*STATISTICS IN TRANSITION new series, August 2020*
*Special Issue, pp. 163–169, DOI 10.21307/stattrans-2020-036*
Submitted – 31.01.2020; paper accepted for published – 30.06.2020

# Model selection in radon data fusion

**Xuze Zhang**[1], **Saumyadipta Pyne**[2], **Benjamin Kedem**[3]

## ABSTRACT

Fitting parametric models or the use of the empirical cumulative distribution function are problematic when it comes to the estimation of tail probabilities from small samples. A possible remedy is to fuse or combine the small samples with additional data from external sources and base the inference on the so called density ratio model with variable tilt functions, which widens the support of the estimated distribution of interest. This approach is illustrated using residential radon concentration data collected from western Pennsylvania.

**Key words:** Tail probabilities, density ratio model, variable tilt functions, Appalachian Plateau, Forest County, Pennsylvania.

## 1. Introduction

In general, the estimation of tail probabilities requires large samples. However, in many cases the available samples are relatively small, a problem which can be overcome to a reasonable extent by fusing the available data from several independent sources. This is illustrated here using residential radon concentration data collected from counties in western Pennsylvania (PA). We used county-level indoor radon concentrations based on records collected by the Pennsylvania Department of Environmental Protection (PA DEP), Bureau of Radiation Protection, Radon Division. For more details about the data see Zhang, Pyne, and Kedem (ZPK) (2019), and the appropriate references including PA Department of Environmental Protection, Rack-Amber (2013), Wikipedia contributors (2019).

The range of values of a small sample may not be large enough to shed light on the tail behavior of the distribution which gave rise to the sample. In that case more data are needed. However, in many cases, more data are not available. Our goal is to demonstrate that the problem can be ameliorated to a reasonable extent when the sample is fused or combined with data from other sources, as the range of values of the combined data is larger. Technically, this can be achieved by appealing to the so called *density ratio model* (DRM), where the distributions of the various sources are connected by fixed *tilt functions*. The novelty of the paper is the use of *variable tilts* obtained by model selection.

[1]Department of Mathematics and Institute for Systems Research, University of Maryland, College Park. USA. E-mail: xzhang51@umd.edu. ORCID: https://orcid.org/0000-0002-8672-8515.

[2]Public Health Dynamics Laboratory, and Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh. USA. E-mail: spyne@pitt.edu. ORCID: https://orcid.org/0000-0003-3470-2345.

[3]Department of Mathematics and Institute for Systems Research, University of Maryland, College Park. USA. E-mail: bnk@umd.edu. ORCID: https://orcid.org/0000-0001-8720-3465.

In this paper, we apply a data fusion method in the estimation of residential radon levels in Forest County, located in the Appalachian Plateau in western PA. Its population is small, with 2000 households as per the 2010 census, yielding a small sample of 47 homes only, insufficient for the estimation of tail probabilities, and hence qualifying it as a "small area" problem. To overcome the small sample size, we fuse the Forest data with samples from the two adjacent counties Elk and Warren whose populations are much larger. Tail probabilities can then be estimated by using the density ratio model (DRM) with *variable tilt functions* ZPK (2019). This formulation requires the selection of optimal models out of a large number of models. In ZPK (2019), the selection of tilt function was done via a long process of hypothesis testing while here we use a more efficient model selection advocated in Fokianos (2007). The DRM is discussed in detail in Kedem, De Oliveira and Sverchkov (KDS) (2017) and Qin (2017).

Fusing data from Forest, Elk, and Warren counties is sensible as they share the geographical features of the "High Plateau Section" in northwestern PA in the region of Appalachian Plateau (Rack-Amber 2013, Wikipedia contributors 2019).

Radon is an odorless cancer-causing radioactive gas released from decaying uranium, thorium and radium in rocks and soil, and is the cause of thousands of deaths each year (Rack-Amber 2013). Approximately 40% of PA homes have radon levels exceeding EPA's action guideline of 4 picocuries (pCi) per liter (PA Department of Environmental Protection).

Therefore, it is of great importance to public health and policy that the residential radon exposure data be analyzed to produce robust tail or exceedance probabilities.

The organization of the paper is as follows. Section 2 deals with the semi-parametric estimation of the parameters and the probability densities of the density ratio model. It also addresses the selection of the tilt functions. A case in point in terms of residential radon is discussed in Section 3. A summary is provided in Section 4.

## 2. Methodology

### 2.1. Density Ratio Model

To make use of the data from neighboring counties, a multi-sample DRM is proposed to fuse the data from the county of interest and its $m$ neighbors such that

$$\frac{g_k(x)}{g(x)} = \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(x)\right) \qquad k = 1, \ldots, m \tag{1}$$

where $g$ represents the density of residential radon levels of the county of interest and $g_1, \ldots, g_m$ represent the densities of its $m$ neighbors.

The semi-parametric estimation of the parameters and densities in (1) is discussed in the next section using the empirical likelihood (Owen 2001). Model (1) was found adequate by a graphical goodness of fit test discussed briefly in Section 3. The model is discussed extensively in the recent books by KDS (2017) and in Qin (2017), which also describe quite a few applications from case-control tests of equidistribution to time series prediction.

Instead of making parametric assumptions on these densities, we propose a parametric structure of their ratios by DRM (KDS 2017, Qin 2017). A proper choice of the tilt functions $\boldsymbol{h}_k$'s is imperative since misspecification of the tilt functions leads to bias, large standard errors, and power loss (Fokianos and Kaimi 2006). We shall commence with a possibly redundant or "global" tilt and then select a reduced form of this tilt. Such a tilt function is specified in section 3.

### 2.2. Estimation and Asymptotic Result

Let $\boldsymbol{X}_0, \ldots, \boldsymbol{X}_m$ be the samples from the county of interest and its $m$ neighbors with sample sizes $n_0, \ldots, n_m$, respectively. The sample $\boldsymbol{X}_0$ is referred to as the reference sample and we shall denote by $G$ the corresponding reference cumulative distribution function (CDF). The fused sample is defined as $\boldsymbol{t} = (\boldsymbol{X}_0^T, \ldots, \boldsymbol{X}_m^T)^T$, with size $n = \sum_{k=0}^{m} n_k$.

Inference can be based on the following empirical likelihood obtained from the fused sample $\boldsymbol{t}$:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) = \prod_{i=1}^{n} p_i \prod_{k=1}^{m} \prod_{j=1}^{n_k} \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(X_{kj})\right) \tag{2}$$

where $p_i = dG(t_i)$ and the estimates $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\beta}}$ and hence the $\tilde{p}_i$'s, are obtained by maximizing (2) with constraints

$$\sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} p_i \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(t_i)\right) = 1 \qquad k = 1, \ldots, m. \tag{3}$$

Subsequently, we obtain the estimated reference CDF $\tilde{G}(t) = \sum_{i=1}^{n} \tilde{p}_i I[t_i \le t]$ and the asymptotic result

$$\sqrt{n}(\tilde{G}(t) - G(t)) \xrightarrow{d} N(0, \sigma(t)), \qquad \text{as } n \to \infty. \tag{4}$$

The expression of $\sigma(t)$ and other details regarding estimation and asymptotic result can be found in KDS (2017), Qin (2017) and ZPK (2019). Therefore, we can construct a 95% confidence interval of the tail probability $1 - G(T)$ for a given threshold $T$ based on (4)

$$(1 - \tilde{G}(T) - z_{0.025}\sqrt{\frac{\tilde{\sigma}(t)}{n}}, 1 - \tilde{G}(T) + z_{0.025}\sqrt{\frac{\tilde{\sigma}(t)}{n}}). \tag{5}$$

### 2.3. Model Selection

As mentioned in 2.1, we aim to select tilt functions that can better specify the density ratio structure. Such selection can be made based on the AIC criterion given by

$$-2\log L(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{G}) + 2q \tag{6}$$

where $q$ is the number of free parameters in the model (Fokianos 2007). Note that the number of free parameters is equal to the number of $\beta$'s due to the constraints (3).

## 3. Illustrative Example: Forest County Radon Data Fusion

Here Forest county is the county of interest. Denote the Forest sample by $\boldsymbol{X}_0$ and its size by $n_0$. The sample size $n_0 = 47$ is relatively small so that the empirical estimate of the CDF $\hat{G}(t) = \frac{1}{n_0}\sum_{i=1}^{n_0} I[X_{0i} \le t]$ may not be satisfactory for the estimation of tail probabilities. That is, we cannot make inference about $G$ based on $\hat{G}$ outside of the range of $\boldsymbol{X}_0$. Also, a smaller sample size leads to higher standard errors and hence wider confidence intervals, and may not be adequate for the estimation of small tail probabilities.

We wish to mitigate these issues by fusing $\boldsymbol{X}_0$ with samples from its two neighboring counties Warren and Elk to obtain an estimate of the reference CDF $\tilde{G}$ based on the DRM (1). The samples from Warren and Elk are denoted as $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. The corresponding sample sizes are $n_1 = 837$ and $n_2 = 1191$.

Observing that the data in the three counties are positive and right skewed, the global or redundant tilt function $(x, \log(x), \log^2(x))^T$ is a sensible choice based on ZPK (2019). Hence, we initially assume that $\boldsymbol{h}_k = (x, \log(x), \log^2(x))^T$ for $k = 1, 2$, and then curtail it using the AIC model selection criterion. The AIC values corresponding to different tilts are shown in Table 1.

**Table 1:** AIC values of models based on different tilt choices. A hyphen "-" indicates that $\boldsymbol{h}_k(x) \equiv \boldsymbol{0}$ and therefore $g_0$ and $g_k$ are identical.

| AIC $\boldsymbol{h}_2$ \ $\boldsymbol{h}_1$ | - | $x$ | $\log(x)$ | $\log^2(x)$ | $(x, \log(x))$ | $(x, \log^2(x))$ | $(\log(x), \log^2(x))$ | $(x, \log(x), \log^2(x))$ |
|---|---|---|---|---|---|---|---|---|
| - | 31696.52 | 31697.86 | 31694.68 | 31697.54 | 31686.85 | 31682.73 | 31694.35 | 31684.24 |
| $x$ | 31698.24 | 31691.11 | 31695.63 | 31699.20 | 31680.96 | 31677.07 | 31696.32 | 31678.58 |
| $\log(x)$ | 31693.46 | 31685.55 | 31695.07 | 31692.86 | 31687.35 | 31683.05 | 31694.81 | 31684.70 |
| $\log^2(x)$ | 31695.67 | 31680.36 | 31696.67 | 31694.28 | 31680.14 | 31680.10 | 31691.31 | 31681.62 |
| $(x, \log(x))$ | 31693.43 | 31684.21 | 31695.04 | 31694.63 | 31682.37 | 31679.01 | 31696.63 | 31680.02 |
| $(x, \log^2(x))$ | 31693.13 | 31682.36 | 31695.03 | 31691.36 | 31681.38 | 31678.75 | 31690.98 | 31680.26 |
| $(\log(x), \log^2(x))$ | 31695.11 | 31681.91 | 31696.71 | 31693.58 | 31680.03 | 31682.06 | 31691.40 | 31681.93 |
| $(x, \log(x), \log^2(x))$ | 31694.44 | 31683.83 | 31696.05 | 31692.66 | 31680.67 | 31680.48 | 31690.13 | 31682.01 |

It is observed that the smallest AIC value of 31677.07 is achieved by the model with tilts $\boldsymbol{h}_1(x) = (x, \log^2(x))$ and $\boldsymbol{h}_2(x) = x$.

We proceed to estimate the parameters and reference CDF according to 2.2 with the chosen tilts. The confidence intervals of the tail probabilities for different thresholds obtained from both $\tilde{G}$ and $\hat{G}$ are shown in Table 2.

**Table 2:** Tail probability $1-G(T)$ estimates and 95% confidence intervals for threshold $T = 5, 10, 25, 50, 100, 150, 200, 250$.

| $T$ | $1-\tilde{G}(T)$ | 95% CI | Length of CI |
|---|---|---|---|
| 5 | 0.4447 | $(0.3773, 0.5121)$ | 0.1349 |
| 10 | 0.2790 | $(0.2004, 0.3577)$ | 0.1573 |
| 25 | 0.1482 | $(0.0693, 0.2271)$ | 0.1578 |
| 50 | 0.0915 | $(0.0201, 0.1629)$ | 0.1429 |
| 100 | 0.0548 | $(-0.0041, 0.1138)$ | 0.1178 |
| 150 | 0.0303 | $(-0.0125, 0.0732)$ | 0.0857 |
| 200 | 0.0264 | $(-0.0135, 0.0662)$ | 0.0798 |
| 250 | 0.0121 | $(-0.0142, 0.0384)$ | 0.0526 |
| $T$ | $1-\hat{G}(T)$ | 95% CI | Length of CI |
| 5 | 0.3191 | $(0.1859, 0.4524)$ | 0.2665 |
| 10 | 0.2553 | $(0.1307, 0.3800)$ | 0.2493 |
| 25 | 0.1277 | $(0.0323, 0.2231)$ | 0.1908 |
| 50 | 0.0851 | $(0.0053, 0.1649)$ | 0.1595 |
| 100 | 0.0851 | $(0.0053, 0.1649)$ | 0.1595 |
| 150 | 0.0426 | $(-0.0152, 0.1003)$ | 0.1154 |
| 200 | 0.0213 | $(-0.0200, 0.0625)$ | 0.0825 |
| 250 | 0.0000 | - | - |

From Table 2, it is readily seen that the lengths of the confidence intervals obtained by the DRM are significantly shorter than those obtained by the empirical CDF for a given threshold $T$. The slightly negative lower bounds are due to computational problems with small probabilities and should be replaced by 0's.

It is worth noting that $1-\hat{G}(250) = 0$ while $1-\hat{G}(50) = 1-\hat{G}(100)$. This is due to the fact that $\boldsymbol{X}_0$ does not contain observations between $(50, 100)$ or larger than 207. However, we can make inferences on these regions based on $\tilde{G}$ since $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ do contain observations between $(50, 100)$ or larger than 207.

**Remark:** The use of the DRM requires a justification in terms of goodness-of-fit tests discussed in KDS (2017) and in Qin (2017). As argued in Voulgaraki, Kedem, and Graubard (VKG) (2012), the DRM may not be valid for heavy tailed distributions. Examples include attempts to fit the model to data from two Cauchy distributions and from Cauchy and uniform distributions.

The graphical checking technique proposed in VKG (2012) is applied to check the goodness-of-fit of the selected model. From Figure 1, it is readily seen that the points roughly form a $45^\circ$-line, indicating the closeness of $\hat{G}$ and $\tilde{G}$ and hence an adequate DRM. A simulation of fusing absolute data from three Cauchy distributions, Cauchy(0,1), Cauchy(1,2) and Cauchy(2,3) with respective sample sizes 47, 837, 1191, and tilts $\boldsymbol{h}_1(x) = (x, \log^2(x))$ and $\boldsymbol{h}_2(x) = x$, has been conducted where the reference sample contains the absolute data from Cauchy(0,1). These are the sample sizes and tilts used in the analysis of the Forest radon data. It is observed in Figure 2 that the points are far away from a $45^\circ$-line, which indicates that the DRM is inappropriate. Such a result agrees with the examples in VKG (2012) mentioned above.

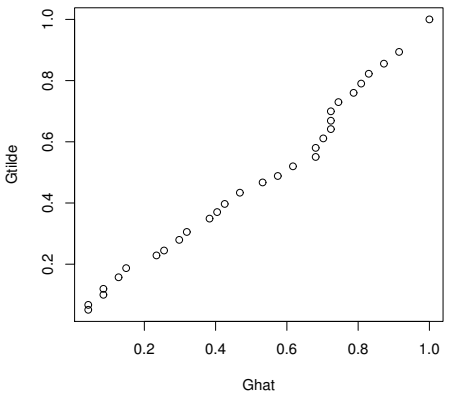


Figure 1: Pairs $(\tilde{G}(T),\hat{G}(T))$ from the selected radon data model

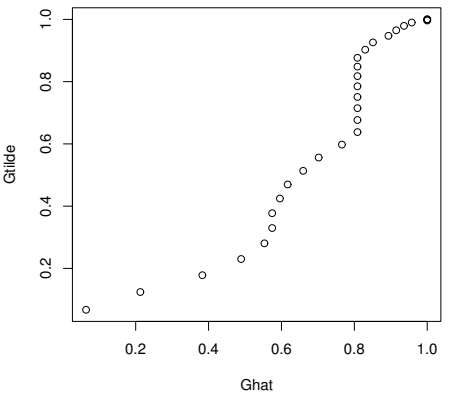


Figure 2: Pairs $(\tilde{G}(T),\hat{G}(T))$ from the DRM fit using absolute data from the three Cauchy distributions

## 4. Summary

When the size of a sample is relatively small, the empirical CDF might be inadequate for inference on distributions, while making parametric assumptions on the distributions can lead to misspecification. The DRM enables us to make semi-parametric inference about the reference distribution based on more observations, that is, based on fused samples with parametric assumptions on the ratios of the densities. These assumptions are generally weaker than the parametric assumptions on the distribution (ZPK 2019). Furthermore, an AIC based model selection renders the assumptions more sensible and hence it mitigates the problem of misspecification.

In the present residential radon application, we have seen that the lengths of the confidence intervals for tail probabilities obtained by the DRM are shorter than those obtained by the empirical CDF for a given threshold $T$.

## Acknowledgements

Research supported by a Faculty-Student Research Award, University of Maryland, College Park.

## REFERENCES

FOKIANOS, K., (2007). Density ratio model selection. *Journal of Statistical Computation and Simulation*, 77(9), pp. 805–819.

FOKIANOS, K., KAIMI, I., (2006). On the effect of misspecifying the density ratio model. *AISM*, 58, pp. 475–497.

KEDEM, B., DE OLIVEIRA, V., SVERCHKOV, M., (2017). *Statistical Data Fusion*, World Scientific, Singapore.

OWEN, A., (2001). *Empirical Likelihood*, Chapman & Hall/CRC, Boca Raton, Fl.

PA DEPARTMENT OF ENVIRONMENTAL PROTECTION. <https://www.dep.pa.gov /Business/RadiationProtection/RadonDivision/Pages/ Radon-in-the-home.aspx>

QIN, J., (2017)., *Biased Sampling, Over-identified Parameter Problems and Beyond*, Springer, Singapore.

RACK-AMBER, T., (2013)., American Lung Association in Pennsylvania to provide free radon testing kits. <https://www.heraldstandard.com/healthy_ living/american-association-in-pennsylvania-to-provide-free-radon-testing/article_ dc55f66a-4c36-588a-87b4-6c4d7448>

VOULGARAKI, A., KEDEM, B., and GRAUBARD, B. I., (2012). Semiparametric regression in testicular germ cell data. *Annals of Applied Statistics*, 6, pp. 1185–1208.

WIKIPEDIA CONTRIBUTORS, (2019). Geology of Pennsylvania - Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/wiki/Geology_of_Pennsylvania> [Online; accessed 20-December-2019].

ZHANG, X., PYNE, S., KEDEM, B., (2019). Estimation of Radon Concentration in Pennsylvania Counties by Data Fusion. *arXiv e-prints*, arXiv:1912.08149.