21st International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
27-30 September 2022, Aveiro, Portugal

A BAYESIAN INFERENCE TECHNIQUE COMBINED WITH A STOCHASTIC WIND MODELING TO IMPROVE MULTIPLE SOURCE PARAMETERS ESTIMATION IN THE ATMOSPHERE

Roseane Albani[1], Vinícius Albani[2], Luiz Eduardo da Silva[3], Hélio Migon[1,3] and Antônio da Silva Neto[1]

[1] Polytechnic Institute, Rio de Janeiro State University, Nova Friburgo (RJ), Brazil
[2] Dept of Mathematics, Federal University of Santa Catarina, Florianopolis (SC), Brazil
[3] Institute of Mathematics, Federal University of Rio de Janeiro, Rio de Janeiro (RJ), Brazil

Abstract: We present a new methodology to estimate multiple source parameters in the framework of Bayesian Inference. In the Bayesian inference technique, the estimation problem is modeled as the determination of the full conditional probability density of the unknown parameters. In turn, the full conditionals are the product of the likelihood function, which relates observed and model-predicted concentrations and a prior density for the unknowns. The full conditionals for the unknown source parameters are not usually easily sampled. Thus, the estimation technique presented in this work combines a Metropolis-Hastings step for those unknowns that cannot be sampled directly with a Gibbs step for those sampled directly from its full conditionals. The proposed algorithm estimates the cartesian source coordinates, the source strength, the number of emissions, the precision of the observational data, and the constant in the Lasso prior distribution. In addition, we proposed a Bayesian dynamic linear model (DLM) to obtain high-frequency average wind data. In this model, we obtain the joint modeling of the wind intensity and direction. We evaluate the algorithm with the multiple tracer releases field experiment FFT07.

Key words: *Source Identification Modelling , Atmospheric Dispersion, Bayesian Inference, MCMC, Finite Element Method, Wind Forecast, Bayesian Dynamic Linear Models.*

INTRODUCTION

Atmospheric pollution is a considerable environmental concern since it strongly affects the population's quality of health and influences the ongoing climate change. In many situations, it is necessary to investigate the geographical localization and magnitude of atmospheric pollutants emissions to decide, possibly based on simulations of a dispersion model, which regions will be affected by their damaging effects, allowing the establishment of emergency measurements.

The methodologies to identify the source parameters via inverse problem modelling require a systematic procedure, which involves some coupled steps, such as the definition of a forward problem or dispersion model, usually an advection-diffusion partial differential equation (PDE). The dispersion model incorporates all the relevant physical processes involved in the case under investigation. The closer the forward problem is in describing such a situation, the better the results of the inverse problem will become. Also, the methodology to solve the forward problem strongly impacts the solution of the dispersion model and, consequently, the accuracy of the source parameters estimation.

In this work, we present a contribution to improving the forward model, more precisely in the pollutant transport by the wind, using a Bayesian Dynamic Linear Model (DLM) to obtain time-dependent wind intensity and direction. Such a model allows us to introduce the local wind changes in the dispersion model, potentially improving their accuracy. A customary procedure in source identification modelling is to substitute the advection-diffusion PDE with its adjoint one. It allows the forward model to be solved only

once for all the iterations of the inverse problem. This strategy considerably reduces the computations of the source identification algorithm, enabling faster responses. The adjoint state dispersion model is solved using the finite element formulation from (Do Carmo and Galeão, 1991) suitable to solve transient advection-diffusion PDEs. Finally, the source parameters are estimated using a Metropolis in Gibbs Monte Carlo Markov chain (MCMC) algorithm to obtain simultaneously the source parameters, the uncertainty in the data, and the parameters in the prior distributions.
In comparison with the previous work, Albani et al. (2021), there are considerable improvements. The dispersion model is now transient and considered a more realistic wind field. The proposed inversion technique is more robust since it appropriately addresses the weight of the prior densities in the estimation.

Description of the case of study

The proposed methodology to estimate simultaneously the source parameters and the number of sources is simulated and evaluated against the case 55 of the multiple emission tracer experiment FUSION Field Trial 2007 (FFT-07) (Storwald, 2007). During Trial 55, the tracer gas polypropylene (C3H6) was released continuously for 10 minutes at a constant emission rate from four-point sources and then sampled over an array of 100 sensors. The sensors were arranged considering a distance of 50 m from each other, over a sampling grid of 450 x 475 m and 2 m above the ground level. The C3H6 concentrations were sampled over about 15 minutes. High frequency 10Hz three-dimensional sonic anemometers were displayed on three 32-m lattice towers. In this work, we used the meteorological datasets measured in the centre of the grid of sensors.
The towers were equipped with five three-dimensional sonic anemometers at levels 2, 4, 8, 16, and 32 m. The sonic anemometer measured the three wind speed components \$u,v, w\$, in m/s, and the air temperature, in K. Although the meteorological dataset's time-series length is 50 minutes, just 14 minutes were applied to consider the concentration sampling period. As the concentration gas sampler units were arranged at 2 m above the ground level, we used only the zonal and meridional wind components data records measured at this same height. The arrangements of the concentration sampling units and emission sources are depicted in Fig. 1.

METHODOLOGY

The proposed methodology to identify multiple atmospheric releases is based on three main steps: the development of a Bayesian dynamic linear model (DLM) to forecast high frequency average wind data, considering Trial 55 of the FFT07 experiment. This model simultaneously takes into account the wind intensity and direction through the joint modelling of the zonal (u) and meridional (y) wind components.This model is partially based on in García et al. (2020). This wind field is one of the input data to the dispersion model. Considering the flat terrain topography, the turbulent diffusion coefficient in the vertical direction is described according to the Monin-Obukhov similarity theory. An usual procedure to solve source estimation problems is to change the advection-diffusion PDE by its adjoint one. Such a strategy considerably reduces the simulation's burden to solve the inverse problem, since in this case, the adjoint PDE needs to be solve only once, for all the iterations of the inverse problem. We avoid a detailed description of the forward problem modelling procedure for the sake of space, but it can be found, e.g. in a forthcome work and in Albani and Albani. (2019). The resulting adjoint-state PDE for the dispersion model was numerically solved using the finite element formulation proposed by Do Carmo and Galeão (1991).
We briefly describe the technique used to estimate the source parameters. Let $\mathbf{C}^{\text{obs}}$ denote the vector containing n values of observed concentrations of C3H6 at different sensors and time instants. The pollutant was emitted from a set of unknown sources. We aim to find the sources locations and their respectives emission rates. Assume, for simplicity, that the number S is fixed. Then, the vector containing the location (x_s, y_s, z_s) and strength (Q_s) of the sth source, for $s = 1, \cdots, S$, is defined as $\mathbf{w}_s = [x_s, y_s, z_s, Q_s]$. We can parameterize all the sources with respect to a given one. Thus, for $s = 2, \cdots, S$, we can write $\mathbf{w}_s = \mathbf{w}_1 + \mathbf{v}_s$. The aim is to reduce the set where the unknowns are defined, simplifying the inverse problem solution. More precisely, we let the entries of $\mathbf{w}_1$ to vary in the whole computational domain, whereas $\mathbf{v}_s$ varies in a much more restricted set. In other words, we are assuming that all the sources are close to each other.

The estimation is made by a Bayesian technique (Gamerman and Lopes, 2006). Thus, we setup prior densities for the unknowns and likelihood functions linking unknown to observed concentrations to build full conditional probability densities. As we shall see, some of such full conditionals will be sampled by the Metropolis in Gibbs Markov chain Monte Carlo algorithm (MCMC) (Albani et al., 2021). It combines a Metropolis-Hastings step for those unknowns that cannot be directly sampled with a Gibbs step for those that are sampled directly from its full conditional.

The first source has a uniform prior, i.e., $\mathbf{w}_1$ has uniform distribution in the set

$$A_\eta = [x_{min}, x_{max}] \times [y_{min}, y_{max}] \times [z_{min}, z_{max}] \times [Q_{min}, Q_{max}].$$

The xy-part of this set is defined based on the wind mean direction during the data collection and the concentration distribution in the domain given by the isopleth in Figure 2. The zQ-part of the set is the same used to solve the dispersion problem numerically.

For $s = 2, \cdots, S$, $\mathbf{v}_s$ has a Lasso prior truncated in the computational domain, with

$$P(\mathbf{v}_s|\alpha) \propto \exp\left(-\alpha\|\mathbf{v}_s\|_{\ell_1}\right).$$

For simplicity, we use the same α for all $s = 2, \cdots, S$. The scalar α is also unknown and has the prior distribution $\text{Gamma}(1/3, 1/3)$. As the concentration observation is uncertain, we assume a log-normal distribution for the noise, which means that, the relationship between observed and numerically calulated concentrations are given as follows, for $i = 1, \dots, n$,

$$\boldsymbol{C}_i^{obs} = C_i(\mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S)\exp(\varepsilon),$$

where ε is the noise and have Gaussian distribution with mean zero and variance $1/p$, with p called precision. Therefore, the likelihood function of $\mathbf{w}_1, \cdots, \mathbf{w}_S$ is the following

$$P\left(\mathbf{C}^{\text{obs}}|p, \mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S\right) \propto p^{\frac{n}{2}}\exp\left(-p\left\|\log\left(\mathbf{C}^{\text{obs}}\right) - \log(C(\mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S))\right\|_{\ell_2}^2\right).$$

The quantity p is unknown and have the prior distribution $\text{Gamma}(1/3, 1/3)$.

The full conditionals are the following,

$$P(\mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S|p, \alpha, \mathbf{C}^{\text{obs}}) \propto P\left(\mathbf{C}^{\text{obs}}|p, \mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S\right)P(\mathbf{w}_1)P(\mathbf{v}_2|\alpha) \cdots P(\mathbf{v}_S|\alpha),$$

$$P(p|\mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S, \mathbf{C}^{\text{obs}}) \propto P\left(\mathbf{C}^{\text{obs}}|p, \mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S\right)P(p),$$

$$P(\alpha|\mathbf{v}_2, \cdots, \mathbf{v}_S) \propto P(\mathbf{v}_2|\alpha) \cdots P(\mathbf{v}_S|\alpha)P(\alpha),$$

Where $P(\mathbf{w}_1)$, $P(p)$, and $P(\alpha)$ denote the prior densities for the unknowns $\mathbf{w}_1$, p, and α.

The full conditional for p and α can be approximated, as mentioned above by the distributions $\text{Gamma}(a, b)$ and $\text{Gamma}(c, d)$, respectively, where

$$a = \frac{1}{2} + \frac{n}{3}, \qquad b = \frac{1}{3} + \left\|\log\left(\mathbf{C}^{\text{obs}}\right) - \log(C(\mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S))\right\|_{\ell_2}^2,$$

$$c = \frac{1}{3} + (S - 4), \qquad d = \frac{1}{3} + \sum_{s=1}^{S} \|\mathbf{v}_s\|_{\ell_1}.$$

In the MCMC algorithm, the sampling of p and α can be performed directly from the full conditionals, leading to Gibbs steps. Samples for $\mathbf{w}_1, \mathbf{v}_2 \cdots, \mathbf{v}_S$ will be approximated by a Metropolis-Hastings step.

Numerical results

In this section, we present numerical experiments using the Metropolis in Gibbs algorithm applied to the FFT07 dataset. The xy-components of the set A_η are defined based on the 15 minutes average concentration distribution in the xy-plane provided by the isopleth in Fig. 1 (right), the distribution of the sensors in the xy-plane in Fig. 1 (left), and the 15 minutes average wind direction, also in the xy-plane indicated by the arrow in Fig. 1 (left). The resulting region in xy-plane is indicated by the rectangle in Fig. 1 (left). The zQ-components are the same used to solve the forward problem. Thus, the prior density of η is the uniform distribution defined in the set

$$A_\eta = [-100,100] \times [50,600] \times [0,50] \times [0,30],$$

with the spatial parameters defined in meters and the source strengths defined in g/s. Notice that, the actual locations of the sources are inside A_η, as the crosses in the left panel of Fig. 1 show.

The Markov Chain generated by the MCMC algorithm in Algorithm 1 has 100 thousand states, with a burn-in set of 20 thousand states. From the remaining 80 thousand states, we select 4000 using a step-size of 20 states. Table 1 compares the true source parameters with the summary statistics of the chains generated by the MCMC algorithm. For comparison, we included the parameters estimated in Albani et al. (2020). It also presents the relative error of the estimations obtained by the MCMC algorithm and in Albani et al. (2020). In Albani et al. (2020), the source parameters were estimated from the same set of experimental data using Tikhonov-type regularization, where the objective functional was minimized using different techniques, including the combination of the genetic algorithm with a gradient descent method (GA+GD) and a combination of particle swarm optimization with GD (PSO+GD). The convergence of the chains was tested using the Gelman Rubin test (Gelman and Rubin, 1992), where the value 1.00 indicates that the chains converged. The relative error is evaluated as follows,

$$RELATIVE\ ERROR = 100\frac{|TRUE\ VALUE - ESTIMATED\ VALUE|}{TRUE\ VALUE},$$

representing a percentage of the true value of the parameter of interest. The minimum and maximum relative errors for the samples on each chain were evaluated and also presented in Table 1. The average relative error for the estimated values using the MCMC algorithm and the techniques from Albani et al. (2020) were also presented. Table 1 also presents the summary statistics for the precision p, and the parameter of the Lasso prior α.

By comparing relative errors and their average values it follows that the results obtained by the MCMC algorithm were slightly more accurate than those from Albani et al. (2020). The computational time to generate the Markov chains was similar to the time to perform the estimation procedure in Albani et al. (2020). Moreover, the estimations using the MCMC algorithm were obtained from 100 concentration values from 10 different sensors, whereas the estimation in Albani et al. (2020) considered the average of the concentration values from 50 sensors. This means that the proposed inversion tool used the available data more efficiently by considering a transient dispersion model. In addition, the present technique estimates the uncertainty in the data and sets appropriately the weight of the prior distributions simultaneously to the estimation of the source parameters, which improves its accuracy and performance in comparison to the techniques presented in Albani et al. (2020). There, a Morozov-type discrepancy principle (Albani and De Cezaro, 2019) was used to appropriately find the regularization parameter.

The potential stabilization of the Markov chains is illustrated by the histograms in Fig. 3 and values equal to 1.00 in the Gelman-Rubin test presented in Tab. 1. It is worth mentioning that, although the overall accuracy of the estimations was slightly better than in previous works, in some cases, the regions defined by the Markov chains failed to contain the true values of source parameters. Since the problem under consideration is highly ill-posed and the data is rather noisy, such behavior can be expected.

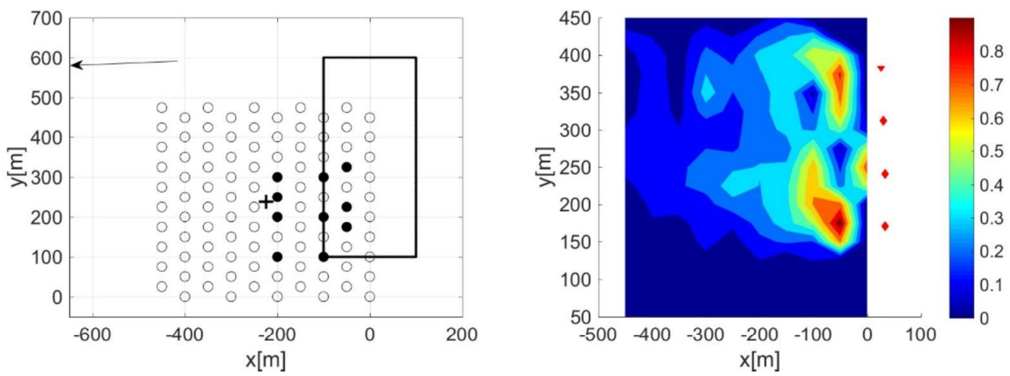

Figure 1: Left: Distribution of the sensors in the *xy*-plane of the computational domain. The arrow indicates the wind direction. The cross represents the position of the tower used to measure the wind speed and direction. The sensors

used in the simulations are the filled circles. The rectangle is the xy-part of the set A_η defining the uniform prior of η. Right: Isopleth showing the average concentration distribution for the trial 55 of the FFT07 in the computational domain. The diamonds represent the actual locations of the sources.

MCMC Summary Results									Results from Albani et al. (2020)			
Param.	True	Min.	Q1	Median	Q3	Max.	GR Test	Error (Min-Max)	GA+GD	Error	PSO+GD	Error
x_1 [m]	33,0	-12.1	-7.47	-6.80	-6.11	-1.53	1.00	121 (105–137)	33.6	1.73	28.7	13.1
y_1 [m]	171	216	225	228	232	249	1.00	33.3 (26.2–45.5)	171	0.26	172	0.59
z_1 [m]	2,00	0.50	0.84	1.02	1.18	2.28	1.00	49.2 (0.00–75.1)	2.32	16.0	0.01	100
Q_1 [g/s]	11,4	8.60	9.33	9.44	9.55	10.1	1.00	17.1 (11.1–24.5)	10.8	5.18	7.46	34.5
x_2 [m]	33,8	-5.07	-0.67	-0.01	0.65	4.70	1.00	100 (86.1–115)	8.51	74.8	28.3	16.3
y_2 [m]	241	287	298	300	302	313	1.00	24.6 (19.1–30.1)	245	1.87	240	0.36
z_2 [m]	2.00	0.01	0.32	0.50	0.66	1.77	1.00	75.2 (11.3–99.4)	1.69	15.5	2.04	2.00
Q_2 [g/s]	11.4	9.26	9.90	10.0	10.1	10.6	1.00	12.2 (7.18–18.7)	3.65	68.0	4.44	61.0
x_3 [m]	30,0	-28.9	-24.4	-23.7	-23.0	-18.6	1.00	179 (162–196)	64.0	113	200	567
y_3 [m]	313	95.5	105	108	112	129	1.00	65.4 (58.9–69.5)	305	2.66	124	60.4
z_3 [m]	2.00	1.14	1.48	1.66	1.82	2.92	1.00	17.2 (0.00–46.0)	8.69	335	1.46	27.0
Q_3 [g/s]	4.65	4.89	5.54	5.65	5.75	6.33	1.00	21.4 (5.15–36.1)	10.1	116	0.01	100
x_4 [m]	26,0	-1.89	2.68	3.34	4.03	7.65	1.00	87.1 (70.6–107)	57.5	121	-5.27	120
y_4 [m]	384	407	424	428	431	438	1.00	113 (5.77–14.1)	360	6.24	372	3.17
z_4 [m]	2.00	0.52	0.87	1.05	1.22	2.31	1.00	47.5 (0–74.0)	11.3	465	6.51	226
Q_4 [g/s]	11.4	8.95	9.64	9.75	9.85	10.4	1.00	14.4 (8.75–21.4)	24.5	115	11.6	2.19
Average								55 (36–69)		91		83
p	-	0.06	0.10	0.11	0.12	910	1.00					
α	-	188	893	1213	1594	4506	1.00					

Table 1: Comparison between the True source parameter values (True), the summary of the MCMC results, and the parameters estimated in Albani et al. (2020). For the MCMC results, we have the minimum value (Min.), the first

quartile (Q1), the median value (Median), the third quartile (Q3), the maximum value (Max.) and the value of the Gelman-Rubin convergence test (GR Test). The percentage error is the relative error multiplied by 100. It is evaluated with respect to the median value and the maximum and minimum possible error values considered in the samples (Min-Max). The relative error for the parameters estimated in Albani et al. (2020). GA+GD stands for the combination of the genetic algorithm with gradient descent (GD) and PSO+GD the combination of particle swarm optimization with GD. The row average presents the average of the relative errors.

CONCLUSION

The proposed methodology presented median values that were at least as accurate as previous approaches that used sofisticated inverse modeling techniques. This means that the presente estimation procedure that uses a more accurate dispersion model can present a better performence. However, in many cases, the credibility regions provided by the Markov chains failed include the true source parameters. This issue shall be addressed in future research by improving the MCMC algorithm and by including additional concentration data from other sensors in the computational domain.

ACKNOWLEDGEMENTS

RA thanks Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for the financial support through the grants E-26/202.932/2019 and E-26/202.933/2019. AJSN acknowledges the financial support provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) through the grants 88887.311757/2018-00 and 88887.194804/2018-00, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) through the grant 308958/2019-5, and the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) through the grant E-26/202.878/2017. HM thanks PAPD/Rio de Janeiro State University for the financial support through the grant E-26/007/10667/2019.

REFERENCES

Albani, R. and V. Albani 2019: Tikhonov-Type Regularization and the Finite Element Method Applied to Point Source Estimation in the Atmosphere. *Atmos. Environ.*, **211**, 69–78.

Albani, R., V. Albani, and A. Silva Neto 2020: Source characterization of airborne pollutant emissions by hybrid metaheuristic/gradient-based optimization techniques. *Environ Pollut.*, **267**, 115618.

Albani, R., V. Albani, H. Migon, and A. Silva Neto 2021: Uncertainty quantification and atmospheric source estimation with a discrepancy-based and a state-dependent adaptative MCMC. *Environ Pollut.*, **290,** 118039.

Albani, V. and A. De Cezaro 2019: A Connection Between Uniqueness of Minimizers and Morozov-like Discrepancy Principles in Tikhonov-type Regularization. *Inverse Probl. Imaging*, **13,** 211–229.

Do Carmo E. and A. Galeão 1991: Feedback Petrov-Galerkin methods for convection-dominated problems. *Comput Methods Appl Mech Eng*, **88,** 1–16.

Gamerman, D. and H. F. Lopes 2006: *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

García, I., S. Huo, R. Prado, and L. Bravo 2020: Dynamic bayesian temporal modeling and forecasting of short-term wind measurements. *Renewable Energy*, **161,** 55–64.

Gelman, A. and D. B. Rubin 1992: Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.

Storwold, D 2007: Detailed test plan for the fusing sensor information from observing networks (fusion) field trial 2007 (fft-07). *US Army Dugway Proving Ground West Desert Test Center Doc. WDTC-TP-07-078*, **46**.