Statstical Variable Inference Sampling in a Prediction Model

Shin ae Kim[a,b], Yunjong Lee [a,*]
[a]*Korea Atomic Energy Research Institute, 29 Geumgu-gil Jeongeup, Republic of Korea*
[b]*Department of Nuclear Engieering, Hanyang University, 222 Wangsimni-ro Seongdong-gu, Seoul, Republic of Korea*
[*]*Corresponding author: yjlee@kaeri.re.kr*

1. Introduction

Predicting the dispersion of contaminants when a large amount of radioactive material is released into the atmosphere, such as during a terrorist activity or an event like the Fukushima nuclear accident, is very important for effective response and emergency post-evaluation[1,2]. In general, accurate local and regional meteorological data can be obtained through monitors installed at local obseravatories. However, in the case of a sudden accident, it takes time to collect data on emission concentration, location , and time. These must be inferred from the local monitor measurements. In this study, to sample the source position information needed, a function was defined of which the output is the data value of x[3]. Then, several exponetial and normal distrbution functions were applied to the more general Gaussian plume model and the functions compared[4,5]. In addition, algoriths for various techniques such as inverse CDF, rejection sampling, and importance sampling were written and implemented. Through this, a preliminary study on the research methodology for regressively estimating the location information to know was conducted and the possibility was reveiwed.

2. Methods and Results

2.1 Sampling

In using a probabilistic model, because it is impossible to infer a probability value accurately, we have no choice but to rely on an approximation technique. Sampling was applied as a method to implement this. For sampling, assuming that there is a certain sereis of distributions p(x), it is possible to obtain the expected distribution E(f) of the desired value by setting f(x) with respect to x. This is expressed in the following way.

$$E[F] = \int f(x)\, p(x)\, dx \quad (1)$$

However, if several measurement data can be extracted from the probability distiribution p(x) through the local monitor, the expected distribution for the source position to be known can be expressed as Ep. 2.

$$f = \frac{1}{L} \sum_{l=1}^{L} f(z^l) \quad (2)$$

In Equation 2, Z is each sample obtained through sampling. A method for implementing such sampling is as follows. The distributions of various physical phenomena, such as movemeny distances and scattering angles of particle, can be expressed as probability density functions (PDF). To extract numerical values randomly according to their probability density, the numerical values having a uniform distribution between 0 and 1 generated by the random number generator should be changed to distributed according to the probability density. For this purpose, there are two methods : the inversion method and the rejection sampling method.

2.2 Simulation

The R program (x64 3.6.2 version) was used to implement the sampling technique. R is an open-source program, a language for statistics/mining and graphing. It is mainly used for research and industry-specific applications for big data analysis. R was introduced in 1993 as a free version of S-PLUS by Ross Ihaka and Robert Gentleman of the University of Auckland, New Zealand. R has now absorbed most of the users of the existing S-PLUS (S language). Although it is open-source, it provides high-speed computing, excellent data processing, and works with various software and APIs with Google and Amazon coloud services, which is good[6]

2.3 Sampling Technique

(1) Inverse Transform Function Sampling

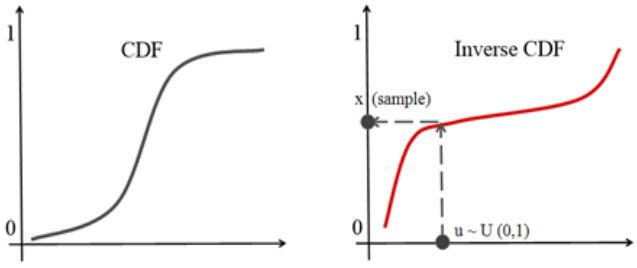

Fig. 1. Output current of the SiC detector for three particles

that have been simulated as interacting in the detector randomly in time, with an average event rate of 108 events/s.

$$X \sim f(x) \rightarrow F(x) = U \sim Unif[0,1]$$

$$F(X) = U \leftrightarrow F(X) \times F^{-1}(X) = F^{-1}(U) \leftrightarrow X = F^{-1}(U)$$

To implement the inverse CDF method, an arbitrary function was defined and a sampling algorithm was implemented. Inverse function sampling is a solution to a relatively easy function that can find the method with a equation 7 to 10.

$$x \sim Exp(\lambda) \quad (\lambda > 0)$$

$$f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \quad (0 \le x \le \infty)$$

$$F(X) \quad U \leftrightarrow F(x) \quad \int_0^x f(t)\,dt \quad \int_0^x \frac{1}{\lambda} e^{-\frac{t}{\lambda}} \quad 1 - e^{-\frac{x}{\lambda}} \quad u$$

$$x = F^{-1}(u) \leftrightarrow e^{-\frac{x}{\lambda}} = 1 - u \leftrightarrow x = -\lambda \log(v)$$

Random sampling times were extracted for the set function in Figure 2. It was confirmed that, as the number of times was repeated, the random sampling value in the function set at random, approximated the exponential function, which was target function. When sampling is performed assuming a uniform distribution in this way, it is possible to extract the model variable to be found.

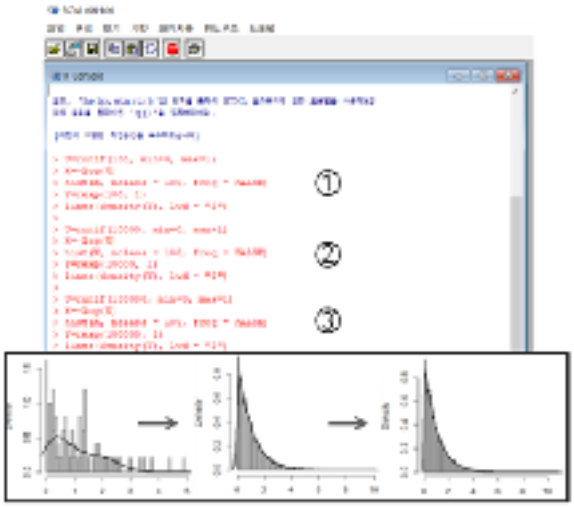

Fig. 2. Fraction of counts lost with voltage and charge sensitive preamplifiers as a function of the true count rate.

Figure 3 was confirmed that the random sampling value in the set function approximates the target function.

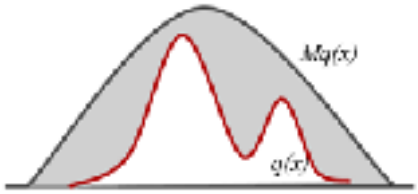

Fig. 3. It was confirmed that the random sampling value in the arbitary set function approximates the target function.

The proposed distribution is denoted as M(x), which acts as a cover function including the target distirbution f(x). Rejection sampling is a technique in which a candidate to be adopted from M(x) is extracte,d after which the sampling probability is corrected by rejecting some samples. The proposed distribution was set using a uniform or normal distribution,.

$$X = (x_1, x_2, \ldots, x_N)$$

$$x_0 \sim q(x),\ u \sim U(0,1) \quad \text{if } u < \frac{p(x_0)}{Mq(x_0)}$$

$$p(\text{accept}|x) = \int \frac{p(x_0)}{Mq(x_0)} q(x)dx = \frac{1}{M} \int p(x)dx$$

To perform sampling following the normal distribution using the exponential distribution, the target function and the target and proposal functions were set as follows.

$$\text{Target Density} : f(x) \sim N(0,1) \quad (-\infty \le X \le +\infty) \qquad eq.10.$$

$$\text{Proposal Density} : g(x) \sim Exp(1) \quad (0 < x < +\infty) \qquad eq.11.$$

$$x = F^{-1}(u) \leftrightarrow e^{-\frac{x}{\lambda}} = 1 - u \leftrightarrow x = -\lambda \log(v)$$

$$F(X) \quad U \leftrightarrow F(x) \quad \int_0^x f(t)\,dt \quad \int_0^x \frac{1}{\lambda} e^{-\frac{t}{\lambda}} \quad 1 - e^{-\frac{x}{\lambda}} \quad u$$

Fig. 4. It was confirmed that the random sampling value in the arbitary set function approximates the target function.

Random sampling times were extracted for the set function in Figure 4. It was confirmed that, as the number of times was repeated, the random sampling value in the function set at random, approximated the exponential function, which was target function. When sampling is performed assuming a uniform distribution in this way, it is possible to extract the model variable to be found.

3. Conclusions

The inverse CDF method is the most basic method and has the disadvantage of having to find the inverse function. During the sampling process, sampling near -∞ and ∞ was difficult. Rejection sampling is when it is said that the purpose of obtaining a sample from a specific target function f(x) is that the distribution of the function f(x) to be sampled is not a commomly known distribution. This gets difficult. In this case, it is a method that uses g(x), which is an approximate function of the function f(x), to create another cover function M(x) that can cover all f(x). Thefore, it can be effective only when the distribution most similar to the target function f(x) is used.

REFERENCES

[1]Connan,O.,Smith.K.,Organo.C.,Solier.L,Maro.D.,Hébert.D.,2013:Comparison of RIMPUFF, HYSPLIT, ADMS atmospheric dispersion model outputs, using emergency response procedures, with 85Kr measurements made in the vicinity of nuclear reprocessing plant. J. Environ. Radioact., 124, 266–277, doi:

[2] Hutchinson M, Oh H, Chen W,2017,A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors, Information Fusion , Volume 36, July 2017, Pages 130-148

[3] Christopher T. AllenGeorge S. EllenHaupt YS, 2007, Improving pollutant source characterization by better estimating wind direction with a genetic algorithm, Atmospheric Environment, Volume 41, Issue 11, April 2007, Pages 2283-2289

[4] Maschio C, Schiozer D J , 2019, A new parameterization method for data assimilation and uncertainty assessment for complex carbonate reservoir models based on cumulative distribution function Journal of Petroleum Science and Engineering, Volume 183, December 2019, 106400

[5] Xia Y, Yang X ,Zhang Y, 2018, A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending, Electronic Commerce Research and Applications , Volume 30, July–August 2018, Pages 111-124

[6] Braun WJ, Murdoch DJ, 2016, A first course in statstical programming with R, Cambridge University Press.