ESTIMATING SOURCE TERM PARAMETERS THROUGH PROBABILISTIC BAYESIAN INFERENCE: AN APPROACH BASED ON AN ADAPTIVE MULTIPLE IMPORTANCE SAMPLING ALGORITHM

Harizo Rajaona[1, 3], Patrick Armand[1], François Septier[2, 3], Yves Delignon[2, 3], Christophe Olry[4], and Jacques Moussafir[4]

[1]CEA, DAM, DIF, F-91297 Arpajon, France
[2]Institut Mines-Télécom / Télécom Lille, 59650 Villeneuve-d'Ascq, France
[3]LAGIS UMR 8219, 59650 Villeneuve-d'Ascq, France
[4]ARIA Technologies, 92100 Boulogne-Billancourt, France

Abstract: This paper presents an adaptive approach based on probabilistic Bayesian inference to estimate the parameters of an atmospheric pollution source term. After introducing the problem and assessing the computational framework, we present an Importance Sampling based algorithm called Adaptive Multiple Importance Sampling (AMIS). It performs an efficient calculation of the source parameter posterior distribution by iteratively upgrading the proposal's parameters and recycling all generations of weighted samples, thus allowing a faster convergence and reducing the number of necessary iterations. We highlight the results of the AMIS by comparing it to a MCMC estimation in a simple example.

Key words: *Source term estimation, Bayesian inference, Monte Carlo techniques, Adaptive Multiple Importance Sampling.*

INTRODUCTION

The threat of Chemical, Biological, Radiological, and Nuclear (CBRN) releases in the atmosphere is a key issue. Such incidents may be due to terrorist acts, using non-conventional methods such as dirty bombs in order to create panic. The origin of these events can also be accidental, for example given a leak of hazardous material on an industrial site. Either way, the development of tools to detect the source and assess the parameters of the release is a major concern for the population's safety. Scientifically speaking, the problem of source term estimation (STE) is quite challenging, because obtaining the most accurate estimation within the shortest amount of time is crucial.

There are currently several approaches to solve the STE problem, each of them using a specific set of skills. One line of study focuses on adjoint-transport modelling and retro-transport, as further developed in Pudykiewicz (1998) or Issartel and Baverel (2003) where backward simulations are computed using the principle of time-symmetry in atmospheric transport to reconstruct the source. These methods perform well, but most of them cannot quantify the uncertainty relative to the estimated source. Using deterministic Bayesian inference, it is also possible to solve the problem at a global scale as mentioned in Issartel (2005).

Another way of dealing with STE problems consists in coupling Bayesian inference with stochastic sampling. The Bayesian framework allows encompassing errors (from the model and from the observations) and dealing with both the presence and absence of possible prior information regarding the source. Sohn *et al.* (2002) used a general Bayesian Monte Carlo (BMC) method to reconstruct indoor sources successfully. Delle Monache *et al.* (2008) showed that Monte Carlo Markov Chains (MCMC) also perform well for a STE problem at continental scale by correctly estimating the Algeciras incident source term. Chow *et al* (2008) applied the MCMC methodology to an urban scenario, highlighting the need of an accurate dispersion model to generate complex flows: Computational Fluid Dynamics (CFD) was used, and came with a heavy cost in computation time. Keats *et al.* (2007) emphasized that issue and coupled MCMC and backward modelling to gain computation efficiency; this work has later been extended in Yee (2008) to multiple-source scenarios.

In this paper, we present a method in the context of probabilistic Bayesian inference, based on Importance Sampling (IS) principles. This method, called *Adaptive Multiple Importance Sampling (AMIS)* and presented in Cornuet *et al.* (2012) adds an adaptive layer on the IS basis by adjusting the parameters of the proposal distribution over an iterative scheme and enabling a recycling process over all the previously generated results at each iteration. It has proven to give good results for instantaneous releases, as shown in Ickowicz (2013). In our case, we extend the application of the AMIS algorithm to a STE situation for non-instantaneous releases with a simple example.

THE BAYESIAN FORMULATION OF THE PROBLEM

We define Y the concentration measurements given by a set of N_C sensors dispatched over a network. Y is defined within each time step by:

$$Y = (y_{1,t1}, y_{1,t2}, \dots, y_{1,tT}, y_{2,t1}, \dots, y_{Nc,tT})$$

We are interested in estimating the position $\theta = (x_s, y_s)$ of the source term, and q which is the release-rate vector. We assume that we deal with a source located on ground level, so the z_s coordinate is not mentioned. In practice, q is divided into T_S time steps, and formally represents the discretization of a non-instantaneous source over the time dimension. The number of parameters to be estimated is then $T_s + 2$. This concept of discretization is inspired by the work of Koohkan *et al.* (2012), where such a process is applied on both time and spatial domain. In our case, we narrow it down to a time-only discretization. Our data model is hence defined by:

$$Y = C_\theta q + b$$

C_θ is a source-receptor matrix of the concentrations obtained from a unitary release of a source at potential location θ: q then acts as a modulator for the matrix C_θ switching from unitary concentrations to the actual concentrations by a multiplication factor. b is the noise vector which we assume to be independently and identically distributed over the N_c sensors: it unifies the model error, the measurement error and the model representativeness error in one term. We consider it as a centred, Gaussian-distributed noise with a σ_{obs}^2 observation variance. In our Bayesian reasoning, we aim at estimating a posterior distribution $(q, \theta|Y)$ which can be rewritten as:

$$p(q, \theta|Y) = p(q|\theta, Y)p(\theta|Y)$$

$p(q|\theta, Y)$ is defined as the conditional posterior of q . We know that the measurement noise is Gaussian, so $p(Y \mid \theta, q)$ follows a Gaussian distribution. Given that q and θ are independent and if we make the assumption that the prior distribution is Gaussian, then $p(q \mid \theta, Y)$ is Gaussian. In our case, as in Winiarek *et al.* (2011) we shall assume that the prior statistics on the source is given by a Gaussian profile. This process reduces the dimension of the problem from $T_S + 2$ to 2, because the parameters of $p(q|\theta, Y)$ can be computed analytically: only θ remains to be estimated through its posterior distribution $p(\theta|Y)$. It can be expressed, following the Bayesian rule, as follows:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}$$

$p(\theta)$ is the prior probability density function of the parameter vector θ, $p(Y|\theta)$ is the likelihood function of having the concentrations y using the parameters θ, and $p(\theta|Y)$ is the posterior distribution for the θ variable. We can omit the marginal distribution $p(Y)$ which is just a normalization term in our case, and simplify:

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

The likelihood follows a Gaussian distribution, and its parameters are function of the $p(q)$ parameters as well as C_θ . Each likelihood computation implies running a forward dispersion model, which complexity directly impacts the total amount of computation time. On the other hand, there is no dependence between the problem's formulation and the dispersion model used to compute C_θ: the algorithm itself won't change even if the way of computing C_θ does.

Estimating a posterior distribution is a difficult task, because it often cannot be formulated analytically. That's why we have to rely on Monte Carlo methods to compute an approximation of our posterior. One classical way to sample from the posterior is based on Markov Chain Monte Carlo (MCMC) algorithms: there has already been extensive work on their application to source term estimation problems: see Keats *et al*., (2007), Yee E. (2008), Wade *et al*. (2013) for examples. Even though MCMC have proven to be fit for several cases, they might not provide a sufficient convergence speed, which may be critical for operational situations. In the next paragraph we introduce a different approach which constitutes the core of our method.

THE AMIS ALGORITHM

Another family of Monte Carlo methods aims at sampling as close as possible to a target distribution π, and is called Importance Sampling (IS). It consists in drawing a set of N samples $(x_1, \dots, x_N)$, called *particles*, from a proposal distribution q, and compute importance weights:

$$\forall i \in \{1, \dots, N\}, w_i = \frac{\pi(x_i)}{q(x_i)}$$

By coupling the particles and the importance weights, we can derive an approximation of the target distribution:

$$\pi(x) = \frac{1}{N} \sum_{j=1}^{N} w(x_j) \delta_{x_j}(x)$$

where δ is the Dirac function. IS can be formulated iteratively in order to refine the information at a given time by using the previous generation of particles and weights: such methods are called Population Monte Carlo (PMC) in Cappé (2002). However, if the proposal distribution is badly defined, then the convergence of such algorithms is strongly compromised.

This issue emphasises the need to use an adaptive scheme, and evolve from a static proposal to a dynamic one where the parameters are modified adaptively for every iteration of the algorithm, so that the proposal may closely fit to the target distribution. That is the main goal of the Adaptive Multiple Importance Sampling (AMIS) algorithm where the parameters of the proposal are updated, so that the Kullback-Leibler divergence between the target distribution and the proposal distribution is minimized. The AMIS also uses a recycling process on all the generations of weights and particles to enhance the convergence speed and reduce the necessary amount of iterations.

In our case, the target distribution is the posterior $p(\theta|Y)$, and the proposal is defined as a mixture of D multivariate Gaussian distributions:

$$q(x; \alpha, \Xi) = \sum_{d=1}^{D} \alpha^d q_d(x, \Xi_d)$$

The coefficients α and the distribution parameters Ξ are adjusted iteratively by the AMIS algorithm. A step-by-step presentation of the AMIS algorithm is presented in Ickowicz *et al.* (2013).

SIMULATION AND RESULTS

The AMIS algorithm was run on a simple test case, using synthetic noisy concentration values on the model of (2). We consider a square domain of 50x50 where sensors are uniformly dispatched following a 5x5 grid.

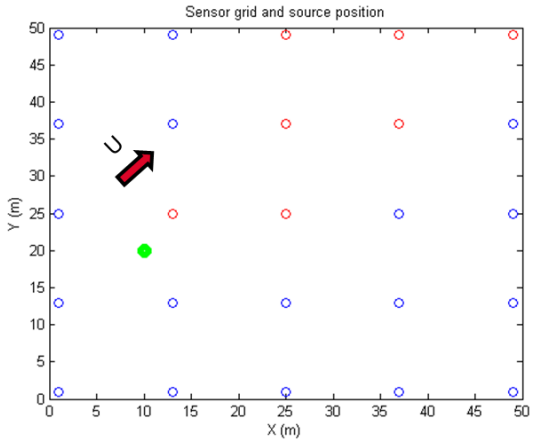

Figure 1. Schematics of the sensors setup, with the position of the source in green, the wind direction represented by a red arrow, and the sensors by coloured dots. Red sensors read nonzero values, on the contrary of blue ones.

We initialized the AMIS algorithm with a mixture of $D = 4$ components, and the only prior information we take into account is the fact that the source is within the 50x50 domain. To compute the source-receptor matrix, we use a Gaussian puff model for faster computation. We then compare the results with a MCMC algorithm based on a simple Metropolis-Hastings sampler.

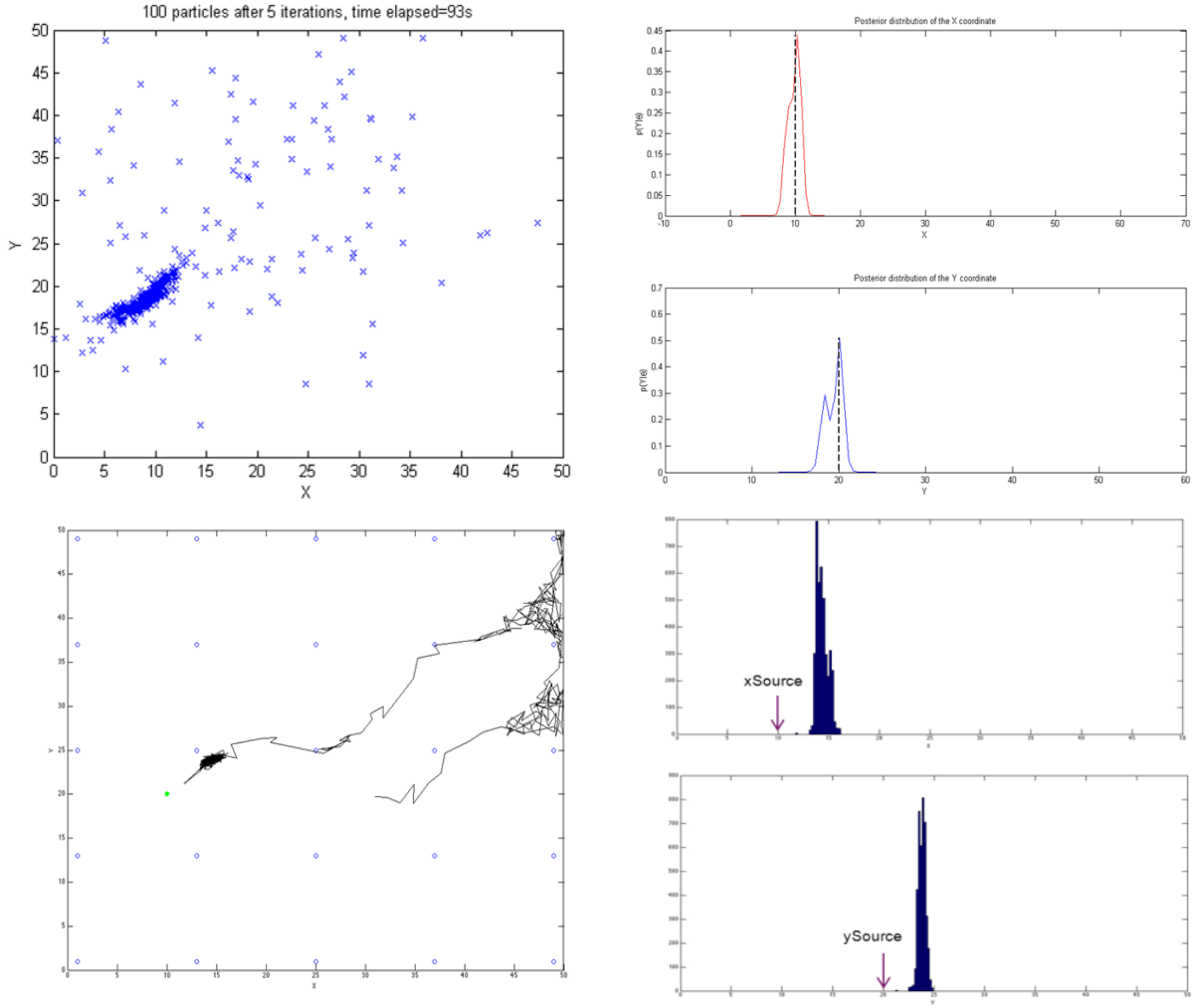

Figure 2. Results of the simulation, for a source term located at position (10, 20) for the AMIS algorithm (first row) and the MCMC algorithm (last row).
First row: final repartition of the 100 particles after 5 iterations (left), approximation of the posterior distribution of the x (top) and y (bottom) positions of the source. The dashed line represents the true value of the parameter.
Second row: trajectory of the Markov chain after 5000 iterations (left), histogram of the x (top) and y (bottom) estimated parameters.

The AMIS provided the quickest estimation within 5 iterations in 93 seconds, compared to the MCMC algorithm which took approximately took 18 minutes to run over 5000 iterations and delivered an estimate less precise than AMIS. By definition the MCMC only takes into account the previous state of the chain while iterating with the AMIS uses all the information available. Moreover, the results of the AMIS can be used starting from the very first iterations, not like the MCMC where the first phase of the algorithm called *burn-in phase* outputs results that have to be discarded. However, we realized that the quality of the AMIS algorithm is strongly conditioned by the choice of the proposal distribution and how we initialize it, as it is for most of IS based techniques. Using a retro-propagation calculation as a first step before running the AMIS, it is possible to narrow down a smaller region of interest in the space of parameters where the proposal distribution can be optimally initialized.

CONCLUSION

In this paper, we have presented an adaptive scheme in the spirit of Bayesian inference to solve STE problems. It has the advantage of converging faster than conventional algorithms such as MCMC, and can consequently fit better for emergency cases where the time of response regarding a CBRN incident must be minimal. After testing the AMIS on a simple case, we are currently in the process of validating it against an experimental setup such as the FUSION Field Trial 2007 (FFT 07) experimental campaign which has proven to be a good benchmark tool for source estimation as mentioned in Platt *et al.* (2010). Another line of work to be pursued is the consideration of non-stationary wind conditions, and how to deal with the impact of the wind uncertainty in our estimation.

The AMIS has also the advantage of being fit to run using parallel computing over the particles, not like the MCMC where there is a conditional link between two consecutive states of the Markov chain. With the possibility of using High Performance Computing (HPC), our future work will aim at coupling the AMIS with more elaborate models, such as the Parallel Micro-Swift-Spray (PMSS) tool described in Tinarelli *et al.* (2013) and apply it to non-trivial topographic cases such as urban scenarios.

REFERENCES

Cappé O., Guillin A., Marin J.M., Robert C.P., 2004: Population Monte Carlo, *Journal of Computational and Graphical Statistics,* **13,** 907-929.

Chow F.K., Kosovic B., Chan S., 2008: Source inversion for contaminant plume dispersion in urban environments using building-resolved simulations, *Journal of Applied Meteorology and Climatology,* **47,** 1553-1572.

Cornuet J.M., Marin J.M., Mira A., Robert C.P., 2012: Adaptive Multiple Importance Sampling, *Scandinavian Journal of Statistics,* **39,** 798-812.

Delle Monache L., Lundquist J.K., Kosovic B., Johannesson G., Dyer K.M., Aines R.D., Chow F.K., Belles R.D., Hanley W.G., Larsen S.C., Loosmore G., Nitao J.J., Sugiyama G.A., Vogt P.J., 2008 : Bayesian inference and Markov Chain Monte Carlo sampling to reconstruct a contaminant source on a continental scale, *Journal of Applied Meteorology and Climatology,* **47,** 2600-2613.

Ickowicz A., Septier F., Armand P., Delignon Y., 2013: Adaptive bayesian algorithms for the estimation of source term in a complex atmospheric release, *15th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes,* 1-5.

Issartel J.P., 2005: Emergence of a tracer source from air concentration measurements, a new strategy for linear assimilation, *Atmospheric Chemistry and Physics,* **5**, 249-273.

Issartel J.P. and J.Baverel, 2003: Inverse transport for the verification of the Comprehensive Nuclear Test Ban Treaty, *Atmospheric Chemistry and Physics,* **3**, 475-486.

Keats A., Yee E., Lien F., 2007: Bayesian inference for source determination with applications to a complex urban environment, *Atmospheric Environment,* **41,** 465-479.

Koohkan M., Bocquet M., Wu L., Krysta M., 2012: Potential of the international monitoring system radionucleide network for inverse modelling, *Atmospheric Environment,* **54,** 557-567.

Platt N., Deriggi D., 2010: Comparative investigation of source term estimation algorithms using FFT07 data, *13th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes,* 901-905.

Pudykiewicz J., 1998: Application of adjoint tracer transport equations for evaluating source parameters, *Atmospheric Environment,* **32**, 3039-3050.

Sohn M.D., Reynolds P., Singh N., Gadjil A.J., 2002: Rapidly locating and characterizing pollutant releases in buildings, *Journal of the Air and Waste Management Association,* **12,** 1422-1432.

Tinarelli G., Mortarini L., Castelli S.T., Carlino G., Moussafir J., Olry C., Armand P., Anfossi D., 2013: Review and validation of MicroSpray, a lagrangian particle model of turbulent dispersion, *Journal of Geophysical Research,* **200,** 311-327.

Wade D., Senocak I., 2013: Stochastic reconstruction of multiple source atmospheric contaminant dispersion events, *Atmospheric Environment,* **74,** 45-51.

Winiarek V., Vira J., Bocquet M., Sofiev V., Saunier O., 2011: Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release, *Atmospheric Environment,* **45,** 2944-2955.

Yee, E., 2008: Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference, *Boundary-Layer Meteorology,* **127,** 359-394.