Narodowe Centrum Badań Jądrowych
National Centre for Nuclear Research
ŚWIERK

MANHAZ

Zastosowanie algorytmów przybliżonych obliczeń bayesowskich (ang. Approximate Bayesian computation, ABC) w problemie identyfikacji źródła uwolnienia gazu w środowisku zurbanizowanym

Piotr Kopka

10.10.2017

Otwock Świerk,
10.10.2017

Motivation

- Accidental atmospheric releases of hazardous material pose great risks to human health and the environment.
- For example, release of CS-137 in a steel mill in Algeciras, Spain, V.1998 r.
- Registered in Switzerland, France and Italy in June.
- It is necessary to have properly fast response to such incidents.
- Emergency responders need quickly recognize the source of contamination.

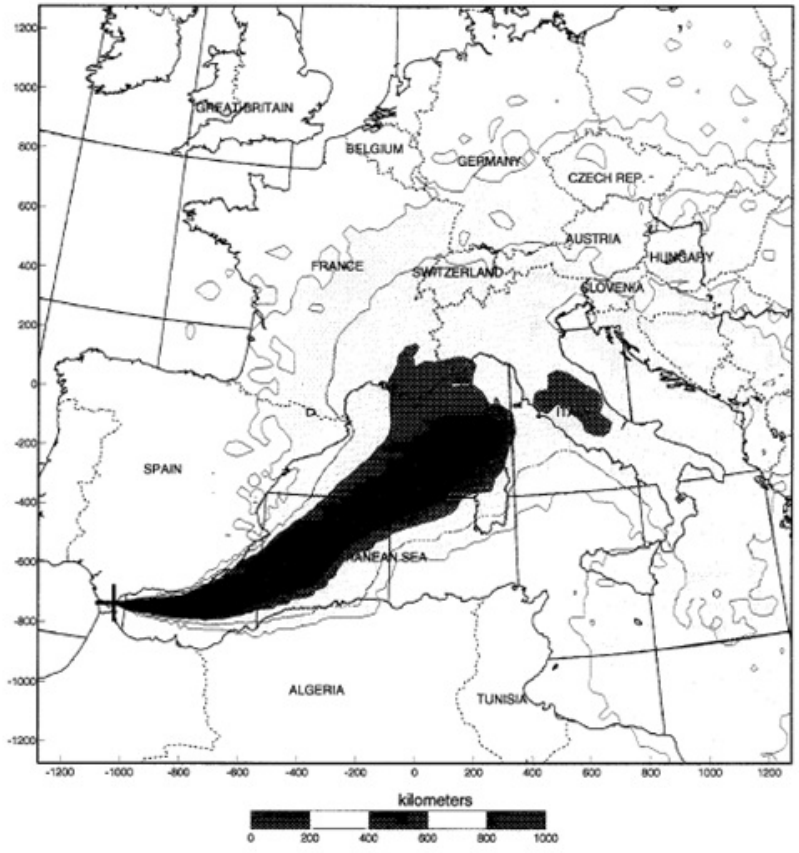


Figure 2. Seven-day average air concentration at 1200 UTC on 5 June 1998. Contours >10 (outermost or lightest), >100, >500, and >1000 uBq/m3 (innermost or darkest).

Estevan, M. (2003). Consequences of the Algeciras accident ..., Security of Radioactive Sources, 357.

Stochastic event reconstruction

- Based on sparse-point substance concentrations we have to answer the questions:

Stochastic event reconstruction

- Based on sparse-point substance concentrations we have to answer the questions:
- **Where? How much? substance was released.**

Stochastic event reconstruction

- Based on sparse-point substance concentrations we have to answer the questions:
- **Where? How much? substance was released.**
- Note: Time of giving the answer to this questions is crucial!
- How to do it?
- Build a model of pollutant transport in the atmosphere and compare point concentrations derived from the model with the measured data obtained from sensor networks

Stochastic event reconstruction

- Based on sparse-point substance concentrations we have to answer the questions:
- **Where? How much? substance was released.**
- Note: Time of giving the answer to this questions is crucial!
- How to do it?
- Build a model of pollutant transport in the atmosphere and compare point concentrations derived from the model with the measured data obtained from sensor networks
- Problem: Find the values of the pollutant transport model, (such as the location of the release source) which will the best „fit” our model to the observational data.

Bayesian inference

- In the framework of Bayesian statistics all quantities included in the mathematical model are modeled as *random variables* with joint probability distributions.

Bayesian inference

- In the framework of Bayesian statistics all quantities included in the mathematical model are modeled as *random variables* with joint probability distributions.
- This randomness can be interpreted as parameter variability, and is reflected in the uncertainty of the true values expressed in terms of probability distributions.

Bayesian inference

- In the framework of Bayesian statistics all quantities included in the mathematical model are modeled as *random variables* with joint probability distributions.
- This randomness can be interpreted as parameter variability, and is reflected in the uncertainty of the true values expressed in terms of probability distributions.
- Bayesian methods reformulate the problem into searching for a solution based on efficient sampling of an ensemble of simulations, guided by comparisons with data.

Bayesian inference

- In the framework of Bayesian statistics all quantities included in the mathematical model are modeled as *random variables* with joint probability distributions.
- This randomness can be interpreted as parameter variability, and is reflected in the uncertainty of the true values expressed in terms of probability distributions.
- Bayesian methods reformulate the problem into searching for a solution based on efficient sampling of an ensemble of simulations, guided by comparisons with data.
- So, in practice we can are looking for the values of parameters which are the most probable base on the data and some priori information

Bayesian inference

- In the framework of Bayesian statistics all quantities included in the mathematical model are modeled as *random variables* with joint probability distributions.
- This randomness can be interpreted as parameter variability, and is reflected in the uncertainty of the true values expressed in terms of probability distributions.
- Bayesian methods reformulate the problem into searching for a solution based on efficient sampling of an ensemble of simulations, guided by comparisons with data.
- So, in practice we can are looking for the values of parameters which are the most probable base on the data and some priori information = **a posteriori probability of dispersion model parameters**

DATA

- DAPPLE - Dispersion of Air Pollution and its Penetration into the Local Environment

DATA

- DAPPLE - Dispersion of Air Pollution and its Penetration into the Local Environment
- *The DAPPLE experimental work took place in central London between 2002-2006*
- *The mean building height in the study area is* $21.6m$ *(range* 10 *to* $64m$*)*
- *Total mass emitted from point-source release was* $323mg$ *of perfluoromethyl-cyclohexane* $(PMCH, C7F14)$
- *Two sets of long-term reference measurements were taken to generate the wind data sets, rooftop City Council (WCC) (18 m) and tower top (190 m) winds.*

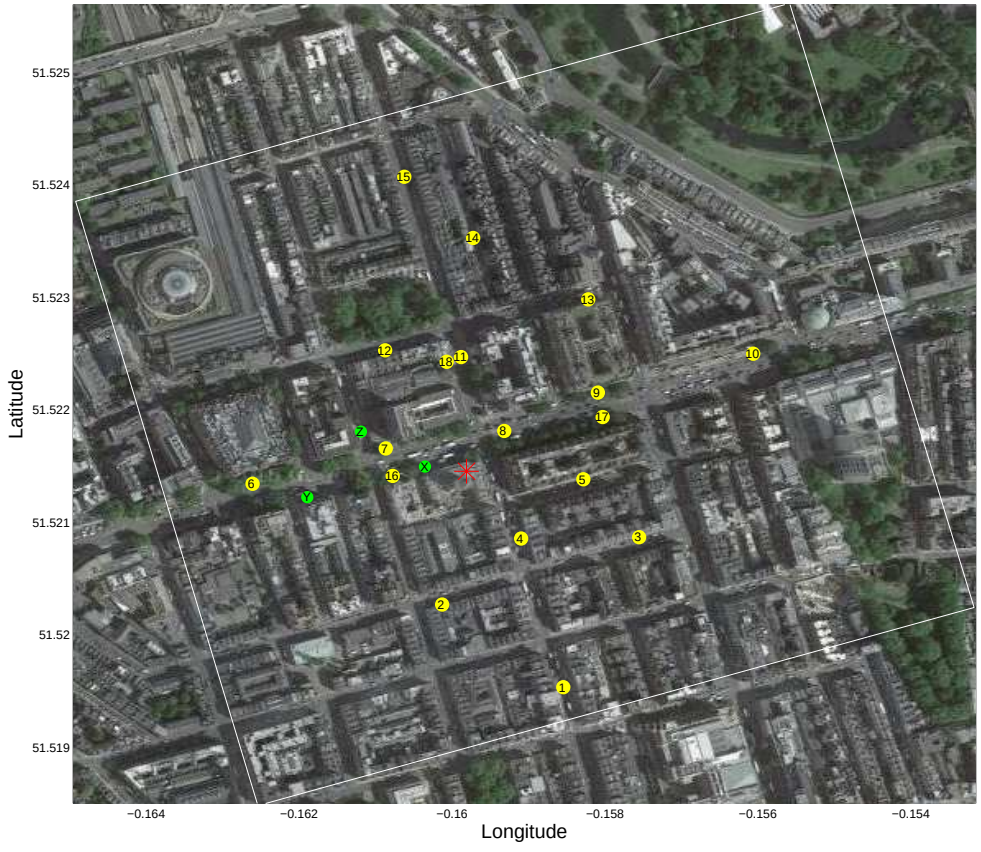


Figure 1: The sampling receptors 1-18 (yellow circles). Three fixed-point tracer sources (green dots X,Y and Z); red star - City Council (WCC); white rectangle - determined computational domain.

Figure 2: Fixed-point tracer source X; City Council (WCC);

MODEL

Requirements for the model of pollutant transport:

- It can not be too complicated, because we might do not know all the necessary parameters such as wind field, the coefficients of turbulence, weather conditions, etc.

MODEL

Requirements for the model of pollutant transport:

- It can not be too complicated, because we might do not know all the necessary parameters such as wind field, the coefficients of turbulence, weather conditions, etc.
- A short computation time.

MODEL

Requirements for the model of pollutant transport:

- It can not be too complicated, because we might do not know all the necessary parameters such as wind field, the coefficients of turbulence, weather conditions, etc.
- A short computation time.
- Adequate to the situation (proper model).

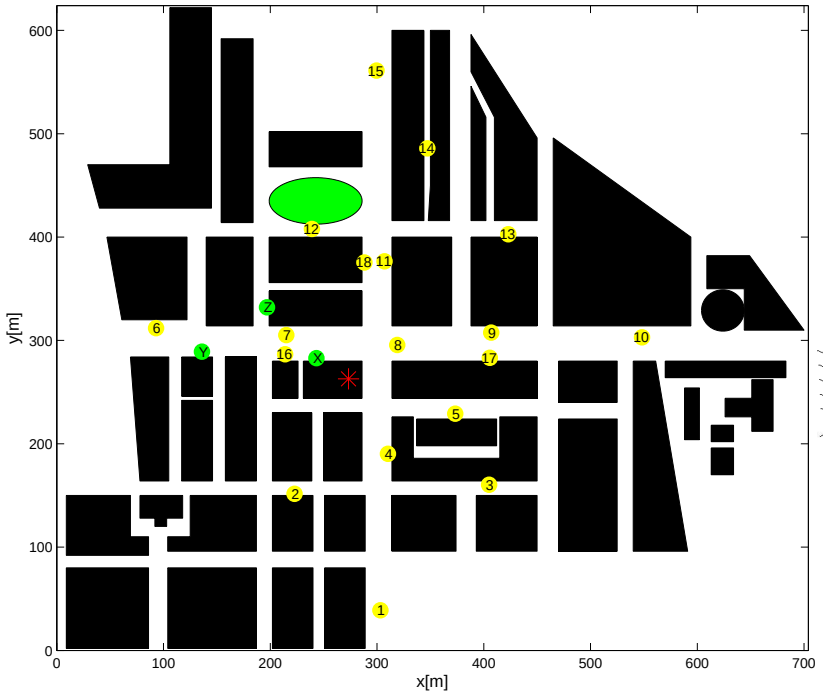
MODEL QUIC

The **QUIC** uses a stochastic Lagrangian random walk approach to estimate concentrations in a gridded domain. The model is designed to use averaged wind fields produced by the QUIC-URB system. Parcels, representing substance are transported with a vector sum of mean winds from QUIC-URB plus turbulent fluctuating winds computed using the random walk equations.

QUIC dispersion model is non deterministic

MLA Williams, Michael D., Michael J. Brown, and Eric M. Pardyjak. "Development of a dispersion model for flow around buildings." Fourth Symposium on the Urban Environment. 2002.

Gowardhan, A. A., et al. "Evaluation of the QUIC urban dispersion model using the Salt Lake City URBAN 2000 tracer experiment data—IOP 10." Sixth Symposium on the Urban Environment/14th Joint Conference on the Applications of Air Pollution Meteorology with the Air and Waste Management Association. Atlanta. January J. Vol. 6. 2006.



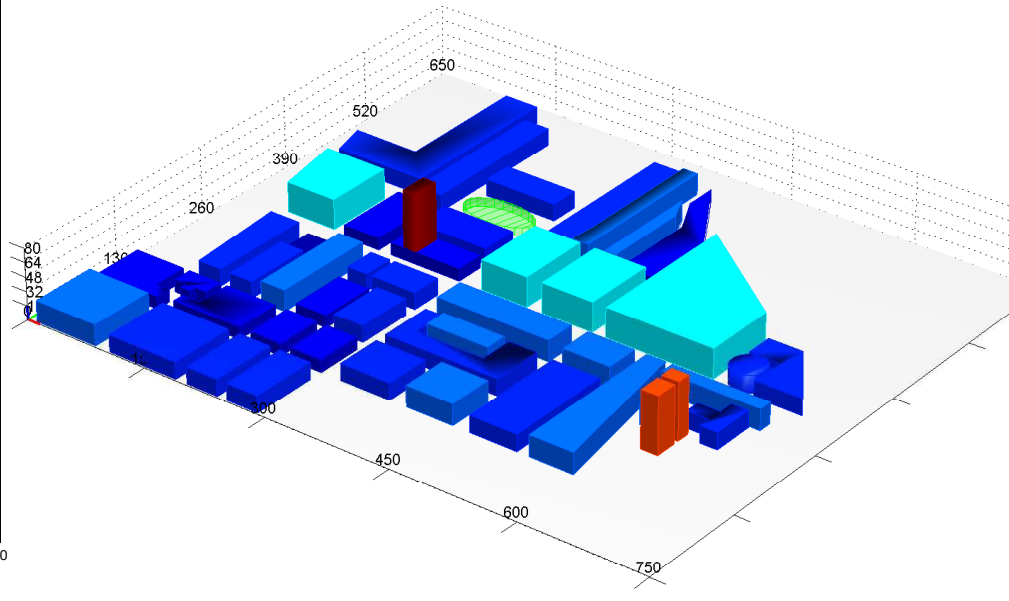
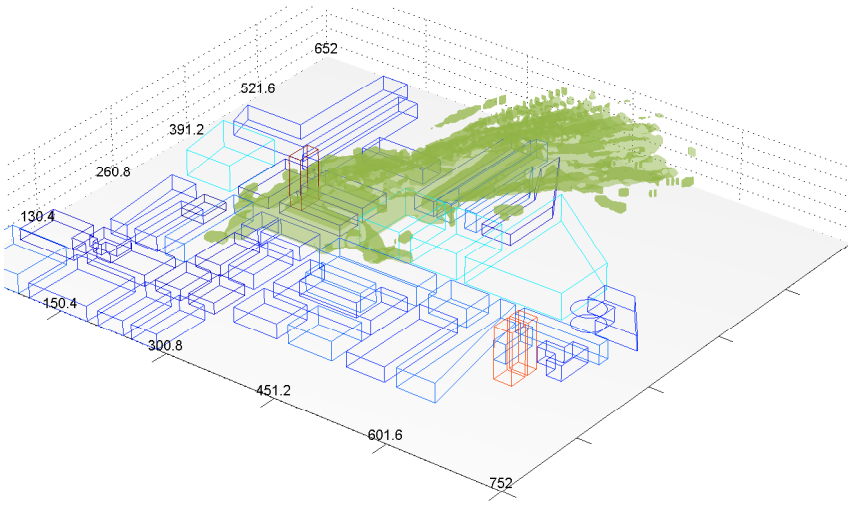


Figure 3: a) Extracted buildings (black rectangle) and greenery (green ellipse) ; sampling receptors are numbered 1-18 (yellow circles), three fixed-point tracer sources (green dots X,Y and Z); red star - Westminster City Council (WCC) b) 3D model of city buildings designed in QUIC-GUI



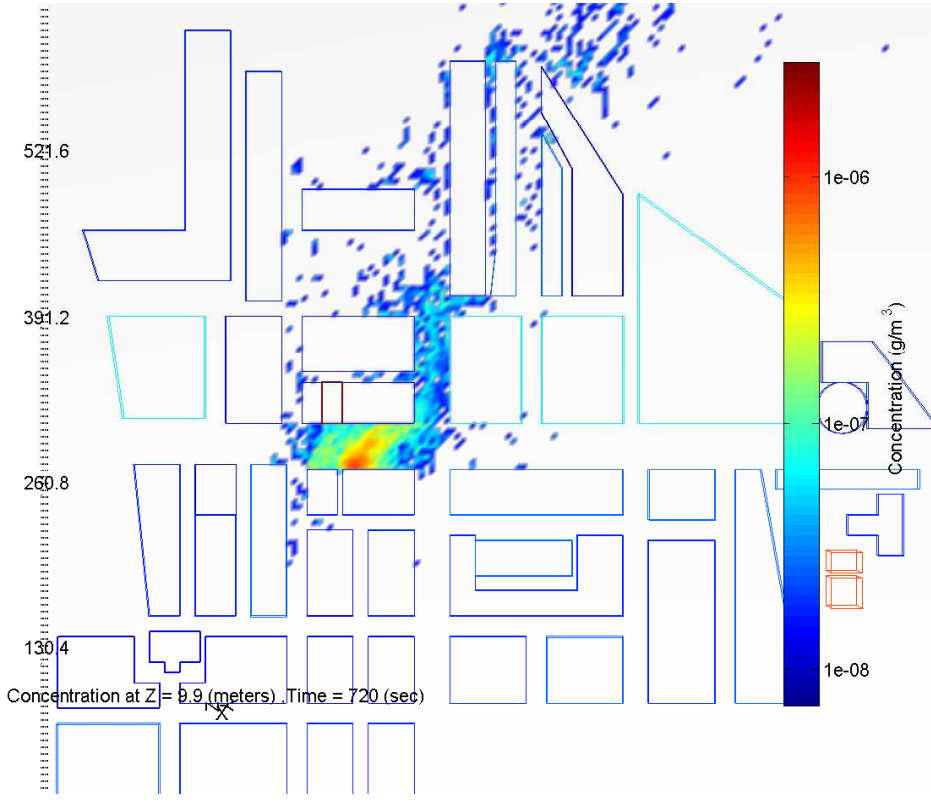


Figure 4: a) The shape of the gas cloud - take from QUIC b) Concentration of the substance at a height of 2 m

Bayesian Inference

Let θ be a parameter vector, given the prior distribution $\pi(\theta)$.

The goal of Bayesian inference is to approximate the posterior distribution, $\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta)$, where $\pi(x|\theta)$ is the likelihood of θ given the data x.

Approximate Bayesian Computation (ABC)

The ABC methods is to accept θ as an approximate posterior draw if its associate data x is close enough to the observed data x_{obs}. Accepted parameters are a sample from $\pi(\theta|\rho(x, x_{obs}) < \epsilon)$ where the $\rho(x, x_{obs})$ is the chosen measure of discrepancy, and ϵ is a threshold defining "closeness margin".

Approximate Bayesian Computation (ABC)

The ABC methods is to accept θ as an approximate posterior draw if its associate data x is close enough to the observed data x_{obs}.
Accepted parameters are a sample from $\pi(\theta|\rho(x, x_{obs}) < \epsilon)$ where the $\rho(x, x_{obs})$ is the chosen measure of discrepancy, and ϵ is a threshold defining "closeness margin".
It is often difficult to define a adequate distance function $\rho(x, x_{obs})$ between the simulate and observe data.

Approximate Bayesian Computation (ABC)

The ABC methods is to accept θ as an approximate posterior draw if its associate data x is close enough to the observed data x_{obs}.
Accepted parameters are a sample from $\pi(\theta|\rho(x, x_{obs}) < \epsilon)$ where the $\rho(x, x_{obs})$ is the chosen measure of discrepancy, and ϵ is a threshold defining "closeness margin".
It is often difficult to define a adequate distance function $\rho(x, x_{obs})$ between the simulate and observe data.
In many cases it is replace with a distance defined on summary statistics, $\rho(S(x), S(x_{obs}))$.

Approximate Bayesian Computation (ABC)

In ABC methods, Sequential Monte Carlo (SMC) is used in order to automatically, sequentially "clean" approximation of posterior distribution to be used to generate proposals for further steps.
In methods, the set of samples with weights, called particles sampled from the prior distribution $\pi(\theta)$, are propagated through a sequence of intermediate posterior distributions $\pi(\theta|\rho(x, x_{obs}) < \epsilon_t)$, $t = 1, ..., T$, until it represents a sample from the target distribution, $\pi(\theta|\rho(x, x_{obs}) < \epsilon_T)$.
These methods aim to generate draws from $p(\theta|\rho(x, x_{obs}) < \epsilon_t)$, at each of a series of sequential steps t, where ϵ_t define a series of thresholds.

ABC Algorithm (1)

- *Initialize threshold schedule* $\epsilon_1 < \epsilon_2 < ... < \epsilon_T$
- *Set* $t = 1$
- *For* $i = 1$ *to* N
- *Simulate* $\theta_i^t \sim \pi(\theta)$ *and* $x \sim \pi(x|\theta_i^t)$
- *Until* $\rho(x, x_{obs}) < \epsilon_t$
- *Set* $w_i^t = \frac{1}{N}$

Bonassi, F. V., West, M. (2015). Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. Bayesian Analysis, 10(1), 171-187

ABC Algorithm (2)

- *For* $i = 2$ *to* T
- *Compute new weights* $v_i^{t-1} \propto w_i^{t-1} K_{x,t}(x_{obs}|x_i^{t-1})$
- *Normalize weights* v_i^{t-1}
- *For* $i = 1$ *to* N
- *Pick* $\tilde{\theta}_i$ *from the* θ_j^{t-1} *set with probabilities* v_i^{t-1}
- *Draw* $\theta_i^t \sim K_{\theta,t}(\theta_i^t|\tilde{\theta}_i)$ *and* $x \sim \pi(x|\theta_i^t)$
- *Until* $\rho(x, x_{obs}) < \epsilon_t$
- *Compute new weights as*
- $w_i^t \propto \frac{\pi(\theta_i^t)}{\sum_j v_j^{t-1} K_{\theta,t}(\theta_i^t|\theta_j^{(t-1)})}$
- *Normalize weights* w_i^t *for* $i = 1 : N$

Bonassi, F. V., West, M. (2015). Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. Bayesian Analysis, 10(1), 171-187.

Data and distance measure

The choice of distance measure or summary statistics is a crucial step in ABC.

Data and distance measure

The choice of distance measure or summary statistics is a crucial step in ABC.

Since distance measure are not sufficient in many cases, this choice involves a trade-off between loss of information and reduction of dimensionality.

Data and distance measure

The choice of distance measure or summary statistics is a crucial step in ABC.

Since distance measure are not sufficient in many cases, this choice involves a trade-off between loss of information and reduction of dimensionality.

In those type of problems we chose normalize approximation error between all the data obtained to the current time step t with is also called Fractional Bias (FB).

The FB is used to indicate a bias towards underprediction or overprediction of concentration data by the model.

Cox, W.M.: Protocol for determining the best performing model. Technical report, Environmental Protection Agency, Research Triangle Park, NC (United States). Technical Support Div. (1992)

Data and distance measure

The $\rho(x^t, x^t_{obs})$ measure in time step t:

$$\rho(x^t, x^t_{obs}) = \frac{1}{18}\sum_{j=1}^{18}\left(\frac{1}{t}\sum_{i=1}^{t}\frac{|C_i^{Sj} - \hat{C}_i^{Sj}|}{C_i^{Sj} + \hat{C}_i^{Sj}}\right), \tag{1}$$

under the assumption, that when $C_i^{Sj} = 0$ and $\hat{C}_i^{Sj} = 0$ then $\frac{|C_i^{Sj} - \hat{C}_i^{Sj}|}{C_i^{Sj} + \hat{C}_i^{Sj}} = 0$;

Data and distance measure

The $\rho(x^t, x^t_{obs})$ measure in time step t:

$$\rho(x^t, x^t_{obs}) = \frac{1}{18}\sum_{j=1}^{18}(\frac{1}{t}\sum_{i=1}^{t}\frac{|C_i^{Sj} - \hat{C}_i^{Sj}|}{C_i^{Sj} + \hat{C}_i^{Sj}}), \tag{1}$$

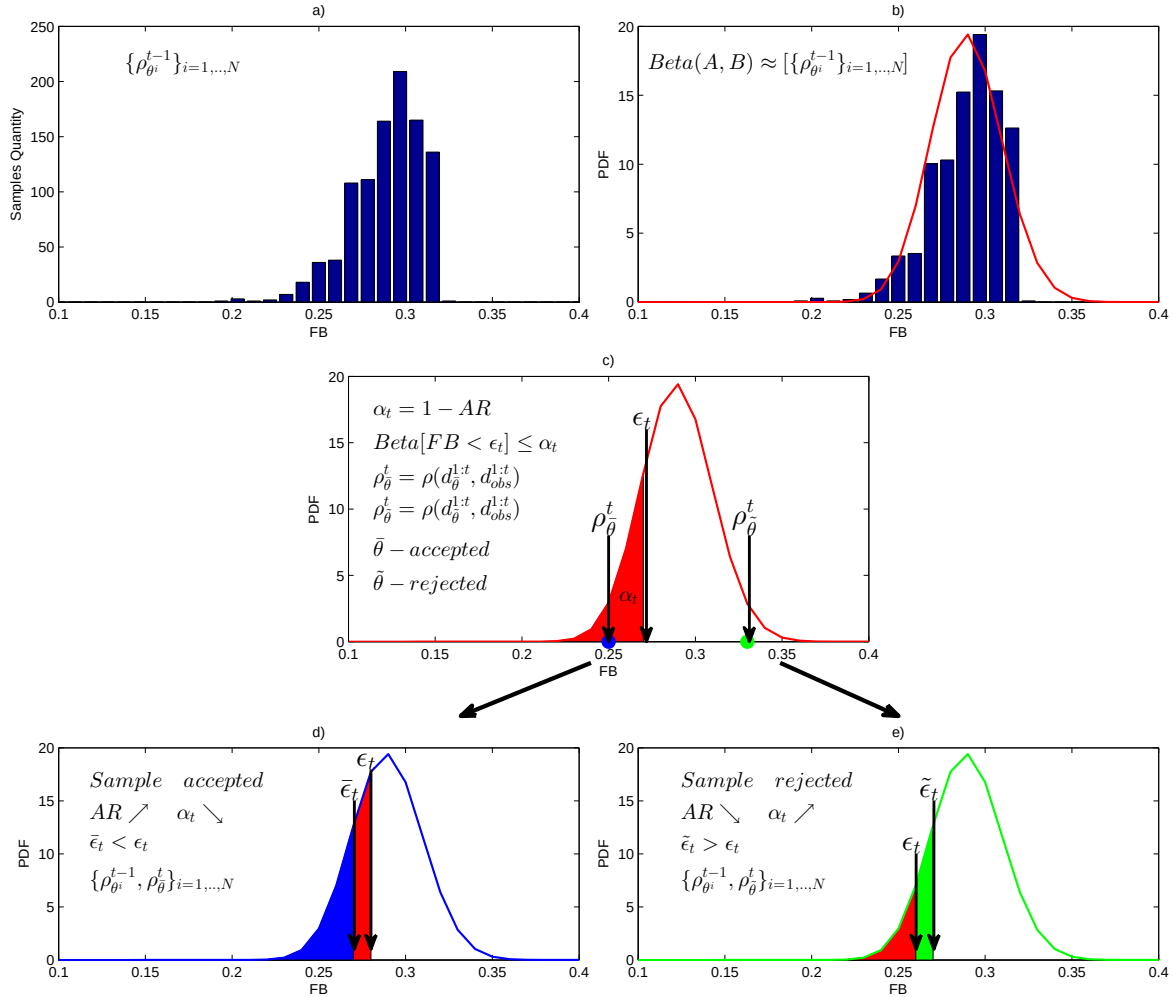
under the assumption, that when $C_i^{Sj} = 0$ and $\hat{C}_i^{Sj} = 0$ then $\frac{|C_i^{Sj} - \hat{C}_i^{Sj}|}{C_i^{Sj} + \hat{C}_i^{Sj}} = 0$;

Given that the concentration C_i^{Sj} has always value ≥ 0, a $\rho(x^t, x^t_{obs})$ is always between 0 and 1 . Let us recognize that $\rho(x^t, x^t_{obs}) = 0$ is the situation when our prediction fit perfectly. In opposite $\rho(x^t, x^t_{obs}) = 1$ does not fit at all.

Threshold schedule and weights

The most commonly used adaptive scheme for threshold choice is based on the quantile of the empirical distribution of the distances between the simulated data and observations from the previous population θ_j^{t-1}.

Threshold schedule



Transition kernel

In this type of inverse problems the parameters are often highly correlated and multimodality is very common.

Transition kernel

In this type of inverse problems the parameters are often highly correlated and multimodality is very common.
Especially, when the (x, y) domain contains a lot of prohibited regions, like buildings. Samples may tend to splitting in a disjointed group by filling out a different canyons streets.

Transition kernel

In this type of inverse problems the parameters are often highly correlated and multimodality is very common.
Especially, when the (x, y) domain contains a lot of prohibited regions, like buildings. Samples may tend to splitting in a disjointed group by filling out a different canyons streets.
Transition kernel $K_{\theta,t}(\cdot|\cdot)$ - multivariate normal kernel. In such cases it is interesting to consider the use of a local mean and covariance matrix. Instead, computing the covariance matrix based on all the samples from $(t-1)$ better idea is use only limited information about the local correlation.

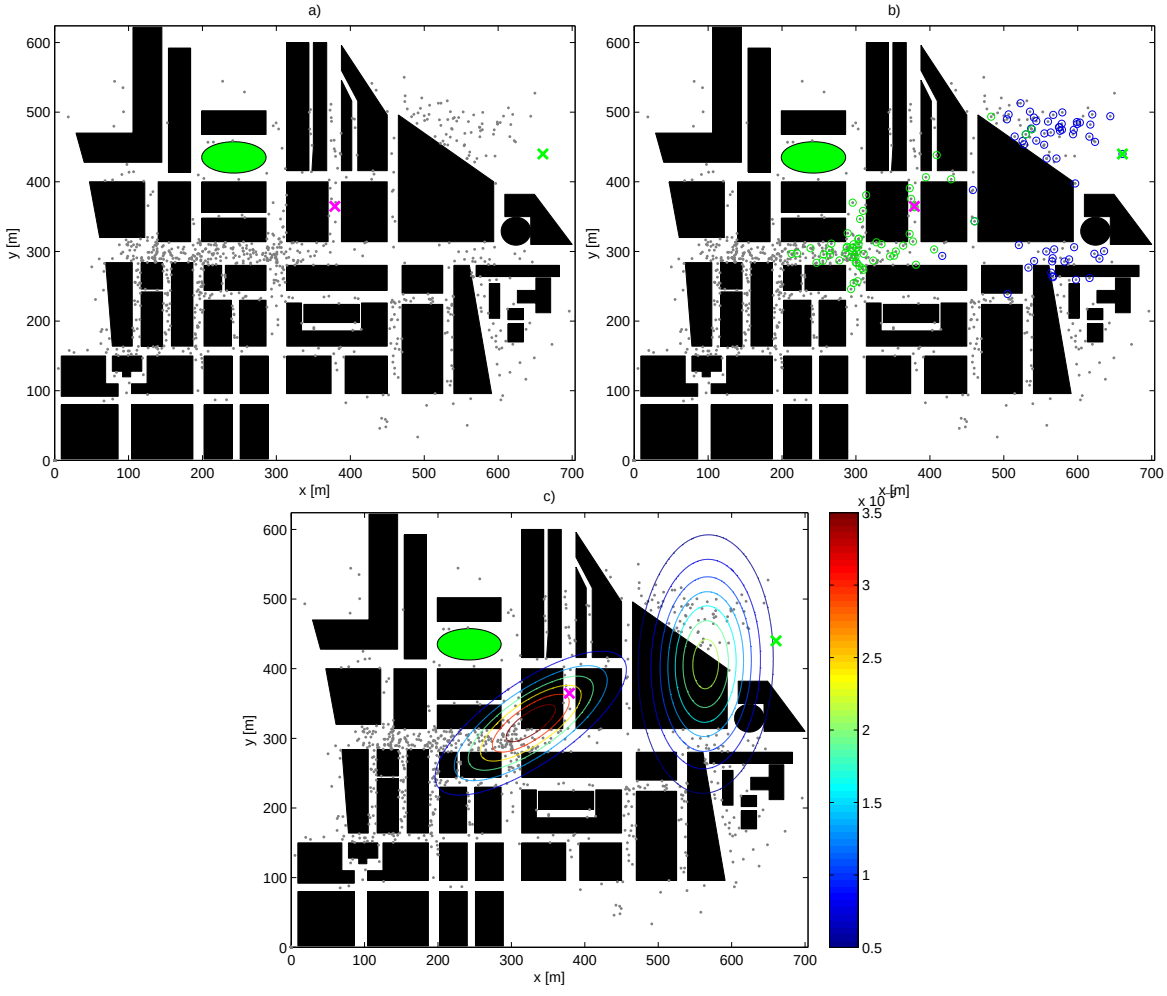
Transition kernel

In this type of inverse problems the parameters are often highly correlated and multimodality is very common.
Especially, when the (x, y) domain contains a lot of prohibited regions, like buildings. Samples may tend to splitting in a disjointed group by filling out a different canyons streets.
Transition kernel $K_{\theta,t}(\cdot|\cdot)$ - multivariate normal kernel. In such cases it is interesting to consider the use of a local mean and covariance matrix. Instead, computing the covariance matrix based on all the samples from $(t-1)$ better idea is use only limited information about the local correlation. In *[Filippi S. and others : On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo* one the proposal methods is to use the multivariate normal kernel base on the M neighbours.

Narodowe Centrum Badań Jądrowych
National Centre for Nuclear Research
ŚWIERK

MANHAZ

Transition kernel - example



Parameters vector

The following parameters vector describe the source of the release:

$$\theta \equiv (x, y, z, q, l, s). \tag{2}$$

The (x, y) is source position within computing domain, (z) is the height of source location above ground level, (q) is a mass of release, (s) is the start time of release and (l) is duration time.
In the presented procedure we declare the following priori distribution on particular parameters:

$$\begin{aligned}
\pi(\theta^1) \equiv & \\
& (x, y) \sim U^{\Theta}([100, 600], [100, 500]) \\
& z \sim Gamma(3, 3) \\
& q \sim U(10, 500) \\
& l \sim U(0, 1800) \\
& s \sim U(0, 180).
\end{aligned} \tag{3}$$

Results of DAPPLE reconstruction experiment

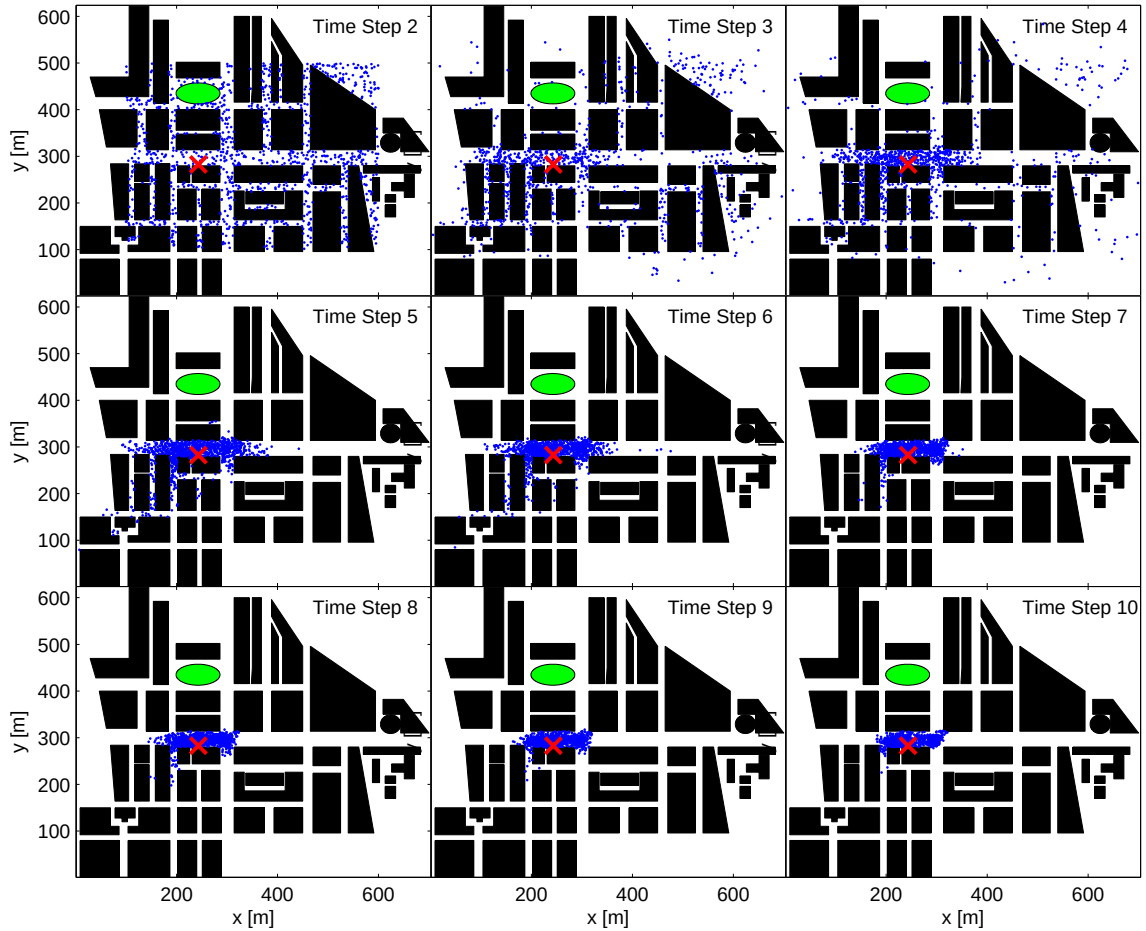


Figure 7: Scatter plot of all samples generate in the subsequent time steps $t = 2, 3, \ldots, 10$ in (x, y) space

Results of DAPPLE reconstruction experiment

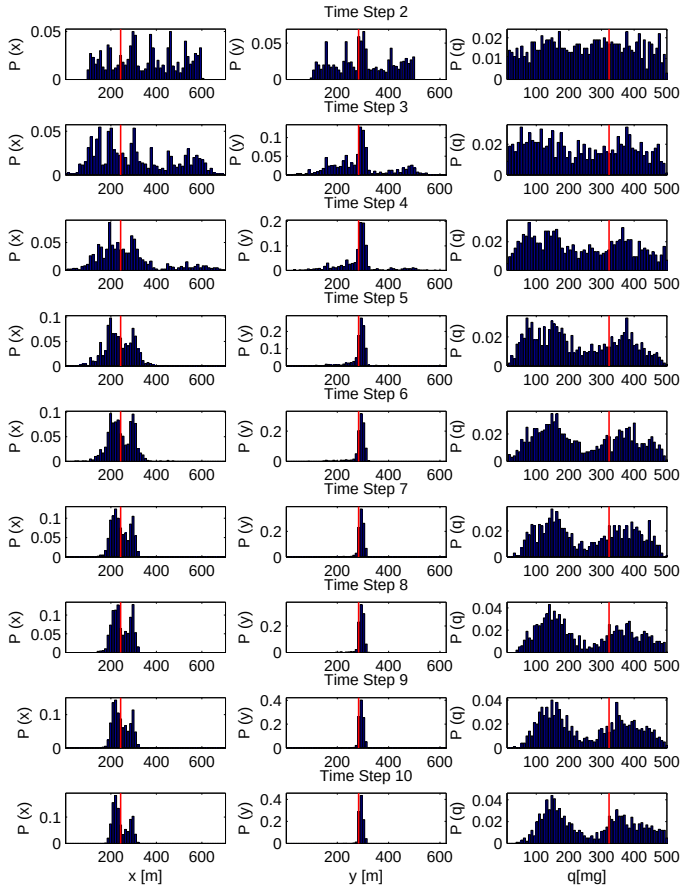
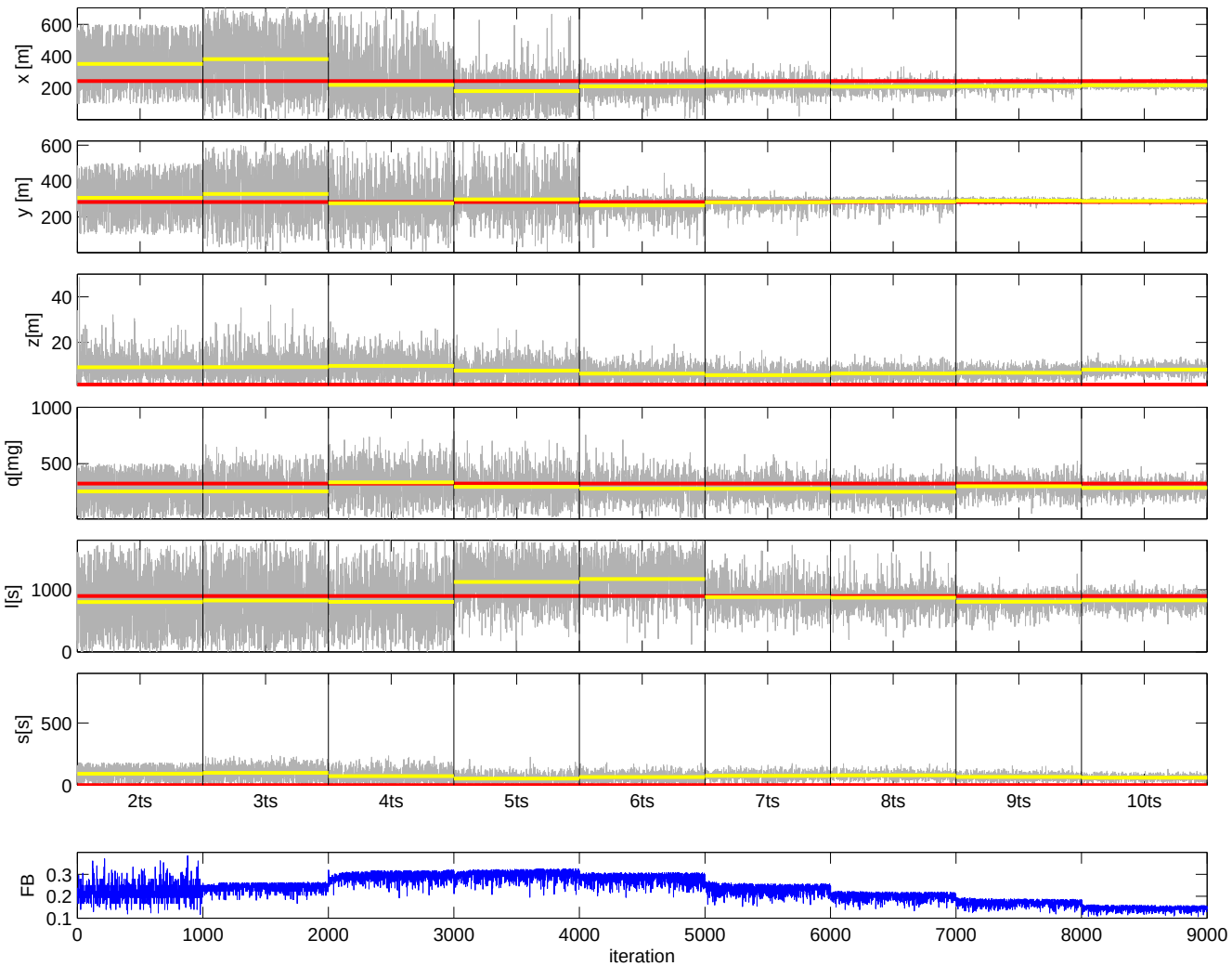


Figure 8: Evolution of the posteriori probability distribution for x, y and q parameters. The red vertical line represents target value of parameters. All plots are taken form for subsequent time steps intervals

Narodowe Centrum Badań Jądrowych
National Centre for Nuclear Research
ŚWIERK

MANHAZ

Results of DAPPLE reconstruction experiment



Results of DAPPLE reconstruction experiment

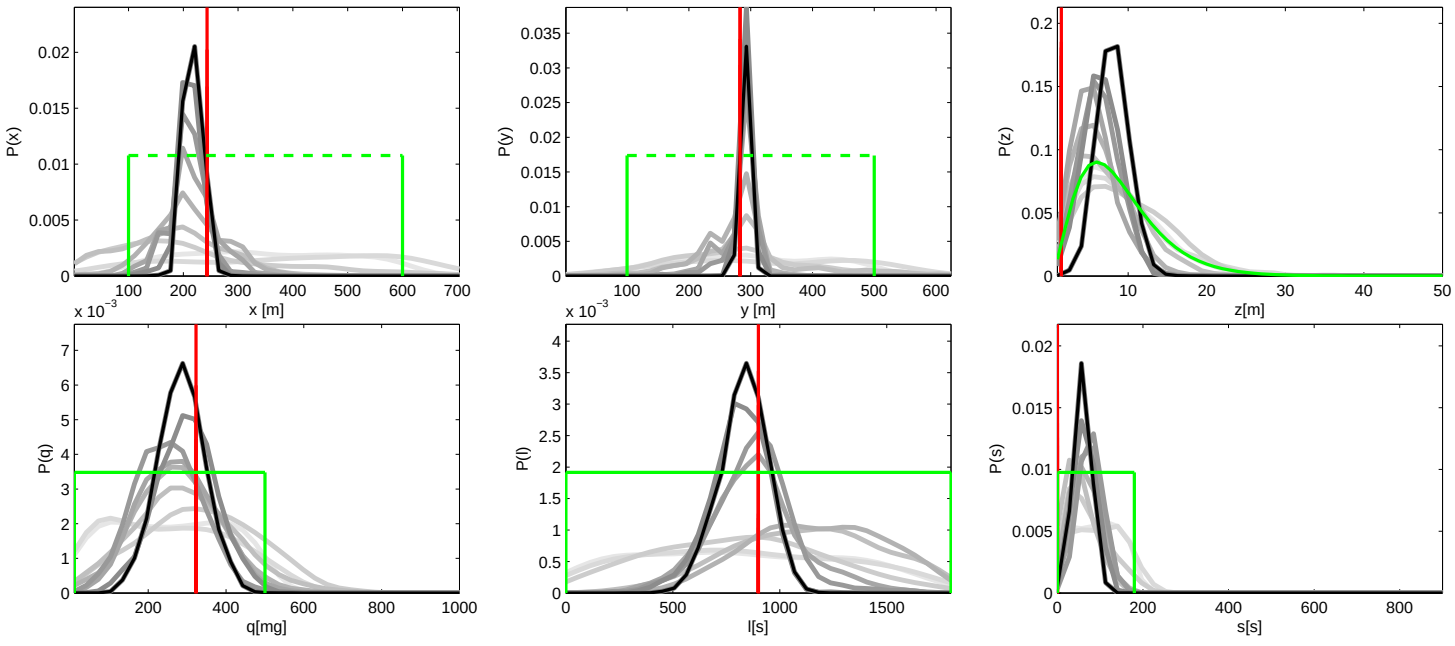


Figure 10: Prior and marginal posterior probability densities for the joint inference of six parameters.

Results of DAPPLE reconstruction experiment

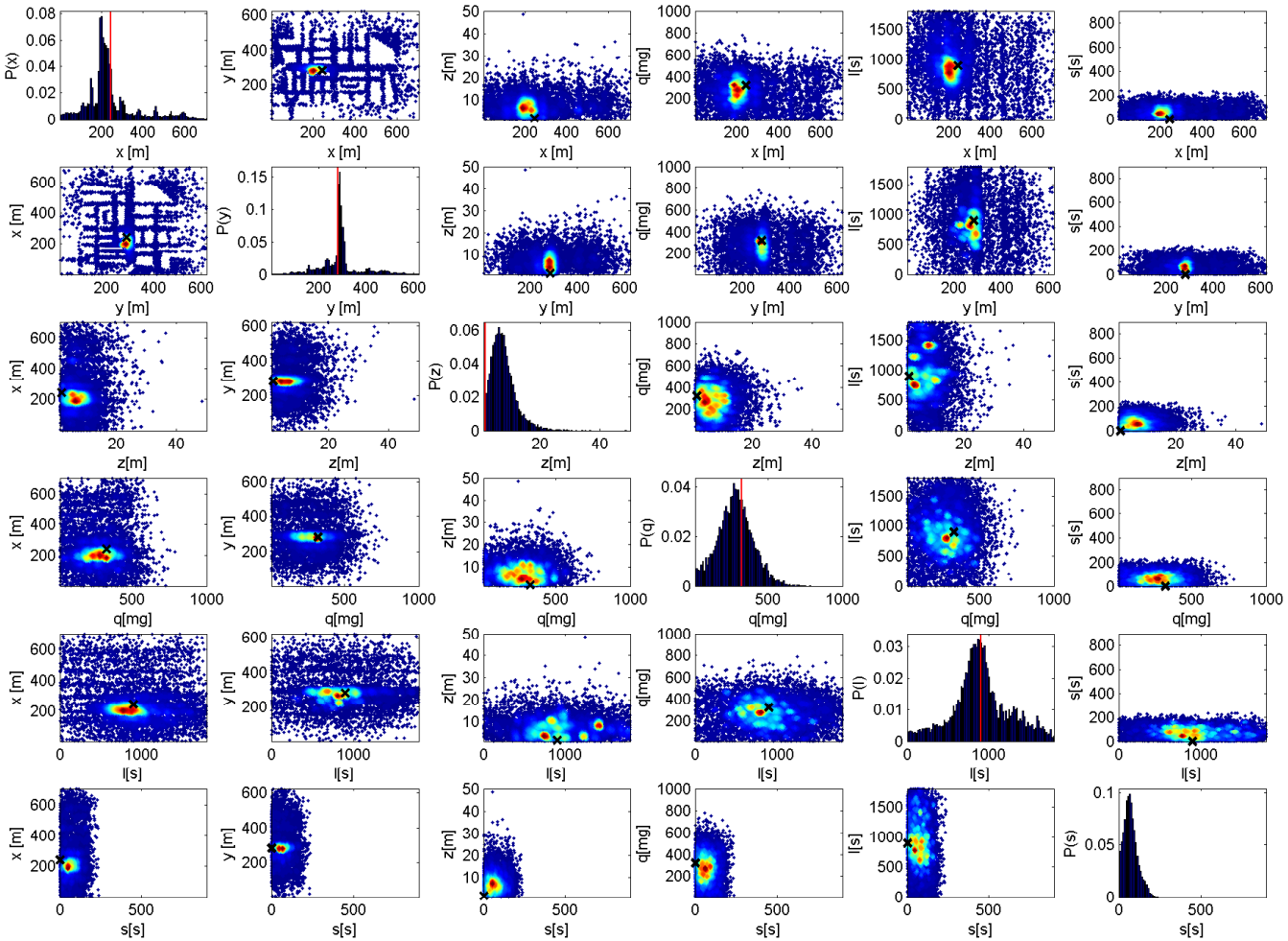


Figure 11: The bivariate and marginal posterior distributions for all searched parameters $\theta \equiv (x, y, z, q, l, s)$.

Publications:

- Approximate Bayesian Computation Methods in the Identification of Atmospheric Contamination Sources for DAPPLE Experiment, P Kopka, A Wawrzynczak, M Borysiewicz International Conference on Bayesian Statistics in Action, 161-171
- In review: Identification of atmospheric contamination source parameter in urban area by Approximate Bayesian Computation methodology, Environmental Modelling and Software

Narodowe Centrum Badań Jądrowych
National Centre for Nuclear Research
ŚWIERK

MANHAZ

THANK YOU
FOR YOUR ATTENTION

Świerk, 10.10.2017