Atmospheric Environment 152 (2017) 519–530

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Evaluation of Bayesian source estimation methods with Prairie Grass observations and Gaussian plume model: A comparison of likelihood functions and distance measures

Yan Wang [a], Hong Huang [a, *], Lida Huang [a], Branko Ristic [b]

[a] Institute of Public Safety Research, Department of Engineering Physics, Tsinghua University, Beijing, China
[b] School of Engineering, RMIT University, Melbourne, Australia

HIGHLIGHTS

- Slice sampling is introduced to solve Bayesian source estimation problems.
- Approximate Bayesian Computation method is applied to source estimation problems.
- Five likelihood functions are evaluated using field experiment data sets.
- Six distance measures are evaluated using field experiment data sets.
- Nemenyi test is used to evaluate all the methods over multiple data sets.

ARTICLE INFO

Article history:
Received 9 October 2016
Received in revised form
17 December 2016
Accepted 5 January 2017
Available online 6 January 2017

Keywords:
Bayesian inference
Source estimation
Approximate Bayesian computation
Markov chain Monte Carlo
Sequential Monte Carlo
Prairie Grass field experiment

ABSTRACT

Source term estimation for atmospheric dispersion deals with estimation of the emission strength and location of an emitting source using all available information, including site description, meteorological data, concentration observations and prior information. In this paper, Bayesian methods for source term estimation are evaluated using Prairie Grass field observations. The methods include those that require the specification of the likelihood function and those which are likelihood free, also known as approximate Bayesian computation (ABC) methods. The performances of five different likelihood functions in the former and six different distance measures in the latter case are compared for each component of the source parameter vector based on Nemenyi test over all the 68 data sets available in the Prairie Grass field experiment. Several likelihood functions and distance measures are introduced to source term estimation for the first time. Also, ABC method is improved in many aspects. Results show that discrepancy measures which refer to likelihood functions and distance measures collectively have significant influence on source estimation. There is no single winning algorithm, but these methods can be used collectively to provide more robust estimates.

1. Introduction

Source term estimation (STE) for atmospheric dispersion refers to the process of determining the emission strength and location by limited information including site description, meteorological data, remotely measured concentrations and prior information about the source term (Singh et al., 2015). When the release of hazardous materials occurs due to industrial accidents or terrorist attacks, the outcome of STE can help enhance the situation assessment as well as act as important input data for atmospheric dispersion models to track and forecast the future atmospheric transport of hazardous components.

The current state of STE is reviewed in Singh et al. (2015) and Hutchinson et al. (2017). We only provide a necessary background here by categorizing STE methods into direct, optimization and probabilistic methods and introducing several previous studies related to the topic of this paper.

In direct methods, the dispersion model is run only once in a reverse direction to determine the unknown source parameters within Eulerian or Lagrangian description (Bady et al., 2009; Flesch

* Corresponding author. Institute of Public Safety Research, Department of Engineering Physics, Tsinghua University, Beijing, 100084, China.
E-mail address: hhong@tsinghua.edu.cn (H. Huang).

http://dx.doi.org/10.1016/j.atmosenv.2017.01.014

et al., 2004).

The basic idea of optimization methods is to find the source parameters which minimize the carefully selected cost function which describes the discrepancy between the output of forward dispersion models and the measured concentration at all the available sensor locations (Zheng and Chen, 2011). Thomson et al. (2007) used Gaussian plume dispersion model and simulated annealing algorithm to locate a gas source, where the robustness of three different cost functions against various forms of noises was investigated. Ma et al. (2013) compared the performances of several optimization techniques with Gaussian plume model and different cost functions. However, STE problems are usually accompanied with a variety of uncertainty, including uncertainty in sensor observations, meteorological data, the choice of forward dispersion models, as well as uncertainty caused by assumptions in source inversion algorithms (Rao, 2007; Wang et al., 2015). To analyze the uncertainty of STE results under the framework of optimization methods, Singh and Rani (2014) approximated the variance of the source parameters by using the Hessian of the cost function.

The Bayesian probabilistic methods, on the other hand, provide a more natural means for incorporating all the uncertainty by estimating the posterior distribution of the source parameters (Keats et al., 2007; Yee, 2012). Senocak et al. (2008) modified the log-normal likelihood function by considering both zero and non-zero concentration measurements with Markov chain Monte Carlo (MCMC) sampling and Gaussian plume model. Rajaona et al. (2015) tackled this Bayesian inference problem by developing an Adaptive Multiple Importance Sampling method with normal likelihood. Lucas et al. (2016) applied a Bayesian inverse technique to quantify the effects of inflow uncertainty on tracer transport and source estimation in urban environment, also under normal likelihood assumption. Ristic et al. (2016) localized a source of hazardous release using binary measurements, where the likelihood function was determined based on Poisson and Bernoulli distribution. Zhang et al. (2014) applied the ensemble Kalman filter method in large scale STE problems like nuclear accident atmospheric dispersion, where the underlying assumption is a normal likelihood function. Ristic et al. (2015a) proposed a multiple-model likelihood-free approximate Bayesian computation (ABC) method for STE, which simultaneously used a set of candidate forward dispersion models to determine the source, where the posterior estimate using the Canberra distance is significantly worse than that using the squared Euclidian distance. Kopka et al. (2016) applied an ABC method to the estimation of mobile sources using Fractional Bias as the distance metric to measure the discrepancy between predicted concentrations and observed ones.

In Bayesian methods, the likelihood functions play a role similar to that of cost functions in optimization methods, which we referred to as Bayesian likelihood methods below to avoid confusion with ABC. In ABC methods, likelihood functions are replaced by distance measures, which have similar form to cost functions in optimization methods. All of these three can be viewed as some kind of discrepancy measure, either deterministic or probabilistic. The assignment of discrepancy measure is very important in indirect STE methods. As stated above, many researchers have compared different cost functions, and ABC is proposed to avoid specifying likelihood functions and their parameters explicitly.

However, this discrepancy always exists and is difficult to be modelled in a precise way because of three major reasons. The first is the randomness and uncertainties contained in turbulent dispersion processes (Rao, 2005). The second is that atmospheric models are imperfect and contribute a modelling error. The third is the noise contained in sensor measurements. Due to these three difficulties, it is hard to justify any discrepancy measure in a theoretical way. In this paper, different discrepancy measures in Bayesian likelihood methods and ABC methods are evaluated empirically using the Prairie Grass field experiment (Barad, 1958). The Nemenyi test (Nemenyi, 1963) is adopted to compare the performance of these different discrepancy measures. Also, efficient sampling methods for both Bayesian likelihood methods and ABC methods are incorporated.

2. Methodology

2.1. Bayesian likelihood method

In Bayesian paradigm, unknown parameters are treated as random variables. Let $\boldsymbol{\theta}$ denote the unknown parameter vector (will be specified in Section 3.1) which may contain source location, emission rate, wind speed, wind direction, etc. Let $\mathbf{z} = [z_1, z_2, \ldots, z_n]$ denote the observations of n gas sensors. According to the Bayes' theorem, the posterior probability of the parameter vector $\boldsymbol{\theta}$ is given by:

$$p(\boldsymbol{\theta}|\mathbf{z}) = \frac{p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{z})} \tag{1}$$

where $p(\mathbf{z}|\boldsymbol{\theta})$ is the likelihood function, $p(\boldsymbol{\theta})$ is the prior probability which reflects the prior knowledge about the parameter vector $\boldsymbol{\theta}$, $p(\mathbf{z})$ is the marginal probability of the observations and is computed as the integral of $p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$.

2.1.1. Likelihood functions

The assignment of prior probability is very problem-specific and will be given in Section 3.2. The likelihood function is expected to describe the information about the measurement noise and modelling uncertainties (Kaipio and Somersalo, 2006). Two choices of likelihood function are frequently used in the STE literature: normal distribution (Keats et al., 2007; Guo et al., 2009; Humphries et al., 2012; Hosseini and Stockie, 2016) and log-normal distribution (Goyal et al., 2005; Senocak et al., 2008; Wade and Senocak, 2013). The parameters of the likelihood function may be treated in different ways, ranging from assigning an empirical value, to including them into the augmented parameter vector to be estimated jointly with the source term (Keats et al., 2009; Wade and Senocak, 2013).

Models based on normal distribution are sensitive to outliers, and the influence of outliers can be reduced by using long-tailed likelihood distributions such as the family of t distribution (Gelman et al., 2013). The t family of distributions $t_\nu(u, \lambda)$ is determined by three parameters: location u, scale λ and degrees of freedom ν. When ν equals 1, t distribution is equivalent to Cauchy distribution. When ν approaches positive infinity, t distribution approaches normal distribution. In response to log-normal likelihood, a super-heavy tail log-cauchy distribution is also evaluated. Assuming that the sensor observations are independent conditioned on parameter vector $\boldsymbol{\theta}$, the likelihood function described by five different distributions are as follows:

- normal distribution (Norm):

$$p(\mathbf{z}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(z_i|\boldsymbol{\theta}) = \left(\frac{\tau}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\sum_{i=1}^{n}\frac{\tau}{2}[F_i(\boldsymbol{\theta}) - z_i]^2\right\} \tag{2}$$

where $F_i(\boldsymbol{\theta})$ is the prediction of the forward dispersion model at sensor i, τ is the precision parameter.

- log-normal distribution (Ln-Norm):