0.0001042	4.2e-06
Radon	200.00	107.00	300	500	0.0001806	0.0002694	0.0002689	5.0e-07
Mercury	22.41	13.80	50	200	0.0033190	0.0010888	0.0010921	3.3e-06
F(1,18)	15.38	7.27	20	100	0.0022574	0.0010000	0.0010513	5.1e-05
Gamma(1,0.05)	184.21	77.62	210	100	0.0001582	0.0001000	0.0001040	4.0e-06
Weibul(1.1,1)	7.53	2.84	12	100	0.0007227	0.0001000	0.0001126	1.3e-05
Weibul(0.8,1)	16.05	5.80	22	100	0.0006309	0.0001000	0.0001049	4.9e-06
N(0,4)	6.18	4.17	10	100	0.0037089	0.0010000	0.0010089	8.9e-06

10,000 fusions of the NO_2 reference sample X_0 with $X_1 \sim Unif(0,150)$ samples of size $n_1 = 400$ gave 10,000 B upper bounds. Their descriptive statistics are summarized in Table **11**.

Table 11: Descriptive Statistics from 10,000 NO_2 B Upper Bounds. $T = 100$

1st Qu.	Median	Mean	3rd Qu.	Max.
.209e-05	9.860e-05	1.039e-04	1.293e-04	5.821e-04

From Table **12**, with $mean(B)/10 = 0.00001039$ as the p-increment, the down-up shift occurs at $\hat{p} = 0.0001344551$ not far from the $3rd\ Quartile(B) = 0.0001293$.

To approximate the variance of $\hat{p}$, we can redo the analysis over and over again. Thus, we have repeated the above analysis ten times, each time with different 10,000 fusions. The resulting $\hat{p}$'s from different reference samples X_0 of sizes $n_0 = 400$, using different p-increments equal to $mean(B)/10$, are given in Table **13**. The exception is an increment of $3rd\ Qu(B)/10$ in entry 9.

The variation among the $\hat{p}$'s is fairly small pointing to consistent results. In fact the sample mean and standard deviation of the $\hat{p}$'s are 0.0001306905 and 1.411819e-05, respectively.

We see that in all cases the $median(B)$ and $mean(B)$ are close to 0.0001, which is not far from the resulting $\hat{p}$'s.

CONCLUSION

The repeated fusion of a given moderately large sample with numerous computer-generated samples

Table 12: X_0 from NO_2, $X_1 \sim Unif(0,150)$, $\max(X_0) = 48.06$, $T = 100$, $n_0 = n_1 = 400$, $h = (x, \log x)$, p-Increment 0.00001039. Down-Up Shift at $\hat{p} = 0.0001344551$

Starting j	Convergence to	Iterations	
999	0.0002175751	6	Down
990	0.0001760151	7	Down
950	0.0001760151	2	Down
910	0.0001656251	1	Down
850	0.0001448451	1	Down
830	0.0001448451	1	Down
800	0.0001344551	1	Down
780	0.0001344551	1	Down
775	0.0001344551	1	Down
774	0.0001344551	1	Down
773	0.0001344551	1	Up
771	0.0001344551	1	Up
770	0.0001344551	1	Up
769	0.0001344551	1	Up
768	0.0001344551	1	Up
767	0.0001344551	1	Up
766	0.0001344551	1	Up
765	0.0001344551	1	Up
750	0.0001344551	1	Up
700	0.0001240651	1	Up

Table 13: Estimates $\hat{p}$ of the Probability that NO_2 in Washington DC Exceeds 100 ppb Obtained from Different Samples. p -Increment is $mean(B)/10$, Except in Number 9 where p-Increment is $3rd\ Qu(B)/10 = 0.00001220.$

Case	1st Qu(B)	Median(B)	Mean(B)	3rd Qu(B)	$\hat{p}$
1.	5.019e-05	7.064e-05	7.459e-05	9.536e-05	0.0001180567
2.	7.979e-05	1.060e-04	1.130e-04	1.403e-04	0.0001464964
3.	5.364e-05	7.645e-05	8.022e-05	1.020e-04	0.0001035439
4.	7.809e-05	1.048e-04	1.106e-04	1.364e-04	0.0001321030
5.	7.385e-05	1.005e-04	1.058e-04	1.307e-04	0.0001264732
6.	8.448e-05	1.122e-04	1.194e-04	1.470e-04	0.0001427179
7.	6.125e-05	8.044e-05	8.408e-05	1.027e-04	0.0001002384
8.	6.744e-05	9.260e-05	9.752e-05	1.220e-04	0.0001262218
9.	6.744e-05	9.260e-05	9.752e-05	1.220e-04	0.0001336458
10.	7.209e-05	9.860e-05	1.039e-04	1.293e-04	0.0001344551

produces many upper bounds for tail or threshold probabilities. The descriptive statistics of the upper bounds already by themselves provide useful information about the probabilities. In fact, reasonably close approximations. The ensued iterative method takes the ordered upper bounds as input to produce estimates of small threshold probabilities, which take values in small neighborhoods of the true threshold probability. The precision obtained by the iterative method is encouraging if not somewhat surprising, as has been observed numerous times with different types of data exhibiting very different probability distributions. Variability of the estimates can be obtained from different sets of, say, 10,000 fusions.

Some words of caution are in order. First, like with any other statistical method, the sample size is important. Thus, in the NO_2 analysis, the excessively high threshold of $T = 200$ proved problematic, which most likely could be overcome with a larger sample. Second, we do not have any guidelines as to the size of the support of the uniform fusion samples, except to say that the support must contain T , as was done in our data analysis.

Additional references about fusion and repeated fusion of a sample with independent external data are Fokianos and Qin (2008), Katzoff *et al*. (2014), Kedem *et al*. (2016), and Zhou (2013) [8, 15, 16, 17].

APPENDIX

Appendix A: Density Ratio Model for Multiple Sources

The appendix addresses the density ratio model (2) for $m+1$ data sources. Thus, we deal with the density ratio model more generally where X_0 is fused with m computer-generated samples. Above we dealt with the special case of $m = 1$.

Assume that the reference random sample X_0 of size n_0 follows an unknown reference distribution with probability density g, and let G be the corresponding cumulative distribution function (cdf).

Let

$$X_1,..., X_m,$$

be additional computer-generated random samples where $X_j \sim g_j, G_j$, with size n_j, $j = 1,...,m$. The augmentation of $m+1$ samples

$$t = (t_1,\ldots,t_n) = (X_0, X_1,\ldots, X_m), \tag{11}$$

of size $n_0 + n_1 + \cdots + n_m$ gives the fused data. The density ratio model stipulates that

$$\frac{g_j(x)}{g(x)} = \exp(\alpha_j + \beta_j' h(x)), \quad j = 1,\ldots,m, \tag{12}$$

where β_j is an $r \times 1$ parameter vector, α_j is a scalar parameter, and $h(x)$ is an $r \times 1$ vector valued distortion or tilt function. None of the probability densities $g, g_1,...,g_m$ and the corresponding G_j's, and none of the parameters α's and β's are assumed known, but, strictly speaking, the so-called tilt function h must be a known function.

Asymptotic Distribution of $\hat{G}(x)$

Define $\alpha_0 \equiv 0, \beta_0 \equiv 0$, $w_j(x) = \exp(\alpha_j + \beta_j' h(x))$, $\rho_i = n_i / n_0$, $j = 1,\ldots,m$.

Maximum likelihood estimates for all the parameters and $G(x)$ can be obtained by maximizing the empirical likelihood over the class of step cumulative distribution functions with jumps at the observed values $t_1,\ldots,t_n$ [18]. Let $p_i = dG(t_i)$ be the mass at t_i, for $i = 1,\ldots,n$. Then the empirical likelihood becomes

$$\ell(\theta, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\alpha_1 + \beta_1' h(x_{1j})) \cdots \prod_{j=1}^{n_m} \exp(\alpha_m + \beta_m' h(x_{mj})). \quad (13)$$

Maximizing $\ell(\theta, G)$ subject to the constraints

$$\sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i[w_1(t_i) - 1] = 0, \ldots, \sum_{i=1}^{n} p_i[w_m(t_i) - 1] = 0 \quad (14)$$

we obtain the desired estimates. In particular,

$$\hat{G}(t) = \frac{1}{n_0} \cdot \sum_{i=1}^{n} \frac{I(t_i \le t)}{1 + \rho_1 \exp(\hat{\alpha}_1 + \hat{\beta}_1' h(t_i)) + \cdots + \rho_m \exp(\hat{\alpha}_m + \hat{\beta}_m' h(t_i))}, \quad (15)$$

where $I(t_i \le t)$ equals one for $t_i \le t$ and is zero, otherwise. Similarly, $\hat{G}_j$ is estimated by summing $\exp(\hat{\alpha}_j + \hat{\beta}_j' h(t_i)) dG(t_i)$.

The asymptotic properties of the estimators have been studied by a number of authors including Qin and Zhang (1997), Zhang (2000), and Lu (2007) [7, 19, 20].

Define the following quantities: $\rho = diag\{\rho_1,\ldots,\rho_m\}$,

$$A_j(t) = \int \frac{w_j(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y),$$

$$B_j(t) = \int \frac{w_j(y) h(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y),$$

$$\bar{A}(t) = (A_1(t),\ldots,A_m(t))', \quad \bar{B}(t) = (B_1'(t),\ldots,B_m'(t))'.$$

Then the asymptotic distribution of $\hat{G}(t)$ for $m \ge 1$ is given by the following result due to Lu (2007) [20].

Theorem 0.2 Assume that the sample size ratios $\rho_j = n_j / n_0$ are positive and finite and remain fixed as the total sample size $n = \sum_{j=0}^{m} n_j \to \infty$. The process $\sqrt{n}(\hat{G}(t) - G(t))$ converges to a zero-mean Gaussian process in the space of real right continuous functions that have left limits with covariance matrix given by

$$Cov\{\sqrt{n}(\hat{G}(t) - G(t)), \sqrt{n}(\hat{G}(s) - G(s))\} = \left(\sum_{k=0}^{m} \rho_k\right)(G(t \wedge s) - G(t)G(s) - \sum_{j=1}^{m} \rho_j A_j(t \wedge s)) + (\bar{A}'(s)\ \rho, \bar{B}'(s)(\ \rho \otimes I_p)) S^{-1} \begin{pmatrix} \rho \bar{A}(t) \\ (\ \rho \otimes I_p)\bar{B}(t) \end{pmatrix}. \quad (16)$$

where I_p is the $p \times p$ identity matrix, and $\otimes$ denotes Kronecker product.

For a complete proof see Lu (2007) [20]. The proof for $m = 1$ is given in Zhang (2000) [19].

Denote by $\hat{V}(t)$ the estimated variance of $\hat{G}(t)$ as given in (16). Replacing parameters by their estimates, a $1-\alpha$ level pointwise confidence interval for $G(t)$ is approximated by

$$\left(\hat{G}(t) - z_{\alpha/2}\sqrt{\hat{V}(t)},\ \hat{G}(t) + z_{\alpha/2}\sqrt{\hat{V}(t)}\right), \quad (17)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution. Hence, a $1-\alpha$ level pointwise confidence interval for $1 - G(T)$ for any T, and in particular for relatively large thresholds T is approximated by

$$(1 - \hat{G}(t) - z_{\alpha/2}\sqrt{\hat{V}(t)}, 1 - \hat{G}(t) + z_{\alpha/2}\sqrt{\hat{V}(t)}). \quad (18)$$

Appendix B: Proof of Lemma 0.3

To prove that the three types of sequences specified by Lemma 0.3 are the only ones that we obtain from the algorithm, we first show how the p-increment determines the relationship between j_1 and j_2.

By Lemma 0.2, $B_{(j_1)} \ge \hat{p}_{j_1}$ almost surely. If there is a p-increment $k\alpha$ such that $\hat{p}_{j_1} \le k\alpha < B_{(j_1)}$, then the approximated solution that the algorithm provides is $k\alpha$ since it is the smallest p-increment that satisfies (7). If there are more than one p-increment between $\hat{p}_{j_1}$ and $B_{(j_1)}$, then the algorithm will give the smallest one among them and we can denote this as $k\alpha$.

The next step of the algorithm is to find the smallest $B_{(j)} \ge k\alpha$. If we have a $B_{(j_2)}$ such that $k\alpha \le B_{(j_2)} < B_{(j_1)}$, then the sequence will go down from

$B_{(j_1)}$ to $B_{(j_2)}$. Now it remains to show that $B_{(j_3)} \le B_{(j_2)}$ to prove that the algorithm can produce the type 1 sequence i.e. $j_1 > j_2 \ge \cdots$. Recall the expression $\hat{p}_j = q^{\hat{F}_B}_{beta(j,n-j+1)}$ and this is non-decreasing in j. Thus, we must have $\hat{p}_{j_2} \le \hat{p}_{j_1}$.

Note that we have $k\alpha$ between $\hat{p}_{j_2}$ and $B_{(j_2)}$ which means that there is at least one p-increment between $\hat{p}_{j_2}$ and $B_{(j_2)}$ so that the sequence will either go down or stay at j_2. To sum up, *if there exists a p-increment $k\alpha$ such that $\hat{p}_{j_1} \le k\alpha < B_{(j_1)}$ and a $B_{(j_2)}$ such that $k\alpha \le B_{(j_2)} < B_{(j_1)}$, then the algorithm will give type 1 seqence.*

Now we shall examine the circumstances under which the algorithm produces type 2 sequence i.e. the sequence stays at j_1. There are two cases: 1. There is no p-increment between $\hat{p}_{j_1}$ and $B_{(j_1)}$ but there is a p-increment $k\alpha = B_{(j_1)}$. In this case, the algorithm will give $k\alpha$ as the solution and then the smallest $B_{(j)} \ge k\alpha$ is still $B_{(j_1)}$; 2. There exists at least one (if more than one then we take the smallest one among them) p-increment $k\alpha$ between $\hat{p}_{j_1}$ and $B_{(j_1)}$ but there is no $B_{(j_2)}$ between $k\alpha$ and $B_{(j_1)}$. In this case, the smallest $B_{(j)} \ge k\alpha$ is also $B_{(j_1)}$.

Based on the analysis above, it is readily seen that if there is no p-increment $k\alpha$ such that $\hat{p}_{j_1} \le k\alpha \le B_{(j_1)}$, then the algorithm gives type 3 sequence i.e. $j_1 < j_2 = \cdots$.

It remains to show that the sequence stays at j_2 after one-step going up from j_1 to j_2. Suppose that $k\alpha$ is the smallest p-increment such that $k\alpha \ge B_{(j_1)}$ and $B_{(j_2)}$ is the smallest $B_{(j)} \ge k\alpha$. Note that $P(B_{(j)} > p)$ is non-increasing in j so that if $k\alpha$ satisfies the inequality (7) for j_1 then it must satisfy it for j_2 since $j_2 > j_1$. Therefore, the solution of p is also $k\alpha$ for j_2 so that the sequence stays at j_2. At this point, we have completed the proof of Lemma 0.3.

ACKNOWLEDGEMENT

The work of B. Kedem and X. Zhang has been supported by a Faculty-Student Research Award, University of Maryland, College Park. B. Kedem dedicates this article to Gerald R. North author of *The Rise of Climate Science*.

REFERENCES

[1] Eckel SP, Cockburn M, Shu Y-H, Deng H, Lurmann FW, Liu L, Gilliland FD. Air pollution affects lung cancer survival. Thorax 2016; 71: 891-898. https://doi.org/10.1136/thoraxjnl-2015-207927

[2] Faustini A, Rapp R, Forastiere F. Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. European Respiratory Journal 2014; 44: 744-753. https://doi.org/10.1183/09031936.00114713

[3] Hamra GB, Laden F, Cohen AJ, Raaschou-Nielsen O, Brauer M, Loomis D. Lung cancer and exposure to nitrogen dioxide and traffic: a systematic review and meta-analysis. Environmental Health Perspectives 2015; 123: 1107-1112. https://doi.org/10.1289/ehp.1408882

[4] Kedem, Benjamin, Victor De Oliveira, and and Michael Sverchkov. Statistical Data Fusion. Singapore: World Scientific 2017. https://doi.org/10.1142/10282

[5] Kedem B, Pan L, Smith P, Wang C. Estimation of Small Tail Probabilities by Repeated Fusion. Mathematics and Statistics 2019; 7: 172-181. https://doi.org/10.13189/ms.2019.070503

[6] Kedem B, Pyne S. Estimation of Tail Probabilities by Repeated Augmented Reality. Journal of Statistical Theory and Practice 2021; 15. https://doi.org/10.1007/s42519-020-00152-1

[7] Qin J, Zhang B. A Goodness of Fit Test for Logistic Regression Models Based on Case-control Data. Biometrika 1997; 84: 609-618. https://doi.org/10.1093/biomet/84.3.609

[8] Kedem B, Pan L, Zhou W, Coelho CA. Interval Estimation of Small Tail Probabilities – Application in Food Safety. Statistics in Medicine 2016; 35: 3229-3240. https://doi.org/10.1002/sim.6921

[9] Wang, Chen. Data Fusion Based on the Density Ratio Model. PhD dissertation, Department of Mathematics, University of Maryland, College Park 2018.

[10] Casella, George and Roger L. Berger. Statistical Inference, 2nd ed. Pacific Grove, CA: Duxbury 2002.

[11] Zhang X, Pyne S, Kedem B. Estimation of Residential Radon Concentration in Pennsylvania Counties by Data Fusion. Applied Stochastic Models in Business and Industry 2020a; 36: 1094-1110. https://doi.org/10.1002/asmb.2546

[12] Zhang X, Pyne S, Kedem B. Model Selection in Radon Data Fusion. Statistics in Transition, new series, 2020b; 21: 159-165. https://doi.org/10.21307/stattrans-2020-036

[13] Beirlant, Jan, Yuri Goegebeur, Jozef Teugels, and Johan Segers. Statistics of Extremes : Theory and Applications. Hoboken, NJ: Wiley 2004. https://doi.org/10.1002/0470012382

[14] Ferreira A, De Haan L. On the Block Maxima Method in Extreme Value Theory: PWM Estimators. The Annals of Statistics 2015; 43: 276-298. https://doi.org/10.1214/14-AOS1280

[15] Fokianos K, Qin J. A Note on Monte Carlo Maximization by the Density Ratio Model. Journal of Statistical Theory and Practice 2008; 2: 355-367. https://doi.org/10.1080/15598608.2008.10411880

[16] Katzoff M, Zhou W, Khan D, Lu G, Kedem B. Out of Sample Fusion in Risk Prediction. Journal of Statistical Theory and Practice 2014; 8: 444-459. https://doi.org/10.1080/15598608.2013.806233

[17] Zhou, Wen. Out of Sample Fusion. PhD dissertation, Department of Mathematics, University of Maryland, College Park 2013.

[18] Owen, Art. Empirical Likelihood. Boca Raton, FL: Chapman & Hall/CRC 2001.

[19] Zhang B. A Goodness of Fit Test for Multiplicative-intercept Risk Models Based on Case-control Data. Statistica Sinica 2000; 10: 839-865.

[20] Lu, Guanhua. Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Model and its Application to Mortality Forecasting. PhD dissertation, Department of Mathematics, University of Maryland, College Park 2007.

Received on 18-04-2021 Accepted on 01-06-2021 Published on 09-06-2021

https://doi.org/10.6000/1929-6029.2021.10.07