Source Term Estimation For An Unknown Number of Sources

Jon Forster and Philip Wilson
University of Southampton

ADMLC Forum – 5 May 2011

Acknowledgement: This research was sponsored by EPSRC and Dstl/MoD

The problem

Source term estimation when:

- there are multiple sources and
- the number of sources is unknown

Data are a dynamically evolving series of measurements from (downwind) detectors.

The approach to estimation will be fully probabilistic (Bayesian), providing both estimates and associated uncertainties.
Bayesian statistical inference is an inverse probability approach, based on the philosophy that all uncertainty statements should be probabilistic.

Let θ represent the quantities about which inference is required (source term parameters and number of sources) and $D_T = (d_1, \ldots, d_T)$ the data up to and including time T.
Then inference is based on the *posterior* probability distribution for θ *given* the observed data, obtained via *Bayes's theorem* as

$$p(\theta|D_T) = \frac{p(D_T|\theta)p(\theta)}{\int p_t(D_T|\theta)p(\theta)\mathrm{d}\theta}$$

so to compute $p(\theta|D_T)$ we only need the likelihood (model) $p(D_T|\theta)$ and $p(\theta)$.

We call $p(\theta)$ the *prior* distribution, as it describes uncertainty about θ without knowledge of (or equivalently *prior* to observing) any D.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

The statistical model describes the probability distribution of data observations $D_T = (d_{ik})$ given a known release scenario. For example (for single release), $\theta = (x, y, m, t)$:

- Robins et al (2009) use independent

$$p(d_{ik}|\theta) = p_0\delta(d_{ik}) + (1 - p_0)\phi(d_{ik}; \mu_{ik}[\theta], \sigma^2_{ik}[\theta])$$

  where $p_0 = \Phi(0; \mu_{ik}[\theta], \sigma^2_{ik}[\theta])$, ϕ and Φ are normal density and distribution functions and $\mu_{ik}[\theta]$ and $\sigma^2_{ik}[\theta]$ are obtained from a computer dispersion model.

- Huang et al (2010) use independent

$$p(d_{ik}|\theta) = \phi(d_{ik}; \mu_{ik}[\theta], \sigma^2)$$

  where $\mu_{ik}[\theta, a, b, c, v_x, v_y] =$

$$I[t_k > t]\frac{am}{t_k - t}\exp\left\{-b(t_k - t) - \frac{(x_i - v_x(t_k - t) - x)^2 + (y_i - v_y(t_k - t) - y)^2}{c(t_k - t)}\right\}.$$

- Delle Monache et al (2008) use independent

$$p(d_{ik}|\theta) = \phi(\log\{d_{ik}\}; \log\{\mu_{ik}[\theta]\}, \sigma^2)$$

where $\mu_{ik}[\theta]$ is obtained from a computer dispersion model.

- In our simulations, we use

$$p(d_{ik}|\theta) = p_0\delta(d_{ik}) + (1-p_0)\phi(d_{ik}; \mu_{ik}[\theta], \sigma^2\mu_{ik}[\theta]^2)$$

where

$$p_0 = \Phi(0; \mu_{ik}[\theta], \sigma^2\mu_{ik}[\theta]^2)$$

and

$$\mu_{ik}[\theta, a, v_x, v_y] = \frac{am(1+0.0004(t_k-t))(1+0.0015(t_k-t))^{1/2}}{(t_k-t)^3} \times \exp\left\{-\frac{(x_i-v_x(t_k-t)-x)^2+(y_i-v_y(t_k-t)-y)^2}{0.22(t_k-t)(1+0.0004(t_k-t))^{-1}}\right\}.$$

Multiple sources

Generally, models have:

- 'Measurement error' component $p(d_{ik}|\theta)$ (usually independent) requiring mean concentration $\mu_{ik}[\theta]$
- Dispersion model component for $\mu_{ik}[\theta]$

For multiple sources, we have

$$\mu_{ik}[\theta] \equiv \mu_{ik}[\theta^1, \ldots, \theta^S] = \sum_{s=1}^{S} \mu_{iks}[\theta^s]$$

where s indexes source (with measurement variance also adjusted appropriately)

Bayesian estimation in complex models generally proceeds using *Markov Chain Monte Carlo (MCMC)* methods. Such methods have proved very effective.

MCMC computation provides posterior summaries, by *generating* a *dependent* sample $\{\theta_1, \ldots, \theta_J\}$ from the posterior distribution of interest. Then, any posterior expectation can be estimated by the corresponding Monte Carlo sample mean, densities can be estimated from samples etc.

Standard (Metropolis-Hastings) 'recipe' generates proposal θ' given current θ_j using arbitrary density $g(\theta'|\theta_j)$ and accepting $\theta_{j+1} = \theta'$ with probability

$$\alpha = \min\left\{1, \frac{p(\theta'|D_T)g(\theta_j|\theta')}{p(\theta_j|D_T)g(\theta'|\theta_j)}\right\}.$$

Otherwise $\theta_{j+1} = \theta_j$. Often, we use symmetric $g(\theta_j|\theta') = g(\theta'|\theta_j)$.

When the object of inference θ is of *variable dimensionality* (as when we have an unknown number of sources) then we use *reversible jump* MCMC.

For dynamically evolving data sets, we have

$$p(\theta|D_{T+1}) = \frac{p(d_{T+1}|\theta)p(\theta|D_T)}{\int p(d_{T+1}|\theta)p(\theta|D_T)\mathrm{d}\theta}$$

So a Monte Carlo sample from $p(\theta|D_{T+1})$ can be obtained by *weighting* a sample $\{\theta_1, \ldots, \theta_J\}$ from $p(\theta|D_T)$ using weights

$$w_j = \frac{p(d_{T+1}|\theta_j)}{\sum_\ell p(d_{T+1}|\theta_\ell)}.$$

In such examples, it can be effective to run a *population* of MCMC samplers, reweighting (and possibly resampling) when new data arrives and hence the posterior changes.

- Delle Monache et al (2008) use standard MCMC for a single source
- Robins et al (2009) use population MCMC for a single source
- Huang et al (2010) use standard MCMC for multiple sources (potential difficulties with *labelling*)

We combine population Monte Carlo with reversible jump MCMC to allow for an unknown number of sources (Jasra, Holmes and Stephens, 2005).

The critical aspect is the construction of efficient *jump proposals* to move between scenarios θ with different numbers of sources.

Proposals are constructed by considering the move from K to $K+1$ sources. The corresponding reverse move is then largely defined.

- Naive birth/death: The extra source term parameters are proposed from the prior distribution.
- Split merge: An existing source is split into two in such a way as to 'match' the main characteristics at the detectors.

$$x_1 = x_0 - dx_1 \qquad x_2 = x_0 - dx_2 \qquad y_1 = y_0 + dy_1 \qquad y_2 = y_0 - dy_2$$

$$(t_2 - t_1)/2 = dt \qquad m_2 dy_2 = m_1 dy_1 \qquad \frac{m_0}{\sigma(t_0)} = \frac{m_1}{\sigma(t_1)} + \frac{t_2}{\sigma(t_2)}$$

$$(t_1 + t_2)/2 = t_0 + v^{-1}(x_0 - x_1/2 - x_2/2)$$

where $\sigma(t') = (T - t')^3 (1 + 0.0004(T - t'))^{-1} (1 + 0.0015(T - t'))^{-1/2}$.

- Population: The proposal is constructed using those population members which already have $K+1$ sources.

A simulation

We used a (true, but assumed unknown) 3-source release with

$$(t, x, y, m) = \begin{cases} (4, 410, 150, 400) \\ (5, 350, -150, 350) \\ (6, 290, 0, 300) \end{cases}$$

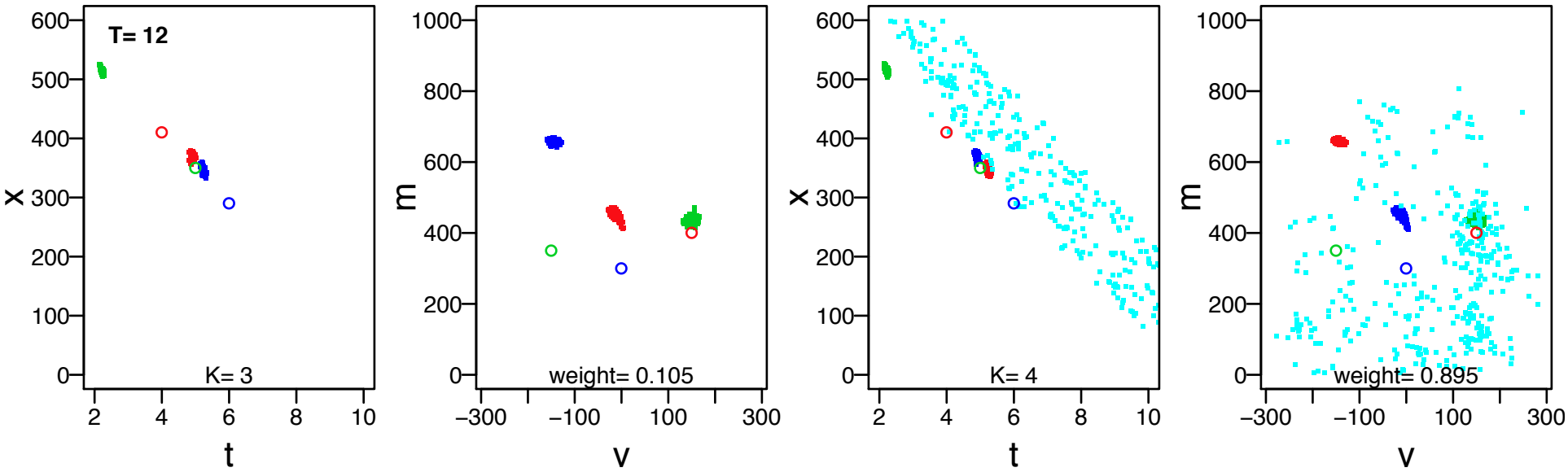
The prior was uniform over $-10 < t < T$, $0 < x < 1000$, $-300 < y < 300$ and $0 < m < 1000$, where T denotes the time of the latest data.

The sensors were in 2 columns of 12 sensors each, located at $x = 0$ and $y = -287.5, -237.5, \ldots 262.5$m for one column and at $x = -50$ and $y = -262.5, -212.5, \ldots 287.5$m for the other.

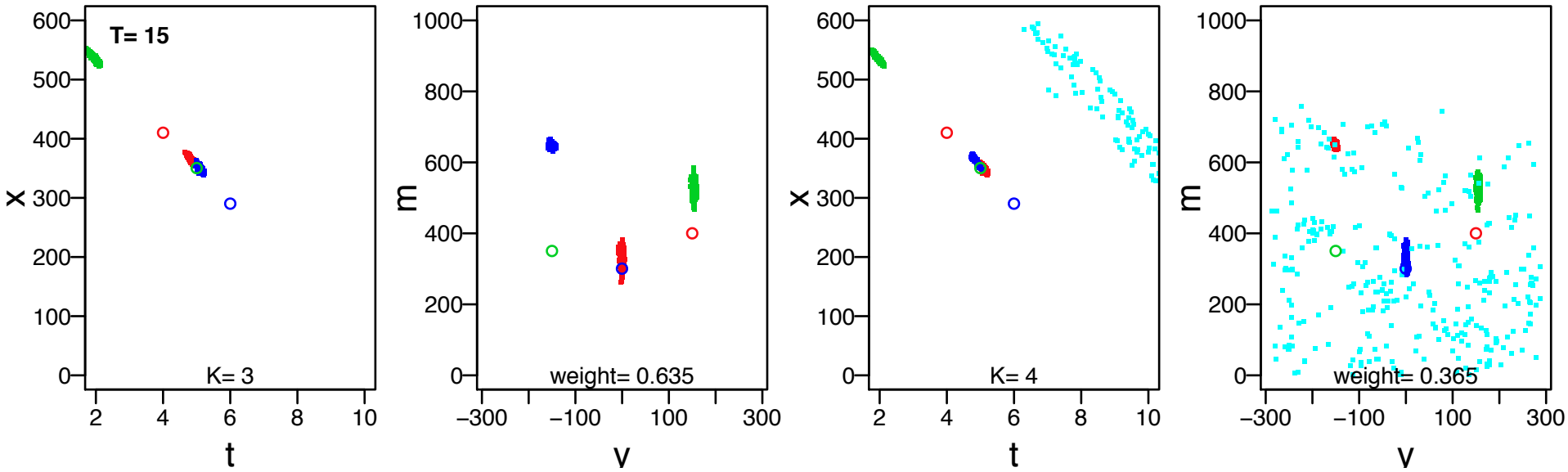
Other parameters include a wind speed of 1ms^{-1}, background concentration mean 2×10^{-6} gm^{-3} and standard deviation 4×10^{-7} gm^{-3},

The population MCMC used 25 chains, 4000 likelihood calculations per time period and a maximum 4 sources.

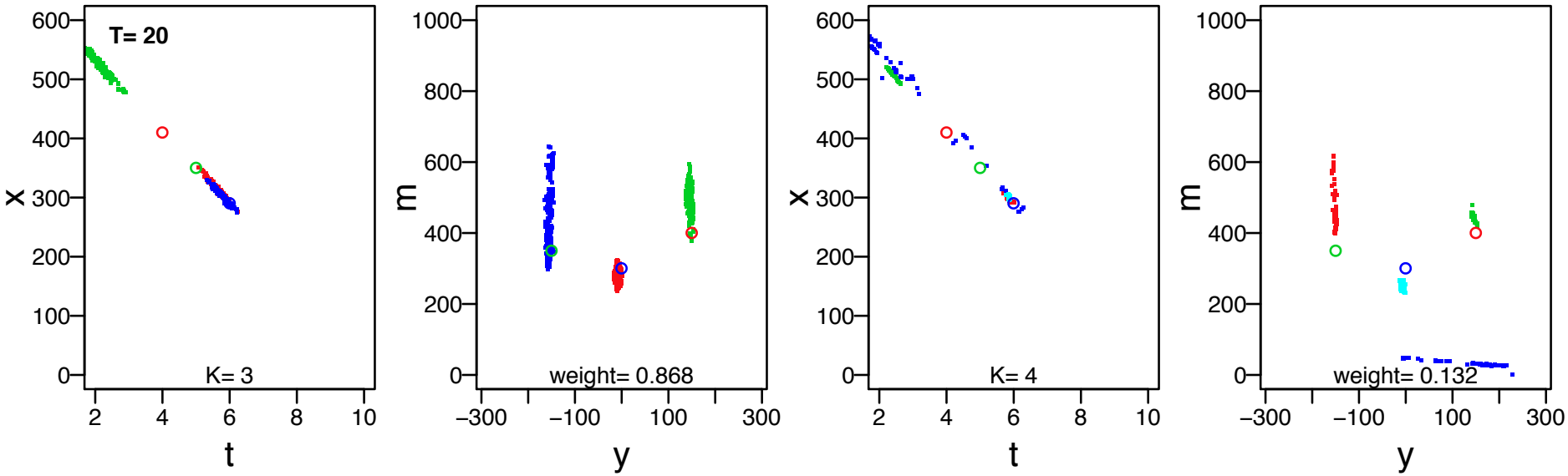
Results (1)



Results (2)

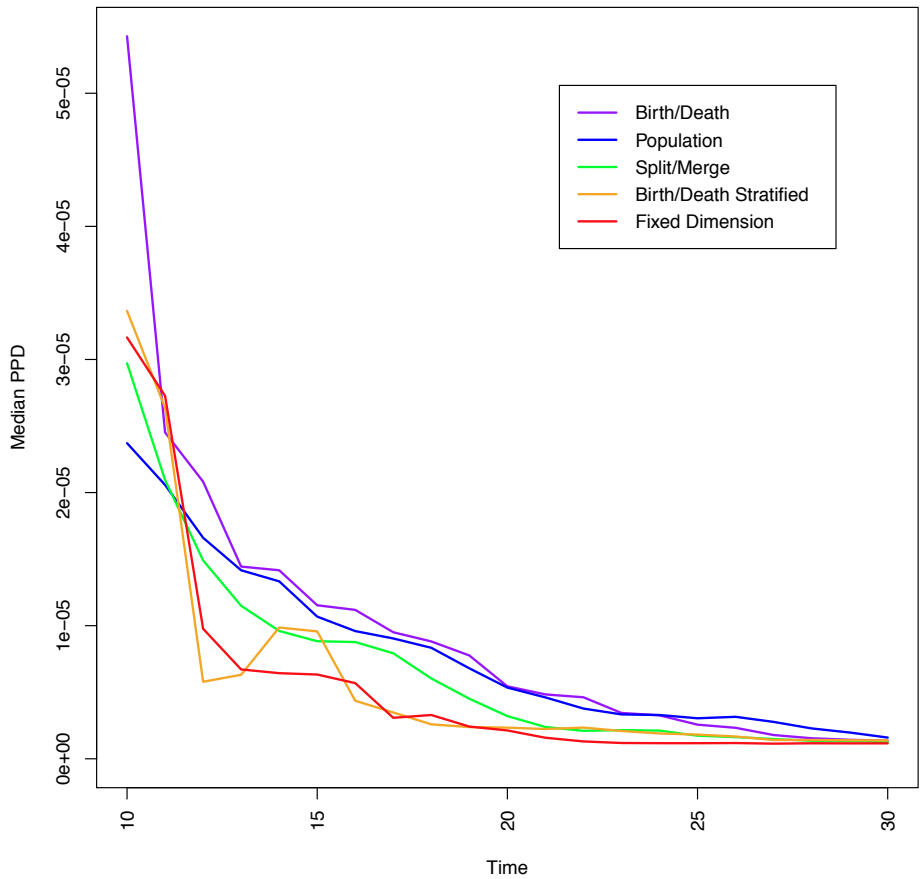


Results (3)



Assessed by considering 'model-averaged' posterior-predictive deviance

$$PPD = \sum_{j} \sum_{ik} (d_i k - \mu_{ik}[\theta_j])^2$$



Future work/challenges

- Assessing robustness across a range of dispersion models
- Real data modelling
- Forward prediction
- ...