PREDICTION OF INFLUENZA RATES BY PARTICLE FILTERING

Pau Closas[(1)]*, Mónica F. Bugallo*[(2)]*, Ermengol Coma*[(3)]*, and Leonardo Méndez*[(3)]

[(1)]Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)
Av. Carl Friedrich Gauss 7, 08860 Castelldefels, Barcelona, Spain
[(2)] Department of Electrical & Computer Engineering
Stony Brook University, Stony Brook, NY 11794, USA
[(3)]Institut Català de la Salut (ICS), Sistema d'Informació dels Serveis d'Atenció Primària (SISAP)
Gran Via de les Corts Catalanes, 587-589, 08007 Barcelona, Spain

e-mail: pclosas@cttc.cat, monica.bugallo@stonybrook.edu, {ecomaredon,lmendezboo}@gencat.cat

ABSTRACT

Predicting the course of influenza rates is extremely useful for the efficacy of planned vaccination programs. In this paper we address this problem by stating a dynamic state-space model that mathematically describes both the evolution of influenza rates and the observations obtained by a surveillance system. We then propose a prediction method based on particle filtering that accommodates the nonlinear nature of the model. Using real data we estimate the necessary model functions prior to the prediction step. Computer simulations reveal promising results of the proposed method.

Index Terms— Time series prediction, nonlinear systems, particle filtering, influenza.

1. INTRODUCTION

Influenza is a well known and common human respiratory infection. It is responsible of significant morbidity and mortality every year. The World Health Organization estimates that annual epidemics result in about 3 to 5 million cases of severe illness and about 250,000 to 500,000 casualties worldwide [1]. Surveillance systems are important tools for detection and monitoring of infectious diseases since they allow for quick response and planning of health resources. Furthermore, the prediction of the evolution of influenza is of major interest if one aims at advancing the response time.

The objective of surveillance systems is to provide an indicator as soon as data gives enough evidence to assess the disease outbreak [2]. In [3], the authors presented a tool to analyze epidemiological data. They used non-epidemic training data to fit a time-series model for the periodic baseline level. Another method proposed to detect deviations from the baseline based on fitting historical data to a time-series regression model [4], [5]. In [6] a switching Markov model was used with an autoregressive process to model epidemic data and a white Gaussian process for non-epidemic modeling. Other references point to imaginative usage of the *big data* such as [7, 8]. The present paper contributes with an online method for outbreak prediction. We propose a procedure to build the model and pay special care in designing the prediction algorithm, which is based on Bayesian theory.

The objective of this paper is to obtain a new algorithm for prediction of the evolution of influenza incidence rates over time. The inputs of the algorithm are the records of influenza cases at the t-th epidemiological week (EW) of the season of interest and a model of the system built using prior knowledge of previous seasons.[1] We aim at predicting the behavior of the influenza incidence rates for later EWs of the season in study. Although we consider linear-noisy measurements of the incidence rates, the dynamics of epidemiological data are nonlinear. Therefore, we use particle filtering [10] to design the prediction algorithm. Particle filters perform a discrete characterization of the posterior distribution of the system based on a properly weighted random set of points, which is suitable in nonlinear/non-Gaussian systems.

The rest of the paper is organized as follows. Section 2 describes the dynamical model of the system and the method used to build the necessary functions from existing data. The prediction algorithm based on the particle filtering methodology is presented in Section 3. Section 4 discusses computer simulations performed to evaluate the method and the obtained results. Finally, Section 5 concludes the paper.

P. C. would like to thank the support of the Spanish Ministry of Economy and Competitiveness project TEC2012-39143 (SOSRAD), by the European Commission in the COST Action IC0803 (RFCSET) and the Network of Excellence NEWCOM♯ (contract n. 318306). M. F. B. would like to thank the National Science Foundation under Award CCF-0953316, the Office of Naval Research under Award N00014-09-1-1154 and the Chair of Excellence Program of Universidad Carlos III de Madrid-Banco de Santander.

[1] In particular, for model building we use real data from Diagnosticat, an existing open-access database of the Catalan Institute of Health [9]. These data is also used for evaluation of the resulting method.

2. SYSTEM MODEL FOR INFLUENZA SURVEILLANCE SYSTEMS

We consider a state-space model for influenza dynamics that is driven by Gaussian processes. Particularly, we assume that the surveillance system is weekly recording noisy samples of the true incidence rates, y_t, modeled as

$$y_t = x_t + n_t, \tag{1}$$

where $t = 1, \cdots, T$, is an index denoting the EW,[2] x_t is the targeted influenza incidence rate at the t-th EW, and n_t is a zero-mean Gaussian noise with variance σ_y^2, which models inaccuracies of physicians when diagnosing cases of influenza. Typically, y_t is normalized per 10^5 population, a unit that allows comparison of diagnoses over different territories independently of the number of inhabitants [11].

The model for the observations of the system given by (1) is rather straightforward. However, the selection of a meaningful model for the time-evolution of influenza is more involved and here we represent it in very general terms as

$$x_t = f_t(x_{t-1}) + \nu_t, \tag{2}$$

where $f_t(\cdot)$ is a known, possibly nonlinear, function of the state x_t, and ν_t is the state noise noise, which gathers any mismodeling effects or disturbances in the state characterization. We assume that $\nu_t \sim \mathcal{N}(0, \sigma_{x,t}^2)$, and that in general, $\sigma_{x,t}^2$ can be time-dependent.

2.1. Building the state function

There have been some attempts to model $f_t(\cdot)$ in the literature. For instance, [4] proposed a method to detect deviations from the baseline based on fitting historical data to a time-series regression model. The resulting function was obtained by combination of a linear term describing the secular trend with sine and cosine terms describing seasonal change.

In this paper we propose a procedure to build $f_t(\cdot)$ based on processing a set of existing influenza season data. We use the open-access database Diagnosticat [9], which contains clinical influenza diagnoses codes updated weekly using an electronic health recording system where primary care physicians routinely register their activity. The website is timely updated a few minutes after the last day of the EW and accounts for entries of over 3,500 physicians collecting data of nearly 6 million people (80% of the population of the surveilled area) [9].

The function $f_t(\cdot)$ is generated by processing a set of L training seasons, denoted as $\mathcal{T} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(L)}\}$, where $\mathbf{y}^{(j)} = y_{1:T}^{(j)}$ corresponds to the observation sequence of the j-th season of the training set. In particular, from each season series $\mathbf{y}^{(j)}$ one could obtain a function fitting the data, $f_t^{(j)}(\cdot)$. We set the condition that the resulting function has the characteristic that each point of the time series depends on the previous one, as in (2). This property allows for building a state-space model with the Markov property, which is then used to design the Bayesian predictor.

Under the assumption that σ_y^2 is small, we can consider that $f_t^{(j)}(x_{t-1}) \approx f_t^{(j)}(y_{t-1})$ and as a result of processing $\mathcal{T}$, we obtain the set of approximated functions per training season $\mathcal{F} = \{f_t^{(1)}, \ldots, f_t^{(L)}\}$. Once each function corresponding to the different seasons is obtained, the function for the state equation (2) can be constructed with weighted mean and unbiased variance as follows

$$f_t(x_{t-1}) = \sum_{j=1}^{L} \alpha^{(j)} f_t^{(j)}(x_{t-1}) \tag{3}$$

$$\sigma_{x,t}^2 = \frac{1}{1 - \sum_{j=1}^{L} \left(\alpha^{(j)}\right)^2} \sum_{j=1}^{L} \alpha^{(j)} (\mathbf{y}_t^{(j)} - f_t^{(j)}(x_{t-1}))^2, \tag{4}$$

where $\alpha^{(j)} \geq 0$ represents the weight given to the data from the j-th season, with $\sum_j \alpha^{(j)} = 1$. Note that the method is quite versatile and could be adapted to the modeling of other diseases.

As stated, the fitting of $f_t^{(j)}(y_{t-1})$ given $\mathbf{y}^{(j)}$ can be done in different ways [12]. The scope of this work is to propose a new prediction method for influenza incidence rates, but it is necessary to use a model for its functioning. For that reason, we resorted to the software tool Eureqa [13], which allows for detection of equations and hidden mathematical relations in data. We used 4 influenza seasons available in Diagnosticat [9] with a fitness metric minimizing the mean of the absolute value of residual errors. The resulting expressions were

$$\begin{aligned}
f_t^{(1)}(x_{t-1}) &= (1.548 + 0.3443x_{t-1} - \cos(x_{t-1}) \\
&- 0.06626x_{t-1}\cos(0.1253x_{t-1}^2)) \,/ a_1 \\
f_t^{(2)}(x_{t-1}) &= x_{t-1} + (x_{t-1}t\cos(0.4435t) \\
&+ x_{t-1}\sin(6.12 + x_{t-1})\cos(0.4435t) \\
&+ x_{t-1}t\cos(0.4435t)\cos(0.4435t)) \\
&\times (60.6 + 0.8503x_{t-1})^{-1} \\
f_t^{(3)}(x_{t-1}) &= 0.1704 + 0.9154x_{t-1} \\
&+ (0.03797x_{t-1}t - 1)\sin(0.09343x_{t-1}t) \\
&- \frac{0.5854x_{t-1}t\sin(0.09343x_{t-1}t)}{7.625 + x_{t-1} - \cos(-0.5854x_{t-1}t)} \\
&+ 0.09343x_{t-1}\sin(0.1353x_{t-1}t) \\
f_t^{(4)}(x_{t-1}) &= 0.1365t\cos(0.1081t) \\
&+ \frac{2.173 + x_{t-1} + t}{0.2446t + \cos(0.4403t)}
\end{aligned}$$

with $a_1 = \cos(4.931 + 0.05115t)$. Eureqa also reported

[2] In epidemiology and surveillance of infectious diseases the EW is the time unit for interpretation of data. The EW is a group of seven days that begins on a Sunday and ends on a Saturday. One year may have 52 or 53 EWs depending on the beginning of the first week.

the following R^2 goodness-of-fit measures: 0.99757973, 0.99664044, 0.99718319, and 0.9968961 corresponding to each season in the database, respectively. Recall that $0 \leq R^2 \leq 1$ is used to describe how well a regression line fits a set of data with values close to 1 indicating agreement.

The obtained fitting, along with the data used, can be seen in Fig. 1. It is important to notice that the influenza season 2009-2010 presented a different temporal pattern due to the A(H1N1) Influenza virus pandemic [14, 15, 16]. That epidemic season had higher incidence rates and took place some weeks before than regular seasonal influenza. This fact should be taken into consideration when building the model for the prediction method, as it could bias the results. The corresponding weight, $\alpha^{(2)}$, will reflect a low importance value with respect to the rest of seasons in the database.

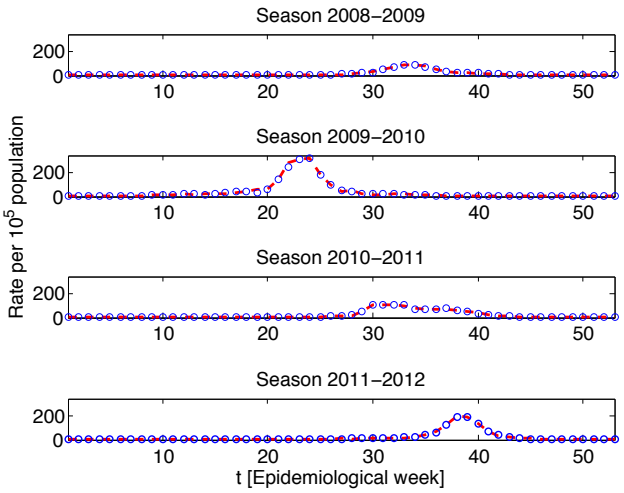

Fig. 1. Weekly recorded data (circles) and approximated functions $f_t^{(j)}(x_{t-1})$ using Eureqa (dashed lines) for the influenza seasons 2008-2012 in the Diagnosticat's database.

3. PREDICTION BY PARTICLE FILTERING

In this section, we propose a prediction algorithm that uses both the noisy observations and the dynamic model previously described. Since the latter is clearly nonlinear, we use the particle filtering methodology [17, 18], which is specially suitable for this type of problems. The objective is to approximate the step-ahead prediction distribution, $p(x_{t+\tau}|y_{1:t})$, with $\tau \geq 1 \in \mathbb{N}$ being the number of step-ahead EWs that we want to predict and $y_{1:t} = \{y_1, \ldots, y_t\}$ the available data.

Generally speaking, particle filtering approximates the filtering distribution $p(x_t|y_{1:t})$ by a set of N weighted random samples, forming the random measure $\left\{x_t^{(i)}, w_t^{(i)}\right\}_{i=1}^{N}$. These random samples are drawn from an importance density

$$x_t^{(i)} \sim \pi(x_t|x_{0:t-1}^{(i)}, y_{1:t}), \tag{5}$$

and weighted according to

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(y_t|x_{0:t}^{(i)}, y_{1:t-1})p(x_t^{(i)}|x_{t-1}^{(i)})}{\pi(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t})}. \tag{6}$$

The choice of the importance density is critical in designing an efficient particle filtering method. It is well-known that the optimal importance density is $\pi(x_t|x_{0:t-1}^{(i)}, y_{1:t}) = p(x_t|x_{t-1}^{(i)}, y_t)$, and it minimizes the variance of importance weights. In that case, the weights in (6) reduce to $w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t|x_{t-1}^{(i)})$. This choice requires the ability to draw from $p(x_t|x_{t-1}^{(i)}, y_t)$ and to evaluate $p(y_t|x_{t-1}^{(i)})$. In general, the two requirements cannot be met and one needs to resort to suboptimal choices. However, the state-space model assumed here is Gaussian, with a nonlinear process equation while the observations are linear. Therefore, we are able to use the optimal importance density [17] and the proposal distribution tuns out to be

$$p(x_t|x_{t-1}^{(i)}, y_t) = \mathcal{N}(\mu_{\pi,t}^{(i)}, \sigma_{\pi,t}^2) \tag{7}$$

with

$$\mu_{\pi,t}^{(i)} = \sigma_{\pi,t}^2 \left(\frac{f_t(x_{t-1}^{(i)})}{\sigma_{x,t}^2} + \frac{y_t}{\sigma_y^2} \right) \tag{8}$$

$$\sigma_{\pi,t}^2 = \left(\frac{1}{\sigma_{x,t}^2} + \frac{1}{\sigma_y^2} \right)^{-1}, \tag{9}$$

and the importance weights can be updated using

$$p(y_t|x_{t-1}^{(i)}) = \mathcal{N}(f_t(x_{t-1}^{(i)}), \sigma_{x,t}^2 + \sigma_y^2). \tag{10}$$

The particle filter provides a discrete approximation of the filtering distribution of the form $p(x_t|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(x_t - x_t^{(i)})$. However, in the problem of predicting the time-course of influenza cases, we are interested in the estimation of the step-ahead prediction distribution

$$\begin{aligned} p(x_{t+\tau}|y_{1:t}) &= \int p(x_t|y_{1:t}) \left(\prod_{k=t+1}^{t+\tau} p(x_k|x_{k-1}) \right) dx_{t:t+\tau-1} \\ &\approx \sum_{i=1}^{N} w_t^{(i)} \int p(x_{t+1}|x_t^{(i)}) \left(\prod_{k=t+2}^{t+\tau} p(x_k|x_{k-1}) \right) dx_{t+1:t+\tau-1} \end{aligned} \tag{11}$$

where one uses the approximation of the filtering distribution given by the particle filter. In order to evaluate the integrals in (11), we extend the particle trajectory $x_{0:t}^{(i)}$ with $x_{t+1:t+\tau}^{(i)}$. For each particle, $i = \{1, \ldots, N\}$, the predicted trajectory is sequentially computed as

$$x_k^{(i)} \sim p(x_k|x_{k-1}^{(i)}), \qquad x_{0:k}^{(i)} \triangleq \left(x_{0:k-1}^{(i)}, x_k^{(i)} \right) \tag{12}$$

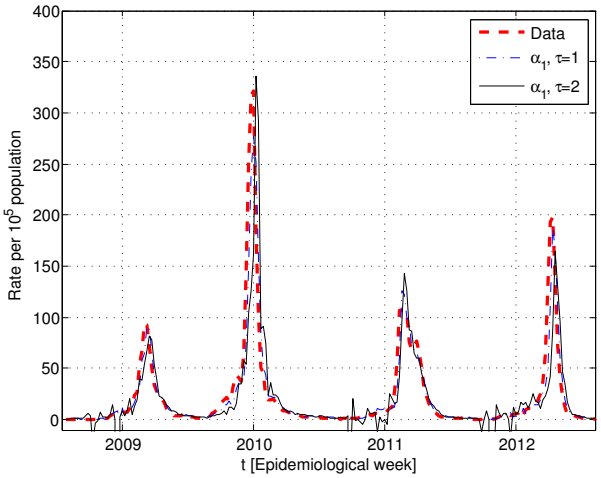

Fig. 2. Prediction results for the 4th epidemiological season. Parameters: $\boldsymbol{\alpha}_1 = [1/3, 1/3, 1/3, 0]$; $\tau = \{1, 2\}$.

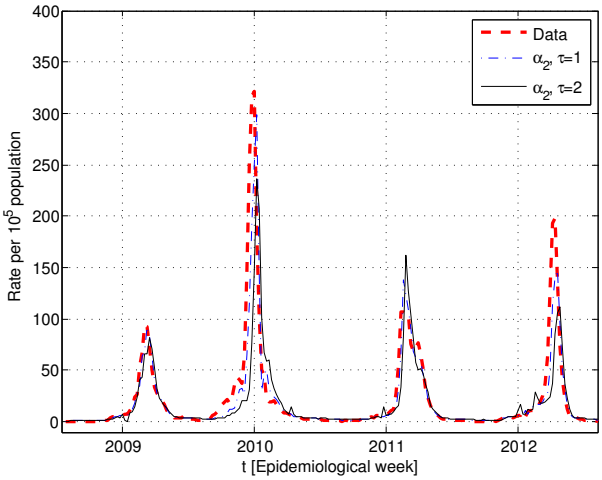

Fig. 3. Prediction results for the 4th epidemiological season. Parameters: $\boldsymbol{\alpha}_2 = [1/2, 0, 1/2, 0]$; $\tau = \{1, 2\}$.

with k from $(t+1)$ to $(t+\tau)$. Then, the algorithm provides an estimate of the τ step-ahead prediction distribution as

$$p(x_{t+\tau}|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(x_t - x_{t+\tau}^{(i)}) \tag{13}$$

from which one can predict the time-course of $x_{t+\tau}$ given measurements up to time t as

$$\hat{x}_{t+\tau|t} = \sum_{i=1}^{N} w_t^{(i)} x_{t+\tau}^{(i)}. \tag{14}$$

As a final step, particle filters incorporate a resampling strategy to avoid collapse of particles into a single state point. Resampling consists in eliminating particles with low weights and replicating those in high-probability regions [19].

4. RESULTS

We used the open database described earlier to train the model, test the prediction algorithm, and assess its performance. Recall that we have four seasons in the database. The goal of the experiments was to predict the evolution of rates in the fourth season. Therefore, it is important to keep in mind when inspecting Figs. 2–3 that only results from the fourth season are meaningful for practical use, and the rest of the seasons' data were used to build the model.

The step-ahead prediction was adjusted via the parameter τ (in units of EW). The larger the value of τ, the poorer the results are expected if the model is not accurately known. We fixed $N = 1000$ particles and $\sigma_y^2 = 10^{-4}$. Figures 2 and 3 show the prediction of influenza incidence rates over time (EWs) with 2 configurations of the weight vector to build the model (equations (3)-(4)), defined as $\boldsymbol{\alpha}_1 = [\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, \alpha^{(4)}]$. Namely, we considered a first configuration where the model was trained using the first three seasons, $\boldsymbol{\alpha}_1 = [1/3, 1/3, 1/3, 0]$; and another configuration where the second season was not considered, since it had a pattern differing from the nominal behavior due to the A(H1N1) pandemic, $\boldsymbol{\alpha}_2 = [1/2, 0, 1/2, 0]$. The results were obtained for τ-step ahead predictions of 1 and 2 weeks.

We evaluated the performance of the predictive method by obtaining the root mean square error (RMSE) of the fourth season, in same units as the influenza incidence rates. With $\boldsymbol{\alpha}_1$, the RMSE was 18.83 and 41.10 for $\tau = 1$ and 2, respectively. For $\boldsymbol{\alpha}_2$, the RMSE was 19.02 and 31.27, respectively. One can conclude that including the second season provides the state-model with larger variances that might help the method when data from new seasons does not follow a "regular pattern." This is the case of the fourth season with respect to the first and the third. For $\tau = 1$, the configuration with $\boldsymbol{\alpha}_1$ shows slightly better results. However, for $\tau = 2$, the inclusion of the second season results in noisy predictive results and it seems more convenient to use $\boldsymbol{\alpha}_2$.

5. CONCLUSIONS

In this paper, we introduced a new algorithm for prediction of the evolution of influenza incidence rates over time. The parameters of the system are obtained using real data and the resulting model is nonlinear. The particle filtering methodology is employed for approximation of the predictive distribution of the state of the system, and by use of the optimal importance density. Computer simulations provide promising results and reveal accurate prediction of the influenza rates. Notice that Eureqa was used for data fitting, but future work includes fitting with simple, yet representative, functions to obtain the model of the influenza evolution.

6. REFERENCES

[1] "Influenza (seasonal)," Tech. Rep. Fact sheet no. 211, World Health Organization (WHO), April 2009.

[2] C. Sonesson and D. Bock, "A review and discussion of prospective statistical surveillance in public health," *Journal of the Royal Statistical Society A*, vol. 166, no. 1, pp. 5–21, 2003.

[3] C. Pelat, P.-Y. Boelle, B. Cowling, F. Carrat, A. Flahault, S. Ansart, and A.-J. Valleron, "Online detection and quantification of epidemics," *BMC Medical Informatics and Decision Making*, vol. 7, no. 1, pp. 1–9, 2007.

[4] R. E. Serfling, "Methods for current statistical analysis of excess pneumonia-influenza deaths," *Public Health Rep*, vol. 6, no. 78, pp. 494–506, 1963.

[5] E. J. Crighton, R. Moineddin, M. Mamdani, and R. E. G. Upshur, "Influenza and pneumonia hospitalizations in Ontario: a time-series analysis," *Epidemiol. Infect.*, vol. 132, no. 6, pp. 1167–1174, 2004.

[6] M. A. Martínez-Beneito, D. Conesa, A. López-Quílez, and A. López-Maside, "Bayesian Markov switching models for the early detection of influenza epidemics," *Statistics in medicine*, vol. 27, no. 1, pp. 4455–4468, 2008.

[7] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 9, no. 457, pp. 1012–1014, 2009.

[8] E. Aramaki, S. Maskawa, and M. Morita, "Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter," in *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*. 2011, Association for Computational Linguistics.

[9] E. Coma, L. Méndez, J. Camús, and M. Medina, "Diagnosticat: a disease surveillance system derived from electronic health record data," in *Proc. of the European Scientific Conference on Applied Infectious Disease Epidemiology (ESCAIDE)*, Stockholm, Sweden, Nov 6–8 2011.

[10] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, February 2002.

[11] "World health statistics 2012," Tech. Rep., World Health Organization (WHO), Geneva, Switzerland, October 2012.

[12] E. S. Allman and J. A. Rhodes, *Mathematical Models in Biology: An Introduction*, Cambridge University Press, 2004.

[13] M. Schmidt and H. Lipson, "Distilling Free-Form Natural Laws from Experimental Data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009.

[14] P. Godoy, T. Pumarola, A. Martínez, N. Torner, A. Rodés, G. Carmona, P. Ciruela, J. Caylà, C. Tortajada, A. Domínguez, and A. Plasència, "Surveillance of the pandemic influenza (H1N1) 2009 in Catalonia: results and implications," *Rev. Esp. Salud Publica*, vol. 1, no. 85, pp. 37–45, Jan-Feb 2011.

[15] J. Zarocostas, "World Health Organization declares A (H1N1) influenza pandemic.," *BMJ [http://www.ncbi.nlm.nih.gov/pubmed/19525308]*, vol. 338:b2425, 2009.

[16] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi, "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic," *PLoS ONE*, vol. 6, no. 8, pp. e23610, 2011.

[17] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Stat. Comput.*, vol. 3, pp. 197–208, 2000.

[18] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Míguez, "Particle Filtering," *IEEE Signal Processing Mag.*, vol. 20, no. 5, pp. 19–38, September 2003.

[19] R. Douc, O. Cappé, and E. Moulines, "Comparison of resampling schemes for particle filtering," in *Proc. of the 4th International Symposium on Image and Signal Processing and Analysis, ISPA'05*, Zagreb, Croatia, Sept. 2005, pp. 64–69.