Combining syndromic surveillance and ILI data using particle filter for epidemic state estimation

Taesik Lee & Hayong Shin

Flexible Services and Manufacturing Journal

ISSN 1936-6582
Volume 28
Combined 1-2

Flex Serv Manuf J (2016) 28:233-253
DOI 10.1007/s10696-014-9204-0

Springer

Flex Serv Manuf J (2016) 28:233–253
DOI 10.1007/s10696-014-9204-0

Combining syndromic surveillance and ILI data using particle filter for epidemic state estimation

Taesik Lee · Hayong Shin

Published online: 11 October 2014

Abstract Designing effective mitigation strategies against influenza outbreak requires an accurate prediction of a disease's future course of spreading. Real time information such as syndromic surveillance data and influenza-like-illness (ILI) reports by clinicians can be used to generate estimates of the current state of spreading of a disease. Syndromic surveillance data are immediately available, in contrast to ILI reports that require data collection and processing. On the other hand, they are less credible than ILI data because they are essentially behavioral responses from a community. In this paper, we present a method to combine immediately-available-but-less-reliable syndromic surveillance data with reliable-but-time-delayed ILI data. This problem is formulated as a non-linear stochastic filtering problem, and solved by a particle filtering method. Our experimental results from hypothetical pandemic scenarios show that state estimation is improved by utilizing both sets of data compared to when using only one set. However, the amount of improvement depends on the relative credibility and length of delay in ILI data. An analysis for a linear, Gaussian case is presented to support the results observed in the experiments.

Keywords Epidemic · Syndromic surveillance · Particle filter · Data fusion

1 Introduction

Designing containment and mitigation strategies upon an epidemic outbreak is of critical public health interest. Typical examples of mitigation strategies include

T. Lee · H. Shin (✉)
Department of Industrial and Systems Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Korea
e-mail: hyshin@kaist.ac.kr

clinical interventions (e.g., mass vaccination) and non-clinical interventions (e.g., social distancing) (Ferguson et al. 2005; WHO 2013). All of these strategies incur significant direct and indirect costs to society, and it is therefore important to develop the most effective strategy possible against epidemic outbreaks. Designing mitigation strategies may be done in a preparatory planning phase or in real-time. For preparatory planning, a hypothetical scenario of an epidemic outbreak is assumed and the effectiveness of various containment strategies is assessed. Understanding gained from these analyses is used to develop epidemic disease response plans, such as securing flu vaccines and antiviral stockpiles. In real-time decision making, public health authorities predict the progress of disease spread based on its current state. If the prediction warrants intervention, action plans such as vaccination campaigns or school closures are quickly developed and executed.

The crucial components to a successful response to a disease outbreak are quickly detecting an outbreak of an epidemic and accurately predicting the future course of spreading of the disease. Early detection of a disease outbreak is known to have a critical impact on the effectiveness of mitigation efforts (Ferguson et al. 2005; Gensheimer et al. 1999; Longini et al. 2005), and there has been much research work on detecting outbreaks of epidemic diseases [e.g., Dukic et al. (2012), Que and Tsui (2011), Reis et al. (2007), Singh et al. (2010)]. Following the detection of a disease outbreak, the magnitude and speed of the spread of an epidemic should be accurately estimated for public health authorities to develop mitigation strategies. In particular, two quantities are important indicators for potential impact of an epidemic: I_{peak} and T_{peak}. I_{peak} is the maximum number of simultaneously infected persons in the community. T_{peak} is the time (e.g., x days after the outbreak) of the peak infection.

Prediction of the future course of disease spread requires a high-fidelity disease spread model that, given the current state of spread of an epidemic, produces quantitative estimates on how rapid and severe the epidemic will be. Largely, there are two approaches to model the spreading of an epidemic disease: equation-based models and simulation-based models. In equation-based models, a community consists of a few subpopulation groups, each of which corresponds to different stages of a disease. Disease spreading in a community is then modeled by the flows of population from one group to another, and a set of differential equations is used to compute the change of the population size in each group. The classic equation-based models assume that the underlying population is homogeneous and well-mixed, describing the overall dynamics in an average sense. Many variants have been developed, and Hethcote (2000) provides a comprehensive review of equation-based epidemic models. The other stream of disease spreading models use simulation as their basis [e.g., Bisset et al. (2009), Chao et al. (2010), Eubank et al. (2004)]. Many of these models are built on the agent-based modeling concept. Individual persons (or small group of individuals) are modeled to interact with other individuals and respond to changes in their environment, the *world* they are living in. These models are capable of representing the heterogeneous nature of the underlying population, and provide a relatively easy means for incorporating the complexity of the real world environment. Regardless of whether the model is

equation-based or simulation-based, once we have a high-fidelity disease spread model, we can build a variety of mitigation strategies into the model and assess their effectiveness.

Quality of the prediction of an epidemic spread also depends on the accuracy of the information on the current state of the system. Information on the current state—i.e., how many people have been infected as of today—is provided to an epidemic spread model to develop prediction and identify the most likely scenario. If the information on the current state is wrong, prediction even by the most accurate epidemic model will be wrong, leading to suboptimal response decisions. This paper addresses the problem of state estimation for the spread of an epidemic using a nonlinear stochastic filtering technique.

One of the tools commonly used for epidemic state estimation and prediction is a recursive Bayesian state estimation technique, and many examples of research using this technique can be found in the literature [e.g., Dukic et al. (2012), Jegat et al. (2008), Ong et al. (2010), Vidal Rodeiro and Lawson (2006), Shaman and Karspeck (2012), Skvortsov and Ristic (2012)]. Bayesian state estimation assumes some knowledge on the underlying dynamics of a system (system model), and recursively updates the degree of belief in system states by using sequentially available observation data. Since in most cases the underlying model is not fully known—i.e., epidemic parameters in the model are typically unknown, these methods often estimate epidemic parameters as well as state variables. For example, Dukic et al. (2012), Ong et al. (2010), Skvortsov and Ristic (2012) use epidemic equations as a system model, and formulate a Bayesian filtering problem to estimate epidemic parameters and state variables. In the method developed in Dukic et al. (2012), emphasis is placed on learning of the epidemic parameters. Skvortsov and Ristic (2012), on the other hand, focuses on the inhomogeneous mixing of a population by using a stochastic epidemic model. Our work is also based on recursive Bayesian state estimation, a particle filter in particular, as the underlying modeling technique, but it differs in that we concentrate on the question of combining multiple types of surveillance information.

Our study is motivated by the fact that two types of surveillance data, particularly in terms of their timeliness, are available to the estimation task (Dailey et al. 2007). For an epidemic flu, a traditional source of information for the current epidemic state is Influenza-Like-Illness (ILI) data from a government health agency such as Centers for Disease Control and Prevention (CDC) (FluView 2013; Influenza Weekly Report 2013). ILI refers to a medical diagnosis of a possible influenza case. For example, Korea CDC defines ILI as a sudden fever over 38 °C along with cough or throat pain (Influenza Surveillance 2014). ILI data are gathered from care providers as patients with relevant symptoms visit hospitals and clinics. These data are used as an indicator for the number of people infected with epidemic flu. ILI data have some uncertainty (Jegat et al. 2008); they are based on a diagnosis by symptoms only, not confirmed by lab tests, and thus do not distinguish those infected with flu versus those who have symptoms by other causes. In addition, there are always some flu patients who do not visit physicians. Nevertheless, ILI data are generated from reports by physicians based on their medical diagnosis, and thus are considered a reasonably reliable indicator for flu activity in the community.

That said, there is one important shortcoming when using ILI data for estimating the current state of disease spread. Generally, it takes 1–2 weeks to gather and process data from a large surveillance network. In Korea, for example, about 250 hospitals and labs participate in the surveillance network to report ILI cases to Korea CDC, and the CDC's weekly report is released on the 10th day from the end of the reported week. Owing to this this 1–2 weeks lag, ILI data are outdated by the time they are released. They do not provide real-time information on the current state of flu spreading.

Another source of information is syndromic surveillance data, which are recently getting significant attention from the research community (Chen et al. 2010; Chew and Eysenbach 2010; Ginsberg et al. 2009; Lampos et al. 2010; Skvortsov and Ristic 2012). Examples of syndromic surveillance data include school or work absenteeism, over-the-counter drug sales, and search engine queries (Henning 2004). With proper tools and systems, syndromic surveillance data can be made available in almost real-time, which offers an advantage for making timely state estimation as evidenced by the well-popularized Google Flu Trends (Ginsberg et al. 2009). However, the data are based on the "syndromes," which are largely the population's behavioral responses, and thus have lower credibility than ILI data.

As the two sets of data compliment each other, we naturally expect that combining the two will improve the estimation outcomes. The main goal of this paper is to understand and characterize the improvement in the state estimation from assimilating the two datasets. In other words, we investigate the following questions: given two sets of observation data—immediately-available-but-uncertain data and credible-but-delayed data, which one would yield a better state estimation? Is it always better to use both sets of data than using only one, and if so, how much better? To investigate these questions, we develop a method to combine reliable-but-time-delayed ILI data with less-reliable-but-immediately-available syndromic surveillance data. Our method uses a particle filter with a compartmental epidemic model, which is similar to Dukic et al. (2012), Skvortsov and Ristic (2012). In addition, we use a modified version of the out-of-sequence-measurement particle filter by Orton and Marrs (2005) to handle the delayed measurement data.

This paper is structured as follows: Sect. 2 briefly introduces the basics of the particle filter algorithm and its extension, the out-of-sequence-measurement (OOSM) particle filter. In Sect. 3, the system model and measurement model used in the proposed particle filter are presented. Section 4 discusses the proposed particle filter algorithm and its pseudo-code for implementation. In Sect. 5, experimental results from a hypothetical epidemic outbreak scenario are presented. Section 6 provides an analytic explanation for the patterns observed in the experimental results. Finally, Sect. 7 concludes the paper.

2 Background

In this section, we briefly discuss the basics of the particle filter technique. Section 2.1 presents the principles of the particle filter, and Sect. 2.2 describes a variant of a basic particle filter that handles out-of-sequence measurement (OOSM) data. For

more details on particle filter and OOSM particle filter, refer to Ducet and Johansen (2013), Orton and Marrs (2005), Ristic et al. (2004).

2.1 Particle filter

A particle filter is a recursive Bayesian filter, used for estimating the state of a dynamic system where its state variables are not directly observable. The technique is particularly useful for non-linear, non-Gaussian state estimation problems. A particle filter combines a series of measurement data with a known system dynamics model to update beliefs on the true state of the system.

To formally describe the technique, consider a system whose dynamics is described by the following system model:

$$\mathbf{x_k} = \mathbf{f_{k-1}}(\mathbf{x_{k-1}}) + \mathbf{v_{k-1}} \tag{1}$$

where $\mathbf{x_k}$ denotes a state vector at time index k, f_{k-1} is a possibly non-linear and time-varying function, and $\mathbf{v_{k-1}}$ represents process noise. Suppose for this system a set of observation data $\mathbf{z_k}$ are measured at each time index k, and $\mathbf{z_k}$ is related to $\mathbf{x_k}$ by the following observation model:

$$\mathbf{z_k} = \mathbf{h_k}(\mathbf{x_k}) + \mathbf{w_k} \tag{2}$$

where h_k is a possibly non-linear and time-varying function, and $\mathbf{w_k}$ denotes measurement noise.

Let $\mathbf{z_{1:k}}$ denote a series of measurement data up to k. Then, a recursive Bayesian filter seeks to construct a posterior pdf $p(\mathbf{x_k}|\mathbf{z_{1:k}})$ as an estimate for the true state of the system. This is done in two stages - *prediction* and *update*. Suppose that a posterior pdf at $(k-1)$, $p(\mathbf{x_{k-1}}|\mathbf{z_{1:k-1}})$ is known. The prediction stage computes the distribution of a *predicted* system state at k using our knowledge of the system model (1). Conceptually, the prediction stage is written as follows:

$$p(\mathbf{x_k}|\mathbf{z_{1:k-1}}) = \int \mathbf{p}(\mathbf{x_k}|\mathbf{x_{k-1}})\mathbf{p}(\mathbf{x_{k-1}}|\mathbf{z_{1:k-1}})\mathbf{dx_{k-1}} \tag{3}$$

where $p(\mathbf{x_k}|\mathbf{x_{k-1}})$ can be determined from the system model (1). When new measurement data at k become available, the update stage is carried out to compute a posterior pdf $p(\mathbf{x_k}|\mathbf{z_{1:k}})$:

$$p(\mathbf{x_k}|\mathbf{z_{1:k}}) = \frac{\mathbf{p}(\mathbf{z_k}|\mathbf{x_k})\mathbf{p}(\mathbf{x_k}|\mathbf{z_{1:k-1}})}{\mathbf{p}(\mathbf{z_k}|\mathbf{z_{1:k-1}})} \propto \mathbf{p}(\mathbf{z_k}|\mathbf{x_k})\mathbf{p}(\mathbf{x_k}|\mathbf{z_{1:k-1}}) \tag{4}$$

$p(\mathbf{z_k}|\mathbf{x_k})$ is the likelihood of measurement $\mathbf{z_k}$ if the system state were $\mathbf{x_k}$, and $p(\mathbf{z_k}|\mathbf{x_k})$ can be determined from the observation model (2).

When f_k and h_k are linear and $\mathbf{v_k}$ and $\mathbf{w_k}$ are independent white Gaussian, (3) and (4) can be solved exactly; a Kalman filter provides an optimal filtering solution in an analytic form. In most other cases, however, (3) and (4) cannot be solved exactly, and some form of approximation is required. A particle filter is one such method. A

particle filter uses a sample representation for a posterior density $p(\mathbf{x_k}|\mathbf{z_{1:k}})$, which can be written as:

$$p(\mathbf{x_k}|\mathbf{z_{1:k}}) \simeq \sum_{\mathbf{i=1}}^{\mathbf{N_s}} \omega_{\mathbf{k}}^{\mathbf{i}} \delta(\mathbf{x_k} - \mathbf{x_k^i}) \tag{5}$$

where $\delta(\mathbf{x} - \mathbf{x^i})$ is 1 for $\mathbf{x} = \mathbf{x^i}$, and 0 otherwise. ω_k^i is a weight assigned to sample $\mathbf{x_k^i}$, and N_s is the number of samples (i.e., *particles*). Following the principle of importance sampling, ω_k^i can be written in the following recursive form:

$$\omega_k^i \propto \omega_{k-1}^i \frac{p(\mathbf{z_k}|\mathbf{x_k^i})\mathbf{p}(\mathbf{x_k^i}|\mathbf{x_{k-1}^i})}{q(\mathbf{x_k^i}|\mathbf{x_{0:k-1}^i}, \mathbf{z_{1:k}})} \tag{6}$$

$q(\cdot)$ is an importance or proposal density from which the samples $\mathbf{x_k^i}$ are drawn.

A generic particle filter algorithm, often referred to as Sequential Importance Resampling (SIR) algorithm, then proceeds as follows:

1. For each particle $\mathbf{x_{k-1}^i}$, draw $\mathbf{x_k^i}$ from $q(\mathbf{x_k^i}|\mathbf{x_{0:k-1}^i}, \mathbf{z_{1:k}})$
2. Evaluate the importance weight ω_k^i for $\mathbf{x_k^i}$ by (6)
3. Normalize the weights, $\hat{\omega}_k^i = \omega_k^i / \sum_{j=1}^{N_s} \omega_k^j$
4. Draw a new sample set $\mathbf{x_k^{i*}}$ so that $Prob(\mathbf{x_k^{i*}} = \mathbf{x_k^j}) = \hat{\omega}_{\mathbf{k}}^{\mathbf{j}}$; $\mathbf{x_k^{i*}}$ is assigned an equal weight of $1/N_s$

As new measurement data $\mathbf{z_k}$ arrive, the SIR particle filtering algorithm estimates the posterior density $p(\mathbf{x_k}|\mathbf{z_{1:k}})$ by updating the sample set that approximates the true posterior. It is noteworthy that if we use $p(\mathbf{x_k^i}|\mathbf{x_{k-1}^i})$ for the importance density $q(\mathbf{x_k^i}|\mathbf{x_{0:k-1}^i}, \mathbf{z_{1:k}})$, then (6) reduces to $\omega_k^i \propto \omega_{k-1}^i p(\mathbf{z_k}|\mathbf{x_k^i})$. The fourth step is referred to as a *resampling* procedure. Resampling is necessary to prevent the weights from being concentrated to a few particles. When weight concentration occurs, most other particles become irrelevant to the computation as the algorithm proceeds, rendering the algorithm inefficient and ineffective. This phenomenon is known as the degeneracy problem. A common practice for resampling is to carry out resampling only when the effective sample size is smaller than some threshold. Step (4) here simply assumes that resampling is executed at every time step.

2.2 OOSM particle filter

Out-of-sequence measurements (OOSMs) refer to measurement data that arrive with delay. The delayed arrival of OOSM means that they represent the system state at some point in the past. Let t_k denote the actual time instant when $\mathbf{z_k}$, measurement data arriving in the k^{th} sequence, have been taken. When $t_k > t_{k-1}$, $\mathbf{z_k}$ is in sequence, and when $t_k < t_{k-1}$, then it is out of sequence. The OOSM particle filter provides a means to update the importance weight upon an arrival of an out-of-sequence measurement.

Suppose we have a series of in-sequence measurement data, $\mathbf{z_{1:k-1}}$, and at k, out-of-sequence data $\mathbf{z_k}$ arrives. Let b and a be the time indices immediately before and after t_k. That is, $t_b < t_k < t_a$ and $a = b + 1$. The joint posterior density $p(\mathbf{x_{0:k-1}}|\mathbf{z_{1:k-1}})$ can be expanded as follows:

$$\begin{aligned} p(\mathbf{x_{0:k-1}}|\mathbf{z_{1:k-1}}) &= \frac{p(\mathbf{z_{k-1}}|\mathbf{x_{k-1}})\mathbf{p}(\mathbf{x_{k-1}}|\mathbf{x_{k-2}})\mathbf{p}(\mathbf{z_{1:k-2}})}{p(\mathbf{z_{1:k-1}})} \times p(\mathbf{x_{0:k-2}}|\mathbf{z_{1:k-2}}) \\ &= \frac{p(\mathbf{z_{k-1}}|\mathbf{x_{k-1}})\mathbf{p}(\mathbf{x_{k-1}}|\mathbf{x_{k-2}})\mathbf{p}(\mathbf{z_{1:k-2}})}{p(\mathbf{z_{1:k-1}})} \times \cdots \\ &\quad \times \frac{p(\mathbf{z_a}|\mathbf{x_a})\mathbf{p}(\mathbf{x_a}|\mathbf{x_b})\mathbf{p}(\mathbf{z_{1:b}})}{p(\mathbf{z_{1:a}})} \times \frac{p(\mathbf{z_b}|\mathbf{x_b})\mathbf{p}(\mathbf{x_b}|\mathbf{x_{b-1}})\mathbf{p}(\mathbf{z_{1:b-1}})}{p(\mathbf{z_{1:b}})} \\ &\quad \times p(\mathbf{x_{0:b-1}}|\mathbf{z_{1:b-1}}) \end{aligned} \tag{7}$$

When the out-of-sequence measurement $\mathbf{z_k}$ arrives, (7) is modified to incorporate the new dependence relationships between $\mathbf{x_b}, \mathbf{x_k}$, and $\mathbf{x_a}$ by rewriting (7) with the insertion of $\mathbf{z_k}$. The new posterior $p(\mathbf{x_{0:k}}|\mathbf{z_{1:k}})$ is,

$$\begin{aligned} p(\mathbf{x_{0:k}}|\mathbf{z_{1:k}}) &= \frac{p(\mathbf{z_{k-1}}|\mathbf{x_{k-1}})\mathbf{p}(\mathbf{x_{k-1}}|\mathbf{x_{k-2}})\mathbf{p}(\mathbf{z_{1:k-2}})}{p(\mathbf{z_{1:k-1}})} \times \cdots \\ &\quad \times \frac{p(\mathbf{z_a}|\mathbf{x_a})\mathbf{p}(\mathbf{x_a}|\mathbf{x_k})\mathbf{p}(\mathbf{z_{1:k}})}{p(\mathbf{z_{1:a}})} \times \frac{p(\mathbf{z_k}|\mathbf{x_k})\mathbf{p}(\mathbf{x_k}|\mathbf{x_b})\mathbf{p}(\mathbf{z_{1:b}})}{p(\mathbf{z_{1:k}})} \\ &\quad \times \frac{p(\mathbf{z_b}|\mathbf{x_b})\mathbf{p}(\mathbf{x_b}|\mathbf{x_{b-1}})\mathbf{p}(\mathbf{z_{1:b-1}})}{p(\mathbf{z_{1:b}})} \\ &\quad \times p(\mathbf{x_{0:b-1}}|\mathbf{z_{1:b-1}}) \end{aligned} \tag{8}$$

Comparing (7) with (8), we obtain the following recursive relationship:

$$p(\mathbf{x_{0:k}}|\mathbf{z_{1:k}}) = \mathbf{p}(\mathbf{x_{0:k-1}}|\mathbf{z_{1:k-1}}) \times \frac{\mathbf{p}(\mathbf{x_a}|\mathbf{x_k})\mathbf{p}(\mathbf{x_k}|\mathbf{x_b})\mathbf{p}(\mathbf{z_k}|\mathbf{x_k})}{\mathbf{p}(\mathbf{x_a}|\mathbf{x_b})\mathbf{p}(\mathbf{z_{1:k}})} \tag{9}$$

With (9), we now have a weight update equation similar to (6).

$$\omega_k^i \propto \omega_{k-1}^i \frac{p(\mathbf{z_k}|\mathbf{x_k^i})\mathbf{p}(\mathbf{x_k^i}|\mathbf{x_a^i}, \mathbf{x_b^i})}{q(\mathbf{x_k^i}|\mathbf{x_{0:k-1}^i}, \mathbf{z_{1:k}})} \tag{10}$$

It was shown in Orton and Marrs (2005) that the optimal importance density is $p(\mathbf{x_k^i}|\mathbf{x_a^i}, \mathbf{x_b^i}, \mathbf{z_k})$. However, sampling from the optimal importance density function is quite difficult, and Orton and Marrs (2005) suggests $p(\mathbf{x_k^i}|\mathbf{x_a^i}, \mathbf{x_b^i})$ as a tractable approximation. This choice of importance function reduces (10) to $\omega_k^i \propto \omega_{k-1}^i p(\mathbf{z_k}|\mathbf{x_k^i})$.

It turns out that for our problem, we can use a more straightforward implementation of the above OOSM particle filter framework. This is due to an assumption in our hypothetical problem - OOSMs in our problem have a known, fixed lag interval L. A modified version of an OOSM particle filter is presented in Sect. 4.

3 Epidemic model

We consider a hypothetical outbreak of an epidemic where two streams of observation information - syndromic surveillance data and ILI data - arrive sequentially. Syndromic surveillance data are assumed to be immediately available but have high uncertainty, while more reliable ILI data are delayed by a certain time lag. In this section, we first describe the epidemic model that we use in the present study as a system model for the particle filter formulation. We then discuss the two observation data sets and their measurement model.

Using a set of differential equations to describe the spread of epidemic disease has a long history, dating back to the work by Kermack and McKendrick (1927) in the early 20th century. Typically, these models divide the population into a few compartments, and express the rate at which the population from one compartment flows to other compartments by a set of ordinary differential equations. One of the simplest models is the S-I-R model, where S, I, and R denotes susceptible, infectious, and recovered compartments. Susceptible individuals have not yet been infected by the disease, and may become infected by contacting an infectious person. Infectious persons may infect the susceptible persons until they eventually recover from the disease. Once recovered, they acquire immunity against the disease. Let P be the total population size, and let S, I, and R denote the number of individuals in each compartment. The time scale of an epidemic disease is short relative to the time scale of population size change, and $P = S + I + R$ for a constant P. Flows between the three compartments can be described by the following set of equations:

$$\frac{ds}{dt} = -\beta si; \quad \frac{di}{dt} = \beta si - \gamma i; \quad r = 1 - s - i \tag{11}$$

where s, i, and r denote the size of each compartment normalized by the total population size P. β represents the rate of infectious contacts, and γ is the recovery rate, which is the inverse of the average infectious period for the disease.

One important underlying assumption in (11) is that we have a homogeneous population and they are perfectly mixed - i.e., anyone can have contact with anyone in the community with equal probability. This is quite a strong assumption, far from the real world's heterogeneous and intricate contact behaviors (Rahmandad and Sterman 2008). One approach to address this issue is to introduce stochastic fluctuations to the basic S-I-R model. For example, Skvortsov and Ristic (2012) presents a stochastic version of the basic S-I-R model, and the modified S-I-R model incorporates stochastic fluctuations in (11):

$$\begin{aligned} \frac{ds}{dt} &= -\beta i s^{\nu} + \sigma_q \xi \\ \frac{di}{dt} &= \beta i s^{\nu} - \gamma i - \sigma_q \xi + \sigma_\gamma \zeta \\ r &= 1 - s - i \end{aligned} \tag{12}$$

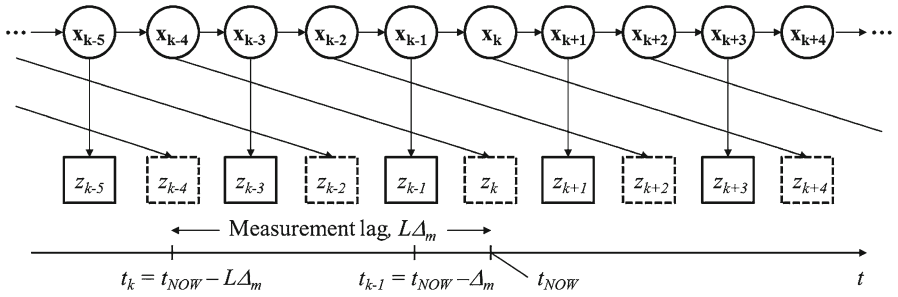

Fig. 1 In-sequence (*solid square*) and out-of-sequence (*dashed square*) measurement data. Out-of-sequence data arriving at t_{NOW} contain information on the system state at $t_{NOW} - L\Delta_m$

$v, \sigma_q\xi$ and $\sigma_\gamma\zeta$ are introduced to account for heterogeneity and stochasticity. ξ and ζ are uncorrelated, white Gaussian noise with zero mean and a unit variance. For the purpose of discussions in this paper, we simplify (12) by assuming that the parameters in the model – $\beta, v, \sigma_q, \gamma$, and σ_γ – are all known constants.

We define a state vector $\mathbf{x} = [\mathbf{s}, \mathbf{i}]^{\mathbf{T}}$, and then we have a discretized state-space model for (12) to use in the particle filter implementation:

$$\begin{Bmatrix} s_{k+1} \\ i_{k+1} \end{Bmatrix} = \begin{Bmatrix} s_k \\ i_k \end{Bmatrix} + \begin{bmatrix} -\beta i_k s_k^v \\ (\beta i_k s_k^v - \gamma i_k) \end{bmatrix} \Delta t + \begin{bmatrix} \sigma_q \xi \\ (-\sigma_q \xi + \sigma_\gamma \zeta) \end{bmatrix} \Delta t^{1/2} \qquad (13)$$

In the analyses in this paper, we assume the parameters in (13) take constant values- $\beta = 0.3, \gamma = 0.1, v = 1.0$, and $\sigma_q = \sigma_\gamma = 0.001$.[1]

We assume that measurement data arrive with a fixed interval Δ_m. We further assume that syndromic surveillance data and ILI data arrive in an alternating sequence. An arrival pattern for the measurement data is shown in Fig. 1.

ILI data have a fixed time-delay of $L\Delta_m$. Suppose that, at $t = t_{NOW}$, we have a new measurement z_k from ILI data reported at the k^{th} sequence. z_k corresponds to the system state at time t_k, which is $(t_{NOW} - L\Delta_m)$. Figure 1 shows an example when $L = 4$. The k^{th} measurement data z_k arrives at t_{NOW}, and it contains information on the system state at $t_{NOW} - 4\Delta_m$.

For a measurement model, we follow Skvortsov and Ristic (2012) to assume the following relationship for both syndromic surveillance data and ILI data:

$$z = b_j i^{\varsigma_j} + \sigma_j \eta_j \qquad (14)$$

where j = synd or ILI to indicate whether a measurement is obtained from syndromic surveillance data or an ILI case report. For simplicity, we assume that b_j and ς_j are known, $b_j = 1.0$ and $\varsigma_j = 1.0$, for both syndromic surveillance data and ILI data. η_j is independent Gaussian noise with a unit variance. σ_j is used to represent reliability of the two data sets. σ_{synd} for the syndromic surveillance data is set to

[1] We also tested other values for σ_q and σ_γ and find the results were qualitatively similar. Result data for the additional tests will be provided upon request.

0.05, and σ_{ILI} for ILI data is varied in the range of 0.0005 to 0.1 to depict the relative difference in reliability between the two data.

Equation (12) and (14) serve as a system model (1) and a measurement model (2) for our particle filter formulation.

4 Particle filter implementation

Our problem is to estimate the current state of epidemic spreading, where we define the system state as the number of susceptible and infectious persons in the community. Two measurement data are available for the estimation task - syndromic surveillance data and ILI case report data. Syndromic surveillance data provide information on the current state while ILI data are delayed by some lag due to data processing by the public health authority. On the other hand, ILI data are more reliable relative to syndromic data because they are obtained based on clinical diagnosis.

Upon an arrival of measurement data $\mathbf{z_k}$, our particle filter algorithm starts by determining whether $\mathbf{z_k}$ is syndromic surveillance data or ILI data. If $\mathbf{z_k}$ is syndromic surveillance data, we know that it is in sequence relative to $\mathbf{z_{k-1}}$ and therefore a standard particle filter is applied as illustrated in Sect. 2.1. If $\mathbf{z_k}$ is ILI data, it is an out-of-sequence measurement and we use an OOSM particle filter.

In this study, we provide a more straightforward implementation for the OOSM particle filter framework. Here, we depart from the OOSM particle filter algorithm of Orton and Marrs (2005) to propose a modified version for two reasons. First, applying Orton's OOSM particle filter Orton and Marrs (2005) as described in Sect. 2.2 requires sampling from $p(\mathbf{x_k^i}|\mathbf{x_a^i}, \mathbf{x_b^i}, \mathbf{z_k})$ or its approximation $p(\mathbf{x_k^i}|\mathbf{x_a^i}, \mathbf{x_b^i})$. Neither is straightforward in our case due to the nonlinearity present in the system model (13). Second, unlike general OOSM cases discussed in Orton and Marrs (2005), we assume a known and fixed amount of lag $L\Delta_m$ for the out-of-sequence measurement data. This assumption eliminates concerns related to having to store the entire history of particles throughout the filtering time horizon. With the assumption of a fixed lag, we only need to store the particle history from t_{NOW} to $(t_{NOW} - L\Delta_m)$.[2]

The basic idea behind our OOSM particle filter approach is as follows: when we obtain a measurement for a past state, the best route is to go back and re-compute from the past point as if a set of in-sequence measurement data are arriving. We call this approach a *roll-back-and-update* scheme. The original OOSM scheme discussed in Sect. 2.2 answers the following question - "Among the set of particles we have at t_{NOW}, which are the ones with high likelihood, given a newly available measurement on its past?" On the other hand, the question addressed by our roll-back-and-update scheme is, "Given a newly available measurement on its past, how would the current particle distribution change?"

[2] In the context of epidemic state estimation, this assumption may not be required since storing the entire history of particles is most likely feasible: (1) measurement sampling frequency is in the order of day or week, and thus the size of measurement data is not huge, and (2) epidemic state estimation does not require real-time computation.

Our roll-back-and-update scheme is illustrated using Fig. 1 up to t_{NOW}. Let us first ignore all out-of-sequence measurement data except $\mathbf{z_k}$. In other words, suppose we have a series of in-sequence measurement data $\mathbf{z_{k-5}}, \mathbf{z_{k-3}}, \mathbf{z_{k-1}}$, and out-of-sequence measurement data $\mathbf{z_k}$. As the in-sequence measurement data arrive, we use the standard particle filter algorithm in Sect. 2.1 to sequentially update the posterior density at $t_{NOW} - 5\Delta_m$, $t_{NOW} - 3\Delta_m$, and $t_{NOW} - \Delta_m$. Thus, at $t = t_{NOW} - \Delta_m$, we have $\{\mathbf{x^i_{k-5}}, \omega^{\mathbf{i}}_{\mathbf{k-5}}\}$, $\{\mathbf{x^i_{k-3}}, \omega^{\mathbf{i}}_{\mathbf{k-3}}\}$, and $\{\mathbf{x^i_{k-1}}, \omega^{\mathbf{i}}_{\mathbf{k-1}}\}$. Now at t_{NOW}, out-of-sequence measurement data $\mathbf{z_k}$ arrives. If we ignore the previously computed posterior densities - $\{\mathbf{x^i_{k-3}}, \omega^{\mathbf{i}}_{\mathbf{k-3}}\}$, $\{\mathbf{x^i_{k-1}}, \omega^{\mathbf{i}}_{\mathbf{k-1}}\}$ - and return to $(t_{NOW} - 4\Delta_m)$, we simply have another instance of standard particle filtering. We have $\{\mathbf{x^i_{k-5}}, \omega^{\mathbf{i}}_{\mathbf{k-5}}\}$ as the posterior density, and a series of in-sequence measurement $\{\mathbf{z_k}, \mathbf{z_{k-3}}, \mathbf{z_{k-1}}\}$: $\mathbf{x^i_{k-5}}$ is propagated to find $\mathbf{x^i_{k-4}}$, its weight ω^i_{k-4} is updated using $\mathbf{z_k}$, $\mathbf{x^i_{k-4}}$ is propagated to $\mathbf{x^i_{k-3}}$, $\omega^{\mathbf{i}}_{\mathbf{k-3}}$ is updated using $\mathbf{z_{k-3}}$, and so on. We re-compute and update the portion of history of the particles from $k-4$ to $k-1$. This roll-back-and-update process is executed every time an out-of-sequence measurement arrives.

The OOSM algorithm used in this paper is summarized below:

1. $\{\mathbf{x^i_{k-1}}, \omega^{\mathbf{i}}_{\mathbf{k-1}}\}^{\mathbf{N_s}}_{\mathbf{i=1}}$ is given, and $\mathbf{z_k}$ arrives
2. If $\mathbf{z_k}$ is a syndromic surveillance data (i.e., an in-sequence measurement)
   - Use a standard particle filter to compute $\{\mathbf{x^i_k}, \omega^{\mathbf{i}}_{\mathbf{k}}\}^{\mathbf{N_s}}_{\mathbf{i=1}}$
3. Else if $\mathbf{z_k}$ is a ILI data (i.e., an out-of-sequence measurement)
   - Let $\mathbf{z_{oosm}} = \mathbf{z_k}$
   - From the stored particle history, retrieve $\{\mathbf{x^i_{k-L-1}}, \omega^{\mathbf{i}}_{\mathbf{k-L-1}}\}$
   - Given an in-sequence measurement $\{\mathbf{z_{oosm}}, \mathbf{z_{k-L+1}}, \mathbf{z_{k-L+3}}, \ldots, \mathbf{z_{k-1}}\}$, execute a standard particle filter to update $\{\mathbf{x^i_{oosm}}, \omega^{\mathbf{i}}_{\mathbf{oosm}}\}$, $\{\mathbf{x^i_{k-L+1}}, \omega^{\mathbf{i}}_{\mathbf{k-L+1}}\}$, $\{\mathbf{x^i_{k-L+3}}, \omega^{\mathbf{i}}_{\mathbf{k-L+3}}\}$, ..., $\{\mathbf{x^i_{k-1}}, \omega^{\mathbf{i}}_{\mathbf{k-1}}\}$
   - Sample $\mathbf{x^i_k}$ using (13), and set $\omega^i_k = \omega^i_{k-1}$ to obtain $\{\mathbf{x^i_k}, \omega^{\mathbf{i}}_{\mathbf{k}}\}^{\mathbf{N_s}}_{\mathbf{i=1}}$

5 Experimental results

We first generate a *true* state sequence, $\mathbf{x_{0:T}}$, using (13) with an initial state $[s^*_0, i^*_0] = [0.99, 0.01]$. Measurement data are then generated according to (14). Syndromic surveillance data arrive with a measurement interval $\Delta_m = 10\Delta_t$, and so do ILI data. Figure 2 shows an example instance of the true system state $i(t)$ and the two measurement data ('$\times$' for syndromic surveillance data and '$*$' for ILI data). In this example, measurement noise for syndromic surveillance data, σ_{synd} is 0.05 and σ_{ILI} is much smaller at 0.005. ILI data have a lag of 50 time units as shown by the shift in the ILI data stream in Fig. 2.

For an initial prior, we use a uniform distribution such that $i_0 \sim U[0, 2i^*_0]$ and $s_0 = 1 - i_0$. The number of particles N_s is 300, and particles are resampled at every step of particle filtering. Figure 3 shows a typical example of particle filtering

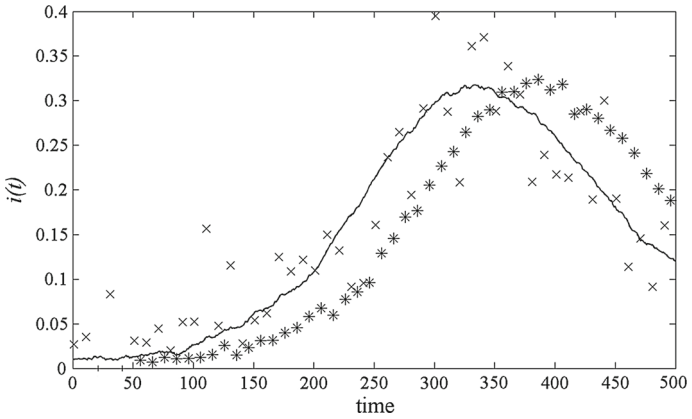

Fig. 2 An example trajectory of true state $i(t)$ (*solid line*) and the two measurement data ('$\times$' for syndromic surveillance data and '$*$' for ILI data). ILI data are delayed by a lag of 50 time units

results. On the left, we use syndromic surveillance data ('$\times$') only, and in the middle, only ILI data ('$*$') are used. Shown on the right is the estimation results when both syndromic data and ILI data are used together.

Comparing the left and middle plots, we see that the relatively high uncertainty of syndromic surveillance data ($\sigma_{synd} = 0.05 > \sigma_{ILI} = 0.005$) manifests as a wider range of particle distribution. A posterior density estimated using syndromic surveillance data shows a larger variance than using ILI data even when ILI data have a non-trivial lag. On the other hand, when we compare the first two plots and the rightmost plot, it is not readily visible whether using both sets of data improves the quality of estimation. To make quantitative comparisons, we measure the RMSE value taken over the whole trajectory as follows:

$$RMSE = \sqrt{mean\{(i_k^{true} - \hat{i}_k)^2\}} \tag{15}$$

where i_k^{true} is the true state value and $\hat{i}_k$ is the average value of N_s particles at k.

We vary the amount of lag from 0 to 140 by an increment of 10. Six levels of σ_{ILI} – $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05\}$ – are tested, while fixing σ_{synd} at 0.05. We run 20 replications for each set of parameter values to obtain the average *RMSE* over the replications. For each case, we evaluate the average *RMSE* under 1) using syndromic surveillance data only ($RMSE_{synd}$), 2) using ILI data only ($RMSE_{ILI}$), and 3) using both sets of data ($RMSE_{both}$). Results are shown in Figs. 4 and 5.

Across all levels of σ_{ILI}, the average *RMSE* curves display a consistent pattern. We make the following observations. First, when using only ILI data ($RMSE_{ILI}$), the average *RMSE* monotonically increases. An intuitive, semantic explanation for the monotonic increase is that for a given level of uncertainty in the measurement data, the *value* of their information decreases as their acquisition is more delayed. Note

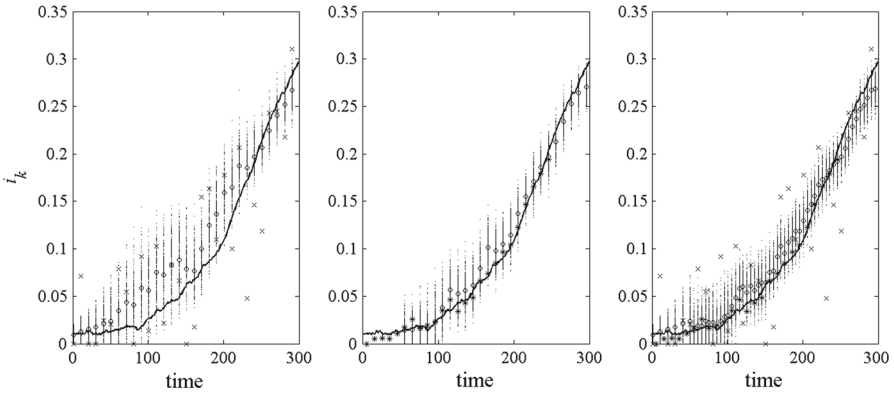

Fig. 3 Particle filter estimates the posterior density of the true state $p(\mathbf{x_k}|\mathbf{z_{1:k}})$ as an approximate density represented by a set of particles. At each k, a set of resampled particles (equal weights) is plotted along the vertical direction, and their mean value, $\hat{i}_k$, is denoted by a *circle*. Filtering results are shown for: (*Left*) syndromic surveillance data only, $\sigma_{synd} = 0.05$; (*Middle*) ILI data with lag = 50, $\sigma_{ILI} = 0.005$ (note that data points have been shifted to indicate their actual measuring point); (*Right*) both syndromic surveillance and ILI data

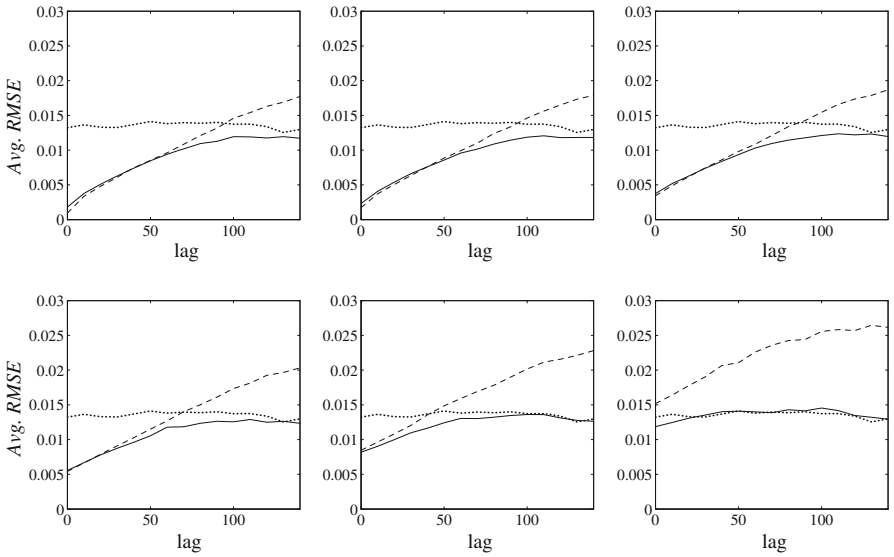

Fig. 4 Average *RMSE* as a function of lag in ILI data: $\sigma_s = \sigma_i = 0.001$; $\sigma_{synd} = 0.05$; $\sigma_{ILI} = 0.001, 0.002, 0.005$ (*top left* to *right*), 0.01, 0.02, 0.05 (*bottom left* to *right*); *dotted line* for a case where only syndromic surveillance is used, *dashed line* for ILI data only, and *solid line* for both sets of data

that, in Figs. 4 and 5, $RMSE_{synd}$ curves (dotted) are identical in all subplots since they are not a function of σ_{ILI}. They remain constant along the x-axis in each subplot since they are not a function of lag either.[3]

[3] Slight variations visible in the figures are due to non-systematic causes.

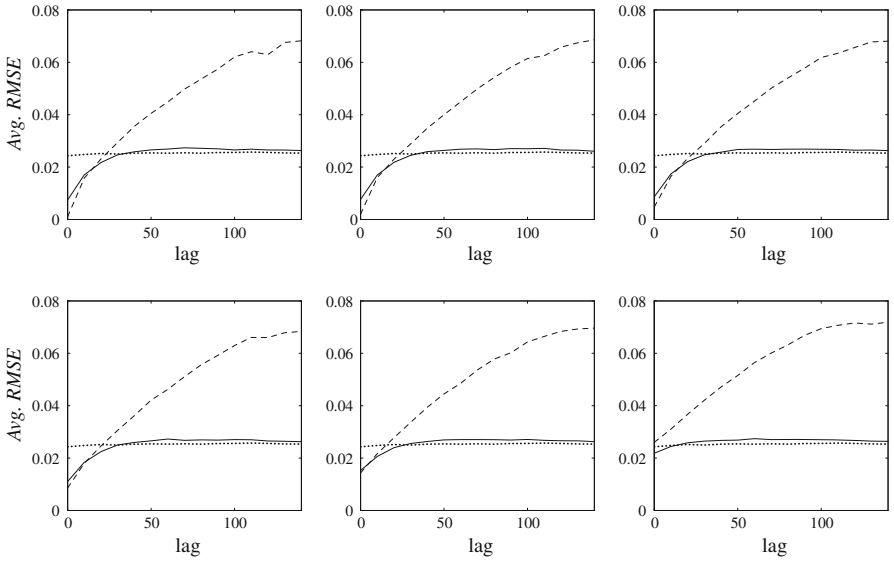

Fig. 5 Average *RMSE* as a function of lag in ILI data: $\sigma_s = \sigma_i = 0.005$; $\sigma_{synd} = 0.05$; $\sigma_{ILI} = 0.001, 0.002, 0.005$ (*top left* to *right*), 0.01, 0.02, 0.05 (*bottom left*to *right*); *dotted line* for the case where only syndromic surveillance is used, *dashed line* for ILI data only, and *solid line* for both sets of data

Second, the monotonic increase observed in the average *RMSE* seems to approach a certain limit. This is particularly visible in Fig. 5. Again, we may offer an intuitive explanation. A very large measurement lag would make measurement information obsolete, and at an extreme, it will be equivalent to having no (useful) measurement at all. In this case, we will be left with a system model only, and our estimation of system states will be no better than the system model's accuracy (i.e., process noise). Thus, the average *RMSE* would approach to a limit, which depends on the underlying process noise.

Third, the average *RMSE* when using both sets of measurement data ($RMSE_{both}$) stays *below* the *RMSE* curves for each data case. While this is rather expected, a closer examination suggests a more interesting behavior. It approaches the *RMSE* curve of ILI-only ($RMSE_{ILI}$) when the lag goes to zero, and it approaches the syndromic-surveillance-only curve ($RMSE_{synd}$) when the lag becomes very large. We also note that the difference between $min\{RMSE_{synd}, RMSE_{ILI}\}$ and $RMSE_{both}$ seems to be maximized when $RMSE_{synd}$ and $RMSE_{ILI}$ curves intersect. The following conjecture for this observation is possible: when the value of one of the two sets of measurements dominates the other, the benefit of using both sets of measurement diminishes and its state estimation is no better than when using the superior measurement data. Using both sets of measurement data is most rewarded when the two have *comparable* value.

Figure 6 presents the same results along the σ_{ILI} axis for a fixed lag. It displays the same pattern as observed in the earlier figures. As σ_{ILI} becomes small, $RMSE_{both}$

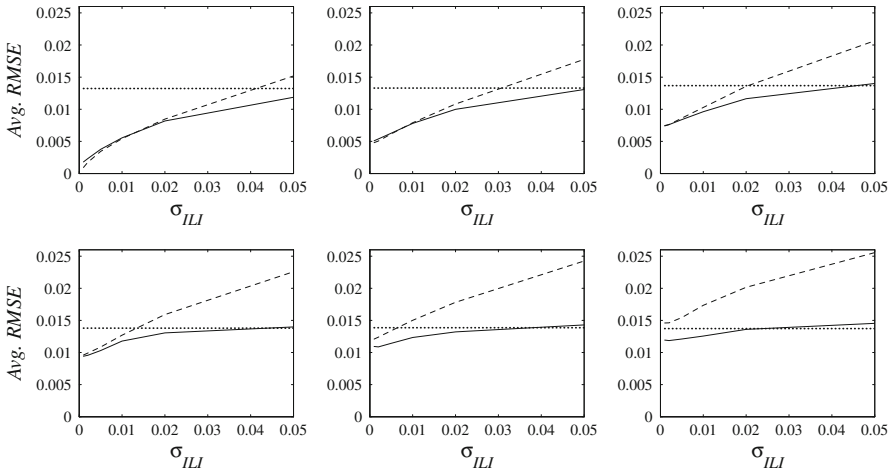

Fig. 6 Average *RMSE* as a function of σ_{ILI} for a fixed lag: $\sigma_s = \sigma_i = 0.001$; $\sigma_{synd} = 0.05$; lag = 0, 20, 40 (*top left* to *right*), 60, 80, 100 (*bottom left* to *right*); *dotted line* for a case where only syndromic surveillance is used, *dashed line* for clinical case report only, and *solid line* for both sets of data

approaches the $RMSE_{ILI}$ curve, and vice versa. The benefit of using both sets of data appears to be maximized when the two *RMSE* curves intersect.

Recall that the first motivating question for our experiment was "given immediately-available-but-uncertain data and credible-but-delayed data, which one would yield better state estimates?" The experimental results suggest that this depends on relative uncertainty and the amount of delay. The second question was "would it always be better to use both sets of data than using only one?" The answer seems to be "not always." It is advantageous to use both sets of data when they have *comparable values*. Otherwise, it is only as good as using measurement data of a higher quality.

All these results appear to indicate that there is a systematic mechanism behind the patterns observed in the above figures. Further investigations using analytic models are warranted to support the observations and conjectures mentioned above. In the next section, we draw an analytic explanation on the behavior of the *RMSE* curves exhibited in Figs. 4, 5, 6.

6 Discussion

We begin our discussion by examining the effect of measurement delay on *RMSE* of state estimates. In Sect. 5, we saw that whether immediately-available-but-uncertain data or credible-but-delayed data would yield better estimates depends on the amount of delay. This leads us to a concept of *equivalent* standard deviation, $\tilde{\sigma}_z$. Suppose we have a series of measurement data with delay L and their standard deviation is $\sigma_z = c$, and we obtain an average $RMSE = r$. Then, we find the standard deviation $\tilde{\sigma}_z$ of hypothetical, no-delay measurements that yields the same *RMSE*.

We call $\tilde{\sigma}_z$ the equivalent standard deviation for the original, delayed measurement z. We expect that the value of $\tilde{\sigma}_z$ will be larger than σ_z. The concept of equivalent standard deviation is illustrated in Fig. 7. Figure 7 (left) is a plot of average *RMSE* as a function of σ_z for various amount of lags: $\{0, 10, 20, \ldots, 140\}$. When, for example, measurement data have a delay of 30 time units and the data's standard deviation is 0.039, the graph shows that the average *RMSE* of the estimation is approximately 0.0165. Looking at the curve for lag = 0 (the bottom-most curve, thick-lined), we see that standard deviation of 0.06 would provide the same level of average *RMSE*. Hence, $\tilde{\sigma}_z$ in this example is 0.06, which is larger than σ_z. Non-delayed measurement data with larger noise provide an *informative value equivalent* to delayed measurement data with smaller noise. Shown in Fig. 7 (right) is a plot of $\tilde{\sigma}_z$ as a function of measurement delay for $\sigma_z = \{0.05, 0.02, 0.01, \ldots, 0.0005\}$. It shows that $\tilde{\sigma}_z$ monotonically increases as a function of measurement delay, which is consistent with our intuitive expectation: the informative value of measurement data decreases as the amount of delay increases.

This monotonicity explains why we observe similar patterns in Figs. 4, 5 and 6. In each subfigure of Fig. 4, the average *RMSE* is shown as a function of the lag in ILI data. Since the increase in the lag implies a higher equivalent standard deviation, Fig. 4 can be redrawn as a function of $\tilde{\sigma}_z$ with more or less the same pattern. This has bearing on our analysis of the behavior of *RMSE* curves: it suggests that we may consider only $\tilde{\sigma}_z$ instead of σ_z (with delay L) when describing the behavior of the *RMSE* curve. As such, in what will follow, we construct an analytic explanation assuming no delay in measurement data.

Consider the following simple model where we are given two types of independent measurements, z_1 and z_2, and attempt to compute the posterior density for a state variable x, $p(x|z_1, z_2)$, . For now, we assume a normal distribution for its prior, $p(x)$, and likelihood, $p(z_1|x)$ and $p(z_2|x)$. That is, $p(x) = N(\mu_0, \sigma_0^2), p(z_1|x) = N(x, s_1^2)$, and $p(z_2|x) = N(x, s_2^2)$. Let β_0 denote the precision of $N(\mu_0, \sigma_0^2)$, i.e., $\beta_0 = 1/\sigma_0^2$. Likewise, $b_1 = 1/s_1^2$ and $b_2 = 1/s_2^2$.

The posterior density of x given a measurement z_1 is written as $p(x|z_1) \propto p(z_1|x)p(x)$. Since we assume a normal prior and normal likelihood, we know the posterior is also a normal density, $N(\mu_1, 1/\beta_1)$, where $\beta_1 = b_1 + \beta_0$ and $\mu_1 = (b_1 z_1 + \beta_0 \mu_0)/\beta_1$; the precision of the posterior is improved by adding the precision of measurement data b_1 to the prior's precision β_0, and its mean is an average of the prior mean and measurement weighted by the relative precision of each. Similarly, $p(x|z_2) \propto N(\mu_2, 1/\beta_2)$, where $\beta_2 = b_2 + \beta_0$ and $\mu_2 = (b_2 z_2 + \beta_0 \mu_0)/\beta_2$.

When both measurement data are given, the posterior density can be written as a factorized form: $p(x|z_1, z_2) \propto p(z_1|x, z_2)p(z_2|x)p(x) = p(z_1|x)(z_2|x)p(x)$. Note that in the second step, the conditional independence between z_1 and z_2 (i.e., $z_1 \perp z_2|x$) is used. $p(x|z_1, z_2)$ is a product of three normal densities, and it is straightforward to show that it is a normal density $N(\mu_{12}, 1/\beta_{12})$ with $\beta_{12} = b_1 + b_2 + \beta_0$ and $\mu_{12} = (b_1 z_1 + b_2 z_2 + \beta_0 \mu_0)/\beta_{12}$. To summarize, we have the following results for our model:

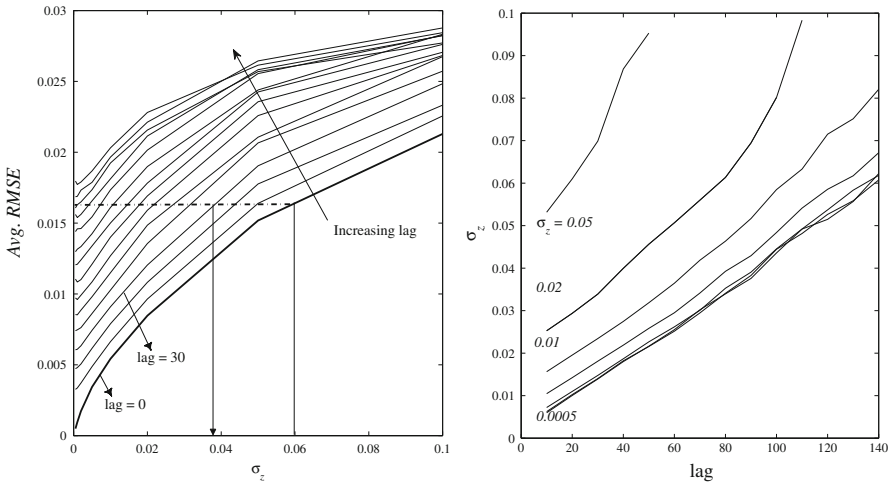

Fig. 7 (*Left*) Same level of avg. *RMSE* is obtained from delayed data with σ_z and no-delay data with $\tilde{\sigma}_z$; (*Right*) Longer lag for a given level of σ_z corresponds to a bigger $\tilde{\sigma}_z$

$$
\begin{aligned}
p(x|z_1) = N(\mu_1, \sigma_1^2) \qquad & \mu_1 = \frac{b_1 z_1 + \beta_0 \mu_0}{b_1 + \beta_0} \qquad & \sigma_1^2 = \frac{1}{1/s_1^2 + 1/\sigma_0^2} \\
p(x|z_2) = N(\mu_2, \sigma_2^2) \qquad & \mu_2 = \frac{b_2 z_2 + \beta_0 \mu_0}{b_2 + \beta_0} \qquad & \sigma_2^2 = \frac{1}{1/s_2^2 + 1/\sigma_0^2} \\
p(x|z_1, z_2) = N(\mu_{12}, \sigma_{12}^2) \quad & \mu_{12} = \frac{b_1 z_1 + b_2 z_2 + \beta_0 \mu_0}{b_1 + b_2 + \beta_0} \quad & \sigma_{12}^2 = \frac{1}{1/s_1^2 + 1/s_2^2 + 1/\sigma_0^2}
\end{aligned}
\tag{16}
$$

With (16), for a fixed value of s_1 and σ_0, we can compute $\sigma_{12}^2, \sigma_2^2, \sigma_1^2$ by varying s_2. Figure 8 shows a plot of $\sigma_{12}, \sigma_2, \sigma_1$ as a function of s_2.

Now, let indexes 1 and 2 denote the syndromic surveillance data and ILI data, respectively. σ_1^2 is then a posterior variance given syndromic surveillance data only, and σ_2^2 for ILI data only. σ_{12}^2 is a posterior variance when both sets of data are used. Figure 8 is then analogous to Fig. 6, and we see that the behavior observed in the experimental results is almost exactly reproduced in Fig. 8.

Furthermore, using (16), we can show that the benefit of using both sets of data is maximized when $\sigma_1^2 = \sigma_2^2$, which is consistently observed in Figs. 4 , 5, 6. Let $\Delta_{1,12}$ (resp., $\Delta_{2,12}$) denote the reduction of posterior variance by using both datasets compared to using only dataset 1 (resp., dataset 2). Subtracting σ_{12}^2 from σ_1^2 gives $\Delta_{1,12}$ as

$$
\begin{aligned}
&\Delta_{1,12}(\lambda) = \sigma_1^2 - \sigma_{12}^2 = \frac{A^2}{AB + B^2\lambda} \\
&A = s_1^2; \quad B = s_1^2 + \sigma_0^2
\end{aligned}
\tag{17}
$$

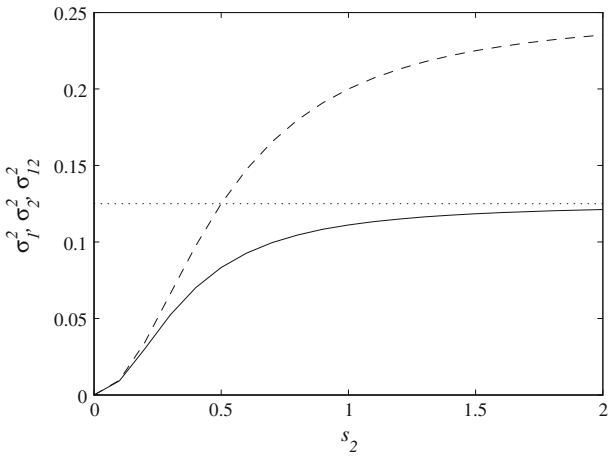

Fig. 8 σ_{12}^2(*solid*), σ_2^2(*dashed*), σ_1^2(*dotted*) as a function of s_2

Note we use λ to denote s_2^2 for a simpler presentation. Since λ is strictly non-negative, $\Delta_{1,12}$ is a monotonically decreasing function, where $\Delta_{1,12}(0) = A/B$ and $\Delta_{1,12}(\infty) = 0$. Similarly, we have an expression for $\Delta_{2,12}(\lambda)$:

$$\Delta_{2,12}(\lambda) = \sigma_2^2 - \sigma_{12}^2 = \frac{(B\sigma_0^2 - A)\lambda^2}{B\lambda^2 + (B\sigma_0^2 + A)\lambda + A\sigma_0^2} \tag{18}$$

$\Delta_{2,12}(0) = 0$ and $\Delta_{2,12}(\infty) = A/B * \sigma_0^2/s_1^2$. The first derivative of $\Delta_{2,12}(\lambda)$ confirms that it is an increasing function of λ. Figure 9 shows a plot of $\Delta_{1,12}$ and $\Delta_{2,12}$ as a function of λ.

The benefit of using both sets of data over using one (superior) set is given by the minimum of $\Delta_{1,12}$ and $\Delta_{2,12}$ - i.e., a reduction in the posterior variance compared to the better one of σ_1^2 and σ_2^2. This is given by the dashed line up to λ^* and by the solid line after λ^*, and its maximum is obtained at λ^*. λ^* is when $\Delta_{1,12} = \Delta_{2,12}$, which is equivalent to $\sigma_1^2 = \sigma_2^2$. This is consistent with the earlier statement made in sect. 5: "Using both sets of measurement data is most rewarded when the two sets have *comparable* informative value".

We end this section by recapitulating the implications and importance of high-fidelity epidemic state estimation especially in the context of future prediction. Figure 10 shows a scenario of predicting the future course of disease spread. Prediction is made on the 14th day ($t = 140$) after the initial outbreak. Syndromic surveillance data arrive daily from day 1, while ILI data arrive with a lag of 0 (top-left), 3 (top-right), 5 (bottom-left), and 7 (bottom-right) days. When the ILI data arrive with a shorter lag, the posterior density on the 14th day shows much smaller variance. Future prediction depends on the posterior density on day 14, and Fig. 10 clearly demonstrates the effect of the goodness of the posterior density on the prediction of future progress of an epidemic spread.

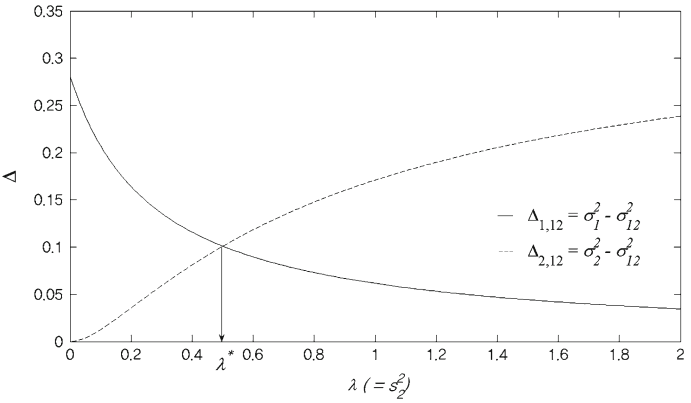

Fig. 9 Reduction of posterior variance, Δ, when using both sets of data. The benefit of using both data sets over one set is maximized at λ^* when the variances of the two sets are equal to each other

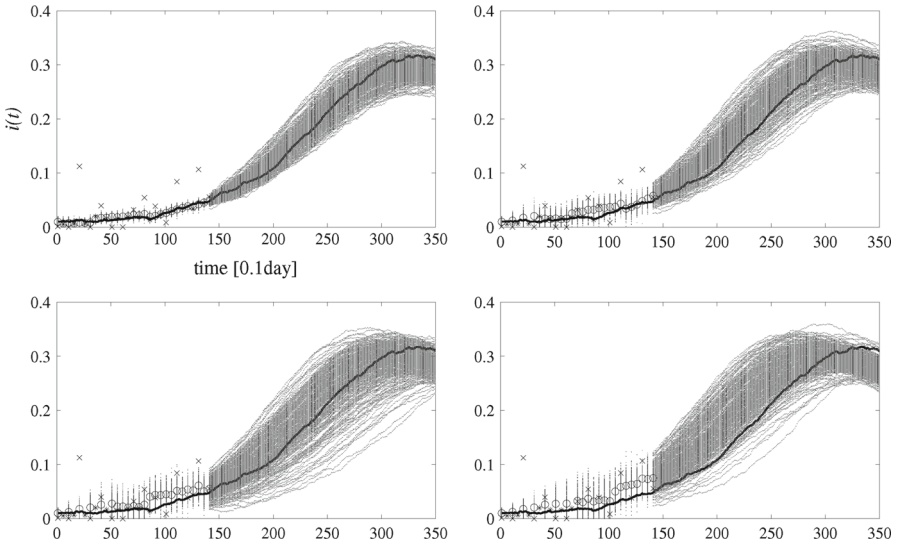

Fig. 10 ILI data are delayed by 0 days (*top-left*), 3 days (*top-right*), 5 days (*bottom-left*), and 7 days (*bottom-right*). Larger lags in ILI report data increases uncertainty in prediction of future states

7 Conclusion

We study a problem of estimating current epidemic state by combining syndromic surveillance data and ILI data through particle filtering. The two sets of data compliment each other: syndromic surveillance data are immediately available but contain large noise while more reliable ILI data are delayed by some lag due to the reporting process. Our experimental results from hypothetical pandemic scenarios show that using both sets of data is advantageous only when the informative value

of the two data sets is comparable. Analysis of a linear, Gaussian case suggests that this behavior is a logical consequence of using a Bayesian stochastic filtering framework. While we believe that the method and conclusions in the paper are not confined to the hypothetical cases tested therein, it will be worthwhile and interesting to conduct further experiments using real data to validate the practical significance and relevance of this work in real world applications.

Considering there is a possible trade-off between timeliness and credibility of clinically validated surveillance data, appropriate design of surveillance data collection and processing is a valid optimization problem. We expect that understanding and insights as well as the state estimation technique presented in this paper will aid in such decision making for public health authorities.

Acknowledgments This research was supported by the Public Welfare & Safety Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (No.2011-0029881) and by Basic Science Research Program through NRF funded by the Ministry of Education (NRF-2010-0025224).

References

Bisset KR, Chen J, Feng X, Kumar VSA, Marathe MV (2009) EpiFast: A Fast Algorithm for Large Scale Realistic Epidemic Simulations on Distributed Memory Systems. Proceedings of the 23rd international conference on Supercomputing. 430–439.

Chao DL, Halloran ME, Obenchain VJ, Longini IM Jr (2010) FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput Biol 6(1): doi:10.1371/journal.pcbi.1000656

Chen L, Achrekar H, Liu B, Lazarus R (2010) Vision: Towards Real Time Epidemic Vigilance through Online Social Networks. ACM Workshop Mobile Cloud Comput Serv, San Francisco, USA

Chew C, Eysenbach G (2010) Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak. PLoS ONE 5(11): doi:10.1371/journal.pone.0014118

Dailey L, Watkins RE, Plant AJ (2007) Timeliness of data sources used for influenza surveillance. J Am Med Inform Assoc 14(5):626–631. doi:10.1197/jamia.M2328

Ducet A, Johansen AM (2013) A tutorial on particle filtering and smoothing: Fifteen years later. http://www.cs.ubc.ca/~arnaud/doucet_johansen_tutorialPF. Accessed 27 December 2013

Dukic V, Lopes HF, Polson NG (2012) Tracking epidemics with google flu trends data and a state-space SEIR model. J Am Stat Assoc 107(500):1410–1426

Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, Wang N (2004) Modelling disease outbreaks in realistic urban social networks. Nature 429:180–184. doi:10.1038/nature02541

Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437:209–214

FluView: A weekly influenza surveillance report. Centers for Disease Control and Prevention. http://www.cdc.gov/flu/weekly/. Accessed 5 February 2013

Gensheimer KF, Fukuda K, Brammer L, Cox N, Patriarca PA, Strikas RA (1999) Preparing for pandemic influenza: the need for enhanced surveillance. Emerg Infect Dis 5:297–299

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014

Henning KJ (2004) Overview of syndromic surveillance: what is syndromic surveillance? Morb Mortal Wkly Rep 53(Supplement):5–11

Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42(4):599–653

Influenza weekly report. Korea Centers for Disease Control and Prevention. http://www.cdc.go.kr/CDC/info/CdcKrInfo0402.jsp?menuIds=HOME001-MNU0003-MNU0727-MNU0045. Accessed 27 October 2013

Influenza Surveillance. Korea Centers for Disease Control and Prevention. http://www.cdc.go.kr/CDC/contents/CdcKrContentView.jsp?cid=17936&menuIds=HOME001-MNU1132-MNU1138-MNU0741. Accessed 2 May 2014

Jegat C, Carrat F, Lajaunie C, Wackernagel H (2008) Early detection and assessment of epidemics by particle filtering. In: Soares A, Pereira M, Dimitrakopoulos R (eds) geoENV VI: geostatistics for environmental applications. Springer, Netherlands, pp 23–35

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond A 115(772):7000–7721. doi:10.1098/rspa.1927.0118

Lampos V, Bie T, Cristianini N (2010) Flu detector: tracking epidemics on twitter. In: Balcazar J, Bonchi F, Gionis A, Sebag M (eds) Machine learning and knowledge discovery in databases. Springer, Heidelberg, pp 599–602

Longini IM Jr, Nizam A, Xu S, Ungchusak K, Hanchaoworakul W, Cummings DA, Halloran ME (2005) Containing pandemic influenza at the source. Science 309:1083–1087

Ong JBS, Chen MI, Cook AR, Lee HC, Lee VJ (2010) Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. PLoS ONE 5(4): doi:10.1371/journal.pone.0010036

Orton M, Marrs A (2005) Particle filters for tracking with out-of-sequence measurements. IEEE Trans Aerosp Electron Syst 41(2):693–702

Que J, Tsui F-C (2011) Rank-based spatial clustering: an algorithm for rapid outbreak detection. J Am Med Inform Assoc 18(3):218–224. doi:10.1136/amiajnl-2011-000137

Reis BY, Kohane IS, Mandl KD (2007) An epidemiological network model for disease outbreak detection. PLoS Med 4(6):e210. doi:10.1371/journal.pmed.0040210

Ristic B, Arulampalam S, Gordon N (2004) Beyond the kalman filter: particle filters for tracking applications. Artech House Publishers, Boston

Rahmandad H, Sterman J (2008) Heterogeneity and network structure in the dynamics of diffusion: comparing agent-based and differential equation models. Manag Sci 54(5):998–1014

Shaman J, Karspeck A (2012) Forecasting seasonal outbreaks of influenza. Proc Natl Acad Sci. doi:10.1073/pnas.1208772109

Singh BK, Savill NJ, Ferguson NM, Robertson C, Woolhouse ME (2010) Rapid detection of pandemic influenza in the presence of seasonal influenza. BMC Public Health 10(726): doi:10.1186/1471-2458-10-726

Skvortsov A, Ristic B (2012) Monitoring and prediction of an epidemic outbreak using syndromic observations. Math Biosci 240:12–19

Vidal Rodeiro CL, Lawson AB (2006) Online updating of space-time disease surveillance models via particle filters. Stat Methods Med Res 15:423–444

WHO checklist for influenza pandemic preparedness planning. Department of communicable disease surveillance and response global influenza programme. http://www.who.int/influenza/resources/documents/FluCheck6web. Accessed 27 December 2013