Early Detection and Assessment of Epidemics by Particle Filtering

C. Jégat, F. Carrat, C. Lajaunie and H. Wackernagel

Abstract Influenza infects from 5% to 20% of the population during an epidemic episode, which typically lasts a few weeks, and, as the conditions leading to an outbreak are not well understood, the moment of its start is difficult to foresee. The early detection of an epidemic would however make it possible to limit its impact by adopting appropriate actions—this is particularly desirable in account of the threat of a major pandemic. As a side-product of the forecasting system an estimation of the total number of infected people in each region at every time step is provided.

We first present the classical epidemiological model for contagious diseases and explain the way regionalization has been incorporated into it. Then we describe the assimilation technique that is used to process the data, which are daily reported cases of influenza-like illness. Tests on simulated data are presented to illustrate the efficiency of the process and finally real observational data from the French Sentinelles network are processed.

Introduction: the *Sentinelles* Network Data

The UMR-S 707 of the INSERM and the Geostatistics Group of the Centre de Geosciences have been collaborating since several years on a prototype system for assimilating data from a sentinel network of voluntary general practitioners into a stochastic epidemiological model (Biboud, 2002; Bui, 2001) The model includes temporal evolution of the epidemic state and a basic regionalization. The improvements made with respect to those studies were mainly concerned regionalization aspects.

The observational data are provided by the French *Sentinelles* network. This network gathers 1200 physicians (among the 60000 general practitioners of metropolitan France) who connect themselves about once every week to the internet to report the number of cases of influenza-like-diseases (as well as 8 other diseases) they have observed among their patients.

H. Wackernagel
INSERM UMR-S 707, Paris, France; Ecole des Mines de Paris - Geostatistics group, Fontainebleau, France
e-mail: hans.wackernagel@ensmp.fr

A. Soares et al. (eds.), *geoENV VI – Geostatistics for Environmental Applications*, 23–35

The observed cases may give an idea of the evolution of the epidemic, but they are subject to many uncertainties. First, the diagnosis leading a physician to report a case is only based on the symptoms of the disease and usually no virologic analysis is carried out. The physician mainly looks for the characteristic symptoms of influenza: sudden fever above 39°C, breathing troubles and muscle soreness. Now, these symptoms could also the consequence of other diseases and it is also possible that some patients could have different symptoms, or no reaction at all, and still be contagious.

Furthermore, an unknown proportion of the population does not consult a doctor when they have got influenza. In this study, given the lack of precise data and according to estimations made at the INSERM, this proportion is assumed to be 50%.

Finally, doctors do not regularly connect themselves to the network, and sometimes forget to report an absence of cases, and yet this would also be an important information. This irregularity also may lead to an accumulation of cases on the day of each connection, whereas the patients that were reported may have consulted already a few days earlier.

The cases have therefore been redistributed over the period prior to correction to take this irregularity into account. A *Sentinelles'* doctor is considered as active if the gap between two consecutive connections is inferior or equal to 12 days. For example, if a doctor connects on a Friday, and his last connection was on the Monday of the same week, he will be considered active on every day from Tuesday to Friday, and if he reports 6 cases, they will be spread out uniformly over these 4 days, that is 1.5 cases per day. This uniform distribution may appear unrealistic since it is more probable that a case reported one day has consulted the same day or the day before, rather than several days earlier. However, this reallocation, even though imperfect, seems acceptable and has been adopted for this study.

In this study, the time scale will be the day, in order to enable an early detection of the epidemic, and the geographic scale will be the French administrative regions (all 21 of the metropolitan, continental regions). Figure 1 shows the number of cases reported in four different regions of France for the last 5 years, as well as the number of active doctors.

1 The Epidemic Model

The SIR model, widely used to model epidemic diseases (Daley and Gani, 1999; Andersson and Britton, 2000) partitions the populations into three categories:

- *Susceptible*: the people free from the disease and without specific immunity,
- *Infected*: the people infected by the virus and who are still infectious,
- *Removed*: the people who have recovered from their illness, are no longer infectious and are immunized.

We will consider a discrete time-space structured model. The population in each region x at time t is partitioned into by $S_t(x)$, $I_t(x)$ and $R_t(x)$. The total population

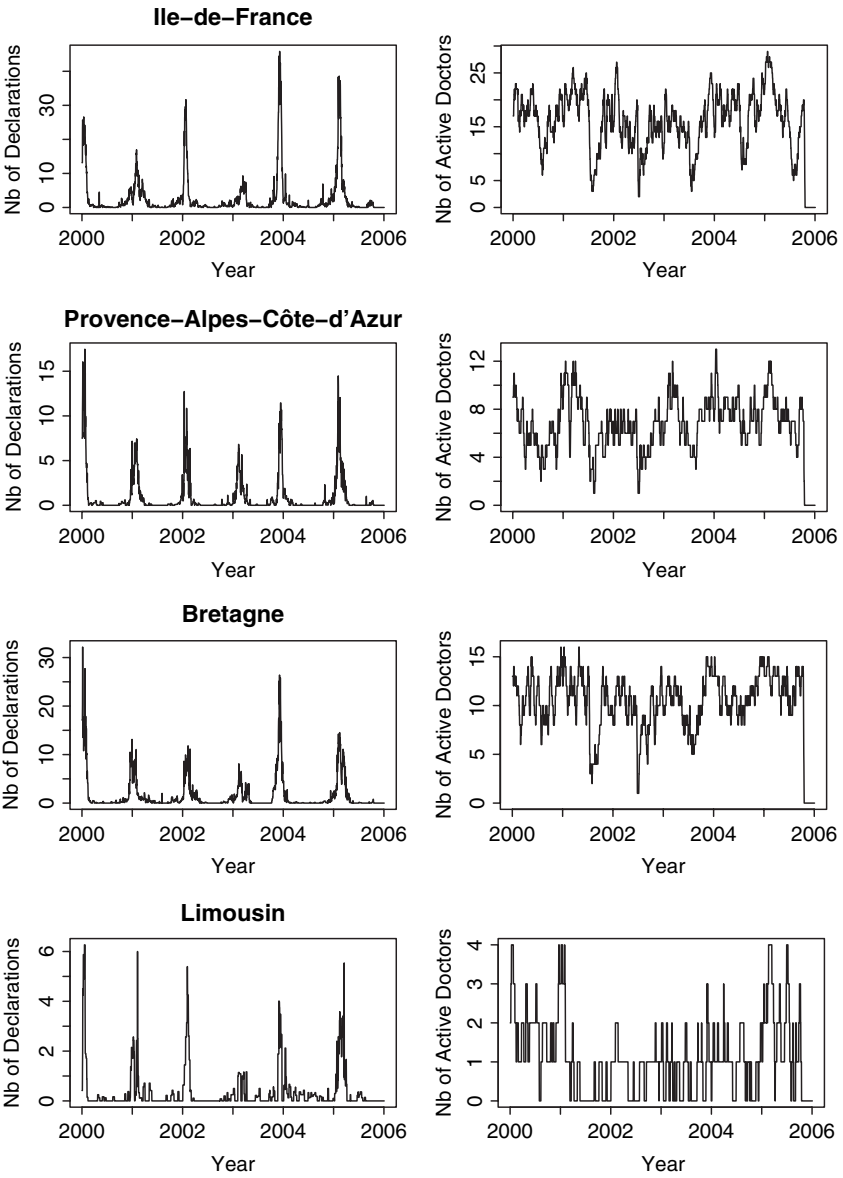

Fig. 1 Daily declarations of the *Sentinelles* network and number of active doctors over the last 5 years, in four French regions

in each region $N(x)$ is assumed constant during the course of an epidemic, so we have $\forall t$: $N(x) = S_t(x) + I_t(x) + R_t(x)$.

The new cases in a region x result from contacts between susceptibles in given region x and infectives in the other regions y. If the contact rate per time unit between x and y is denoted by $a(x, y)$, and if each $S{-}I$ contact produces a

contamination with probability τ, each susceptible avoids contamination during a time interval with probability:

$$q(x,t) = \prod_y \left(1 - \tau \frac{I(y)}{N(y)}\right)^{a(x,y)}$$

Therefore, under this simple model which assumes independence of contamination events, the number of new infectives at x is a binomial variable with parameters $S(x)$ and $1-q(x,t)$.

The observed behavior of real epidemics, in which bursts follow smooth periods, suggests a sharp increase of the probability of contamination τ during epidemic episodes. This increase could be due to the income of a new virus strain. Thus we modeled τ with independent two-state Markov chains for the different regions and write accordingly $\tau_t(x)$.

The next step in the modeling is about the recovery from disease. A Markovian structure, when only one category of infective is considered, would imply an exponential distribution of individual infectious periods. To obtain more realistic distributions, without sacrificing the Markovian property of the model, we further divided $I_t(x)$ into $I_t^0(x)$, ..., $I_t^6(x)$ according to time since infection. Recovery probability for each category in one time step is $p_{rec}^k; k = 1, \cdots, 7$, with $p_{rec}^7 = 1$, since it is assumed that infection cannot last more than 7 days.

Next we need to model how the declarations inform us on the epidemic state. We assume that each new infected subject is registered by *Sentinelles* with a known probability, which depends on the coverage of the region. Thus the number of declarations in x is conditionally binomial, with parameters $S_{t-1}(x) - S_t(x)$, $p_{dec}(x)$.

2 Regionalization of the Model

This section describes the calculation of the contact rate $a(x, y)$, which is a main parameter of the regionalized model.

2.1 The Origin-destination Matrix

A common tool in population flux modeling is the *origin-destination matrix*. This matrix is defined by: $M = (M_{xy})$ with $M_{x,y}$ being the number of people moving from x to y per time unit.

There are several ways to calculate this matrix. To describe precisely the population fluxes between regions, the ideal way would be to have access to the road and rail traffic data. Unfortunately, these data are not easy to find and we will finally use a mathematical model of the fluxes instead.

2.2 The Gravitational Model

The *gravitational model*, inspired by Newton's laws, assumes that the flux from x to y is proportional to the populations in x and y and also inversely proportional to a power of the distance between the regions. We thus write

$$T_{xy} = k\frac{P_x^{\lambda} P_y^{\alpha}}{d_{xy}^{\beta}}$$

with:

T_{xy} the number of people moving from x to y,
P_x the population at x,
d_{xy} the distance between x and y (based on their main cities),
k a proportionality constant depending on the time step,
λ the force of emission of the source,
α the force of attraction of the destination, mainly on the basis of its economic activities,
β the transport friction, depending on the efficiency of the transport system.

The last three parameters should be adjusted with observed data, which however lacked in this study, and so λ and α were set to 1, whereas β was set to 2.

2.3 Contact Probabilities

The *origin-destination matrix* can be set up in different ways. One of them assumes that the people effectively move, leading to a revision of the total population within each region at each step. This approach may be acceptable in the case of important population fluxes, for example on holidays, but it is far too rough to describe daily contacts between regions, which are mainly caused by commuting.

We choose a different approach and define the *contact matrix*. We assume that people from two regions are in contact and therefore can infect each other, but at the end of each time step, everyone is back home. The population within each region therefore remains constant. We further assume that the disease does not affect the population movements.

The contact matrix is generated according to the following choices:

- The total number of contacts for a single person (we set it to 100 in this study).
- We then determine the number of contacts each person has with people in the same region. In the present setting, the model assumes that the more a region is populated, the more it is attractive, and the lower is the within-region contact rate. This led to the formula: $N_{xx} = K/Pop(x)$ with $K = 0.000003$.
- The contacts with people in other regions are split on the basis of the origin-destination matrix obtained from the gravity model. Whatever the constant k,

this does not affect the final result since everything is driven by the choice of the number of total contacts and the rate of internal contacts.
- Lastly, the matrix has to be made symmetric. The numbers of total contacts are allowed to vary from a region to another, but they remain close.

3 Data Assimilation with Interacting Particle Simulation

Data assimilation is a term used by meteorologists and oceanographers to refer to the integration of real time observational data into simulation models. Two ingredients are required for this, one being a mechanistic model of the time evolution of the process. This model, which is not necessarily deterministic, is usually based on the physics of the process. The other ingredient is an observational equation, which should realistically model what the observed data tells us about the state of the process.

Sequential data assimilation is often based on some form of Kalman filtering (Bertino et al., 2003) In this paper, we present the application of an alternative method, known as *particle filtering* (Doucet et al., 2001; Oudjane, 2000; Cauchemez, 2005) which can handle highly non-linear dynamics at the expense of performing a large number of simulations in parallel. Approximations of the state conditional distribution are obtained sequentially and are based on populations of simulations (i.e. the particles).

3.1 The Principles of Particle Filtering

Let N_d be the number of regions and N_s the number of simulations. To notation, we denote the state vector at time t by:

$$X_t = \{\tau_t(x)\,,\, I_t(x)\,; x = 1...N_d\}$$

where I is itself structured according to the elapsed time since contagion. The simulated particles at time t are indexed by an exponent and so X_t^i refers to the i^{th} simulation. Writing $D_{0:t} = D_0, \cdots, D_t$ the observational data up to time t, we look for an approximation of the conditional distribution

$$f(X_t\,|D_{0:t})$$

of the current state given the data. Note that this is a less demanding objective than that of approximating the whole trajectory. Bayes formula, which tells how the exact conditional distribution should be updated, gives a clue on how this should be done. We have:

$$\begin{aligned} f(X_t|D_{0:t}) &= f(X_t|D_t, D_{0:t-1}) \\ &= \frac{f(X_t, D_t|D_{0:t-1})}{\int f(x_t, D_t|D_{0:t-1})\, dx_t} \end{aligned}$$

Now, using two conditional independence properties of the model:

$$\begin{aligned} f(X_t, D_t|D_{0:t-1}) &= \int f(X_t, x_{t-1}, D_t|D_{0:t-1})\, dx_{t-1} \\ &= \int f(D_t|X_t, x_{t-1}, D_{0:t-1})\, f(X_t\,|x_{t-1}, D_{0:t-1}) \\ &\qquad f(x_{t-1}|D_{0:t-1})\, dx_{t-1} \\ &= \int f(D_t|X_t)\, f(X_t\,|x_{t-1})\, f(x_{t-1}|D_{0:t-1})\, dx_{t-1} \end{aligned}$$

we get the update equation:

$$f(X_t|D_{0:t}) = \frac{\int f(D_t|X_t)\, f(X_t\,|x_{t-1})\, f(x_{t-1}|D_{0:t-1})\, dx_{t-1}}{\int f(D_t|x_t)\, f(x_t\,|x_{t-1})\, f(x_{t-1}|D_{0:t-1})\, dx_t, dx_{t-1}}$$

In these equations:

- $f(X_t|x_{t-1})$ represents the system's *evolution* distribution. The term:

$$f(X_t\,|x_{t-1})\, f(x_{t-1}|D_{0:t-1})$$

  can be interpreted as a one step free evolution of the system, starting from a random initial state, generated according to the conditional distribution at $t-1$.
- $f(D_t|X_t)$ represents the *observational process*. It can be interpreted as the likelihood of X_t relatively to data D_t.

The updated distribution can thus be viewed as a reweighted version of the predictive one. The assimilation method follows this algorithm.

To describe the sequence, we assume that at $t-1$ an approximation of $f(X_{t-1}|D_{0:t-1})$ is available in the form of an empirical distribution of N_s particles:

$$\widehat{f_{t-1}} = \frac{1}{N_s}\sum_{i=1}^{N_s} \delta_{X_{t-1}^i}$$

The steps of the algorithm are then as follows:

Prediction step. Free evolution of each particle, according to the model:

$$X_{t-1}^i \to \widetilde{X_t^i}$$

When, as it is the case for the present SIR model, this evolution is stochastic, independent paths are generated.

Conditioning on D_t. This involves the likelihood calculation, $f(D_t|\widetilde{X_t^i})$ and its normalization, which yields the following probability masses:

$$\omega_i = \frac{f(D_t|\widetilde{X}_t^i)}{\sum_j f(D_t|\widetilde{X}_t^j)}$$

At this point, we have the following approximation of the conditional distribution:

$$\widetilde{\mathrm{f}}(X_t|D_{0:t}) = \sum_i \omega_i \, \delta_{\widetilde{X}_t^i}$$

Resampling. The iteration is not yet complete, for we do not have an approximation having the form of an empirical distribution. The idea is to draw a new set of N_s particles independently from $\widetilde{\mathrm{f}}(x_t|D_{0:t})$. This is just a weighted resampling of the set $\widetilde{X}_t^1, \cdots, \widetilde{X}_t^{N_s}$ favoring the particles having the highest likelihood.

There are numerous possible variations of this algorithm. For example the independent resampling, which is not very efficient, can be replaced by better schemes regarding the representation of $\widetilde{\mathrm{f}}$ by empirical distributions.

4 Results

In order to test the efficiency of the model, we have to test the algorithm on simulated data first. The simulation algorithm randomly generates a time series of Markov chain states based on the transition matrix. Then, at each time step, in each region, given this state, the algorithm generates simulated new cases and recovered, computes the susceptible, infected and removed and finally generates the reported cases given the new cases. We then only keep the reported cases and assimilate them. If the algorithm is efficient, we should recover a Markov-chain state time-series similar to the one that generated the cases. We can also compare the estimation of the number of infected to the simulation.

4.1 Results on Simulated Data

In order to test the efficiency of the algorithm, we assimilate simulated observational data (of which we know all the parameters, especially the Markov chain states for all t), hoping to find back the same Markov chain states.

Figure 2 shows the original Markov chain, the simulated cases, the estimated Markov chain and the ratio of the simulated number of infected against the estimated number of infected for two regions.

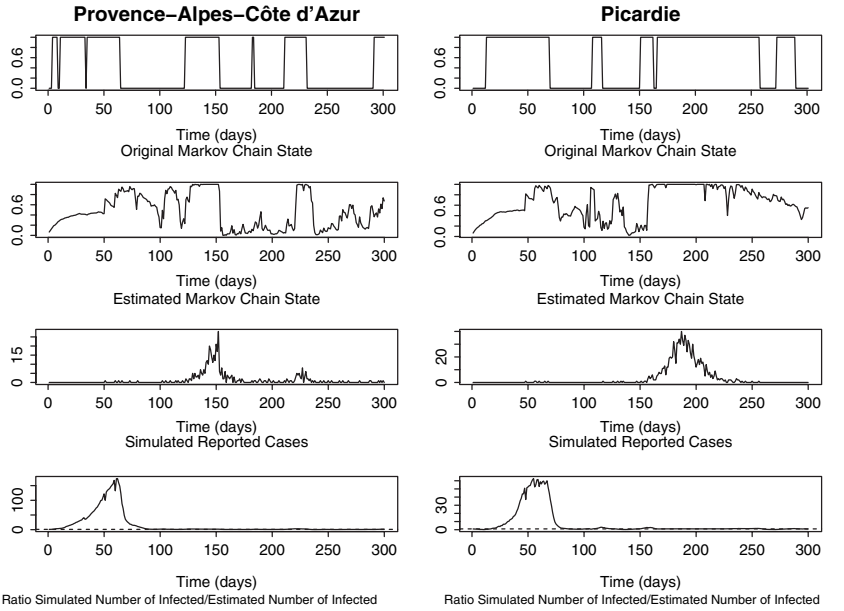

Fig. 2 Assimilation of simulated data (from top to bottom): original Markov chain, estimated Markov chain, simulated declarations, ratio of simulated/estimated number of infected

4.1.1 First Facts

Problems at the Beginning of the Time Period

There are two estimation problems at the beginning. During the time when there are no reported cases, or only few (less than two per day per region), the algorithm is in trouble for finding in which state the Markov chain is, and it underestimates the number of infected.

The underestimation is due to the fact that, in spite of the increasing number of infected in the simulation, the declaration probability is so low that there are no declarations for a long time. Therefore, the algorithm finds out only quite late (at the time of the first reported case) that there are more infected than estimated. Then it catches up on the number of infected by only keeping the particles in an epidemic dynamic and the number of infected is correctly estimated.

In particular, the likelihoods of the particles do not enable to decide between the two types of particles: because the number of infected is low, and because the declaration probability is low, even particles in an epidemic dynamic do not generate enough infected individuals to have a reported case. In that situation both types of particles are just as likely, and the algorithm cannot decide between them.

As said previously the algorithm has to catch up on the number of infected in the early stage of an epidemic. Therefore, only particles in an epidemic dynamic are kept, so that the number of infected progressively increases. Sometimes this is right, at other times the number has increased only because of a contagion within other regions.

4.1.2 Sensitivity Due to Parameter Variations

In this section we discuss the influence of the parameters which are most difficult to estimate.

Transition Probability

In the case of a simple two-state Markov chain model, where we chose to have the same transition probability from 0 to 1 than from 1 to 0, there is only one parameter. It is not actually a parameter which has to be known precisely. If we assume that this probability is totally unknown in reality, then in the assimilation algorithm that probability is only the proportion of particles generated with the same state as the previous particles. Therefore, even if only a few particles are in the right state, these are the ones which are kept in the resampling step. We only have to choose a probability high enough to generate a small number of particle at each step. On the other hand, a too high probability would generate too many epidemic particles at the beginning and lead to an overestimation of the number of infected individuals—even more than it already does. The good news is, however, that in the assimilation of simulations this probability actually has shown to have very little influence on the results.

Contamination Probabilities τ_i

Every contamination probability has to be calculated with reference to a single one. Indeed, in the case of a simple model, an epidemic can start if each infected can infect more than one susceptible. In this model, the number of susceptibles contaminated by an infected is defined as $R = \tau \cdot C \cdot D$, τ being the contamination probability of an S-I contact, C the number of contacts per person, and D the average length of the infection. The probability τ_l above which the epidemic can start is therefore $\tau_l = \frac{1}{C \cdot D} = 0.0018$ in our case, where $C = 100$ and $D = 5.5$. Tests show that even though our model is more complex, this probability does correspond to a case when the epidemic neither starts nor is completely extinct: the number of infected stays constant.

Next, the high and low values of τ must be chosen on each side of τ_l.

The low value does not have a great influence on the results. It only affects how fast the infection stops when dynamics are non-epidemic. However, to avoid total extinction, randomly chosen infected are introduced.

The high value is very hard to estimate from the data. If the state of the Markov chain were known at each time step, the probability could be estimated thanks to the total number of infected and the length of the epidemic. But given that we know neither the state, nor the probability corresponding to the state, this estimation is much harder.

A first rough estimation can be done, prior to the assimilation, by simulating a time series with constant epidemic dynamics. In that case we can avoid to choose a τ for which the epidemic is too strong or too weak.

Another approximate way to adjust this value is simply to look at the estimated Markov chain state. If the value stays high all the time, it is likely that the probability was chosen too low. If the value is oscillating throughout the epidemic period, it is is likely that the probability was set too high.

The Number of Initial Cases

The number of cases in the non-epidemic period is very hard to estimate. Actually, we do not know if the start of an epidemic is caused by a small number of infected remaining from a previous epidemic, or whether it is due to infection coming from abroad. Since we start the assimilation at a time when we know, almost for sure, that the epidemic has not yet started anywhere, this number of infected represents a very small proportion of the population. Moreover, given that we inject randomly a few infected to prevent the extinction of the infection, even if that number is very small, there will be a compensation from this injection. And if it is too high (which can only mean that it generates at least one declaration whereas it should not), it will tend to rapidly decrease.

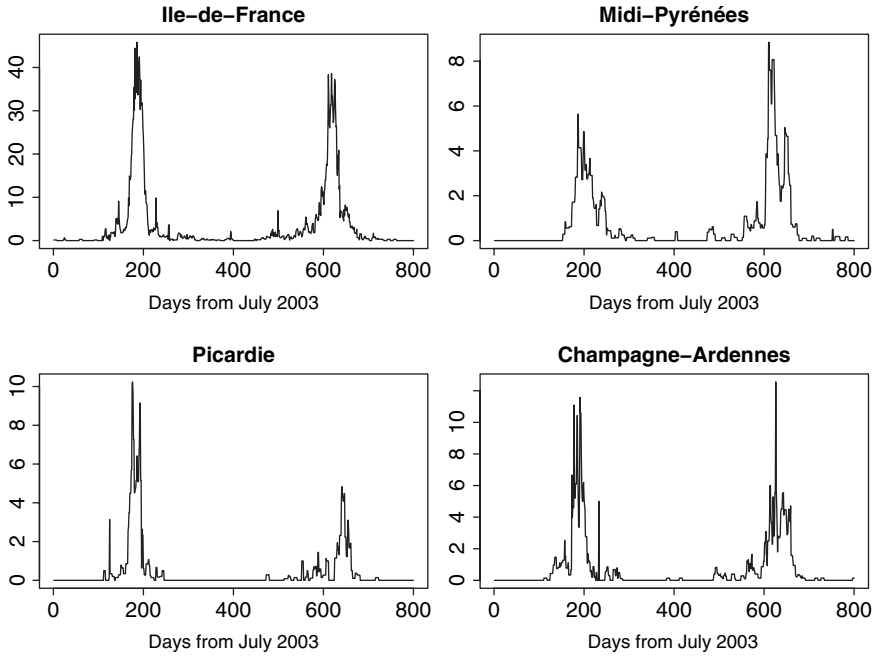

Fig. 3 Reported cases from July 2003 to September 2005

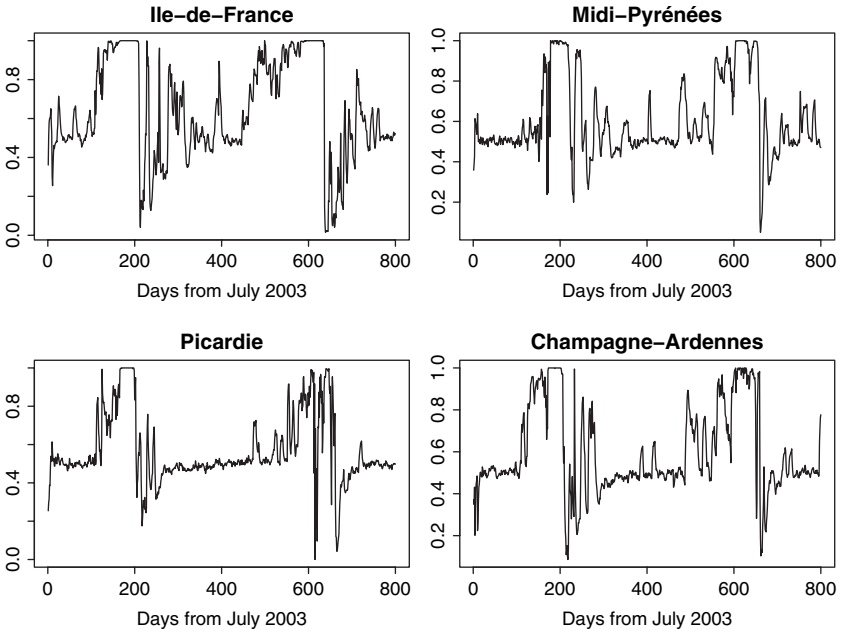

Fig. 4 Assimilation of 2003–2005 data: estimation of the two-state Markov chain

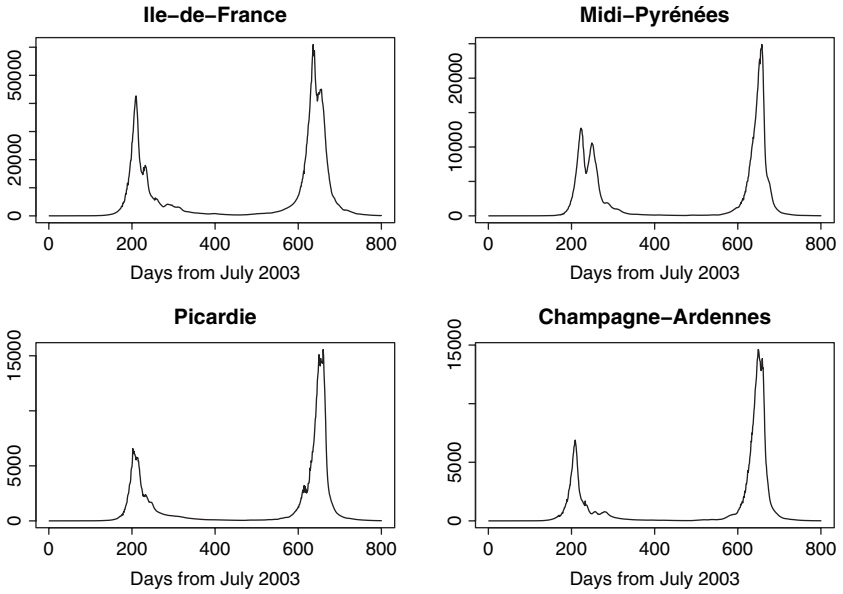

Fig. 5 Assimilation of 2003–2005 data: estimation of number of infected individuals

Other Parameters

The declaration probability (mainly based on the proportion of active doctors) and the recovery probability are relatively well known. This is to say that, given the confidence with which we know them, variations inside that confidence interval did not have a great influence on results.

4.2 Results on Real Data

Finally, we confronted the algorithm to real data taken between July 2003 and July 2005 as shown on Fig. 3. The data were assimilated into a two-state Markov chain model with $\tau_1 = 0.0006$ and $\tau_2 = 0.0027$. Results on Fig. 4 show the seemingly hesitating oscillation of the Markov chain between its two states in non-epidemic periods, but overall the algorithm appears to react fairly quickly to an increase in the number of declarations.

An important side-product of the present forecasting procedure is that it provides an estimation of the total number of infected people in each region at every time step. Fig. 5 displays the time evolution of the estimated number of infected individuals during the two epidemics.

References

Andersson H, Britton T (2000) Stochastic epidemic models and their statistical analysis. Lecture notes in statistics, vol 151. Springer-Verlag, New York, p 137

Bertino L, Evensen G, Wackernagel H (2003) Sequential data assimilation techniques in oceanography. *Int Stat Rev* 71:223–241

Biboud E (2002) Modélisation des épidémies de grippe. Technical Report S-436, Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau

Bui VAP (2001) Modélisation par chaînes de Markov cachées d'une épidémie de grippe. Technical Report S-420, Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau

Cauchemez S (2005) Estimation des Paramètres de Transmission dans les Modèles Epidémiques par Échantillonnage de Monte Carlo par Chaîne de Markov. PhD thesis, Université Pierre et Marie Curie, Paris.

Daley DJ, Gani J (1999) Epidemic modelling: an introduction. Cambridge series in mathematical biology. University Press, Cambridge, p 213

Doucet A, de Freitas N, Gordon N (eds) (2001) Sequential Monte Carlo Methods in Practice. Springer-Verlag, New York

Oudjane N (2000) Stabilité et Approximations Particulaires en Filtrage Non-linéaire: Application au Pistage. PhD thesis, Université de Rennes I, Rennes