11^{e} CONGRES INTERNATIONAL DE GENIE INDUSTRIEL – CIGI2015
Québec, Canada
26-28 octobre 2015

Waiting-time Estimation in Walk-in Clinics

JULIO MONTECINOS[1], MUSTAPHA OUHIMMOU[2], SATYAVEER S. CHAUHAN[3]

[1] Dép. de Génie de la Production Automatisée, ÉTS Montréal
1100 Rue Notre-Dame Ouest, Montréal, QC H3C 1K3, Canada
julio.montecinos@etsmtl.ca

[2] Dép. de Génie de la Production Automatisée, ÉTS Montréal
1100 Rue Notre-Dame Ouest, Montréal, QC H3C 1K3, Canada
Mustapha.Ouhimmou@etsmtl.ca

[3] Supply Chain & Business Technology Management (JMSB), Concordia University
1450 Rue Guy, Montréal, QC H3H 0A1, Canada
Satyaveer.Chauhan@concordia.ca

***Résumé* – Dans la province du Québec, les cliniques sans rendez-vous font en grande partie de l'offre d'assistance médicale. Ces cliniques doivent organiser leurs patients en queues et maintenir leurs patients en ordre. Depuis peu, des compagnies privées offrent un service aux cliniques pour l'émission de tickets numérotés servant à gérer la queue, et aux patients un service payant qui les préviendra lorsque leurs consultations s'approchent. Avec ce service les patients peuvent utiliser leur temps libre qu'ailleurs la salle d'attente. Un modèle basé sur les Filtres Particulaires (PF) et les Modèles de Mélanges Gaussiens permet d'estimer le temps d'attente pour chaque consultation, en utilisant les données historiques et l'information de nouvelles consultations. Le système tient compte de deux types de patients les « réguliers » (non-payants) et les « privés » (payants). Notre méthode donne une estimation plus précise du temps d'attente avant la consultation que les méthodes statiques simples.**

Abstract – In the province of Quebec, medical assistance is offered by walk-in clinics. These clinics must keep track of patients' turn in line. Some private companies offer an extra paid service to walk-in clinic patients that notifies them when their consultation approaches, so patients can use their free time elsewhere than the waiting room. A model based on Particle Filters and mixture models helps to estimate the waiting time for each consultation, using historical and new incoming data from patient consultations. The system considers two types of patients, namely regular and private. Our method gives an estimate of the waiting time for consultation better than simple statistics.

***Mots clés* – Méthode Monte-Carlo, Modèle de Markov caché, Filtres Particulaires, Série temporelles, Simulation.**

***Keywords* - Approximate Monte-Carlo methods, Hidden Markov Models (HMM), Particle filters (PF), Time Series analysis (TS), Computer simulation.**

1 INTRODUCTION

During recent years, several companies have been offering services for patients' queuing and schedule management for walk-in clinics in the province of Quebec which do not work with appointments. These services have in common a system that identifies each patient with a code that puts them into a virtual queue. The queue generally considers the patient's arrival order. An extra and paid follow-up service allows "private patients" to wait elsewhere. This paid follow-up service consists of notifications sent by telephone, message, or displayed in a web application. These systems do not change the general way that walk-in clinics provide their service. Usually walk-in clinics keep patients (regular and private) in FIFO logic, observing very particular rules for their management. After a first triage done by the medical staff, some patients are deferred to emergency and other serious cases are re-scheduled to pass first. The remaining patients retain the FIFO logic of the arrival ordering. The use of service codes makes it easy for clinics to keep a record of the patient in the queue. The company servicing the clinics will continue the follow-up to the patient, allowing them to use their spare time wisely, and especially to avoid extra waiting time in the clinic's facilities. The clinics are also rid of the burden of unsatisfied, frustrated people in their facilities and maintain the presumption of fairness of the FIFO logic for the less severe ones. Another advantage for patients and clinics is the reduced amount of time they are exposed to contagious diseases.

I

The FIFO queue logic maintains the system flow and keeps the medical staff well supplied with patients to their maximum capacity. Every clinic has its capacities driven by the medical personnel in the legal time they have to serve and it is highly variable. Clinics remain open if they have patients to serve, but they generally do not consider new patient arrivals after the first early queuing process and related triage. From the point of view of the clinic and medical staff, the system of "clinic queues" remains very efficient by considering a single queue for several doctors. From the point of view of the patient, the paid follow-up service is very convenient when the price is less than the value of patients' time. The system is not compulsory for "regular patients" that remain in the clinics, so it is inherently fair for them.

The paid service in use is based on the patient's ticket delivery by clinic staff and further ticket-call in the waiting room. Therefore, the service can register "input time" and "output time" for every patient, and indirectly estimate the patient's consultation time. "Waiting time" begins when staff delivers a ticket to the patient. "Service Time" is given by patient's entering/exiting the consultation room. We note that the number of doctors per queue is supposed to be one and staff is in charge of advancing the queue. The final record is imprecise and noisy, but keeping the pace of the process as is, far from ideal. We note that service time is very random (and the process is non-stationary), so it is possible that patients in the follow-up paid system cannot get to the clinics in time because they do not receive the alert soon enough and the staff can penalize delayed patients. The mix among "patients waiting in the clinics" and the "patients' paid follow-up," allows clinics to continuously feed patients to doctors when there are no-shows or delays from "private patients". It is also important to note that medical staff need some extra time to fulfil their duties; time that is usually considered as part of the consultation. Other disturbances as breaks, lunch times, or staff changes etc., introduce discontinuities into the process.

By making some assumptions on patients' characteristics and symptoms and on doctors' behaviour, it would be possible to profile the waiting time for patients with more precision, (along with other internal characteristics of clinics) but today it is not yet possible to have access to symptoms and doctors' information. With the collaboration of one industrial partner running the notification service (ChronoMetriq), we have access to several time series from different clinics where the process has a clearer structure. Based on this data, we propose an algorithm that is able to make estimations of the patient's consultation time as it is seen from the queue perspective, i.e. independent of the number of doctors serving the queue. Estimations should adapt dynamically and contain the anomalies of the process (without explicating them). After all, "waiting time estimation" is the very important information for patients to manage their free time, but also for medical staff and managers to improve their procedures based on this information. For practical considerations, the calculations need to be at a minimum to allow several estimations in different clinics with available computer power.

This paper is organized as follows. Previous research is in Section 2. Section 3 introduces the Mathematical model. Then, we present a Case Study in Section 4. Finally, conclusions are in Section 5.

2 PREVIOUS RESEARCH

For walk-in clinics, the most prominent advantage of using the FIFO logic patients' service resides in not wasting medical and nursing time, in comparison with the rigidity of a schedule-block assignment, which generally presents high variability of patients, and the also common no-shows and delays that a scheduled system of appointments cascades (Barron, 1980; Tai & Williams, 2012). These annoyances or difficulties in scheduling appointments and using appointment scheduling rules (Klassen & Yoogalingam, 2009) and overbooking (Huang & Hanauer, 2014; Zacharias & Pinedo, 2014) are documented in the literature (Cayirli & Veral, 2003; Jerbi & Kamoun, 2011; Qu, Rardin, Williams, & Willis, 2007; Qu & Shi, 2011; Robinson & Chen, 2010; Sampson et al., 2013), even for more flexible same-day scheduling, open access scheduling and advanced access scheduling (Green & Savin, 2008). Some clinics, also allow a few scheduled appointments (fixed appointments) for special patients that get into queue operation without losing efficiency. These operations systems have been studied in (Qu et al., 2007; Qu & Shi, 2011). The systems in use also do not force any constant, minimum or maximum consultation waiting time, leaving this criterion to the medical staff. This seems to assure the smooth development of the consultation, which is not the case in other systems (Silverman et al., 2012).

Time-series analysis (ARIMA models) and queuing theory are common tools in healthcare analysis (Bastani, 2009; Channouf, L'Ecuyer, Ingolfsson, & Avramidis, 2007). These models are especially useful for single departments or care services, e.g. emergency rooms, surgery units, specialty consultations, etc. (Zonderland, 2012). These tools deal with practical problems when data remain stationary; the natural model stays linear and any added noise is bounded and Gaussian. In many cases it is also possible to approximate this behaviour. For more complex processes (patients through several hospital units subject to several chained decisions) (Hulshof, 2013; Patrick, Puterman, & Queyranne, 2008), Markov and Semi-Markov Processes or Bayesian Networks and simulations give strong assistance in decision analysis and decision support. The strength of these techniques seems to be useful for monitoring flow and diagnosis of patients through different facilities, but also useful for outpatient care settings (Côté & Stein, 2007; Salzarulo, Bretthauer, Côté, & Schultz, 2011; Swisher, Jacobson, Jun, & Balci, 2001).

The steps in between seem open to more exploration: the mining of very available non-stationary operational logging data with the interaction of non-documented administrative procedures and human decisions (including mistakes) in small care units. The waiting time of queued patients in walk-in clinics is part of the former. Walk-in clinics as different from major hospital units do not have a constant flow-chart monitoring for patients. Changes in the habits of patients, doctors, nurses and administrative personnel are kept in the operational loggings. This data have many sources of variability. For their nature (non-stationary and weakly structured) is very difficult to analyze and it is almost impossible to use as a decision-support tool when combining with real time data.

I

On the available tools to treat these problems, Filtering, in signal processing, is very promising. Among these very much noticed techniques, Kalman Filters (KF) are used for de-noising streams of observed data and correcting them for inaccuracies, providing a statistical optimal estimation considering the possible system state (Gustafsson, 2010; Li, 2014; Ristic, Arulampalam, & Gordon, 2004). KF need the system to be linear and the noise to be Gaussian, but have several extensions to overcome these limitations. Nonlinear filters are also an alternative to traditional Time Series analysis for complex systems (Carpenter, Clifford, & Fearnhead, 1999; Mukherjee & Sengupta, 2010). Recent studies have rediscovered Cluster Weighted Modelling (N. Gershenfeld, Schoner, & Metois, 1999) and Gaussian Mixtures Models and have also extended PF to Feynman-Kac Particle process.

2.1 Particle Filters

Sequential Monte Carlo family methods are known as PF (Gordon, Salmond, & Smith, 1993; Rosenbluth & Rosenbluth, 1955). Most practical PF consider Bayesian inferences based on approximating posterior distributions of the interesting variables. PF can be very demanding on calculation power, first in choosing appropriate a priori distributions based on observed past data (Shi & Han, 2007) and second, applying Important Sampling (IS) to substitute the posterior distribution by a set of particles which evolves in a loop, adding observed information. If the initial particles “collapse” during the adaptation, auxiliary steps “regenerate” the particles to maintain the ongoing process (Cappe, Godsill, & Moulines, 2007). Most of the practical PF have compromising solutions on the PF parameters (i.e. number of particles, the re-generation step, etc.), the availability of observed data and the available computer power. Its use has been widely extended for dynamical models, signal processing, and dynamical models (Cappe et al., 2007; Chen, 2003)

In this application, we are interested in tracking a time-varying process (consultation time) without knowing the process dynamics (e.g. we cannot observe the process directly). For simplicity, we consider unobserved “consultations” $(x_t; t \in Z)$ modeled as a stochastic process in a measurable space, with initial distribution $p(*)$. Assuming this process as Markovian, with transition distribution $f(x_t|x_{t-1})$. The observations $(y_t; t \in \mathbb{N})$ are conditionally independent of the unobserved process with marginal distribution $p(x_t\,|y_t)$. The model in question is a general state-space model:

$$\begin{array}{ll} Y_t \quad = G_t(X_t, W_t), & \textbf{(1)} \\ X_{t+1} = F_t(X_t, V_{t+1}), & \textbf{(2)} \end{array}$$

where V_t and W_t are vector realizations of independent random variables (like noise, F_t and G_t are functions in R). **Eq. (1)** is called observation equation and **Eq. (2)** is the state transition equation. Because we are interested in the patients’ waiting-room time through the forecasting of the patient’s consultation time of previous patients using a PF. The waiting time will be the summation of consecutive patient consultations. In this situation, there is no other physical model.

The Bayesian theory is applied to the estimation of the posterior distribution $p(x_{0:t}|y_{1:t})$, also called filtering distribution $p(x_t|y_{1:t})$, of the state x_t. We suppose having $y_{1:t}$ observations to use. The filtering distribution is recursively estimated then with,

$$p(x_{0:t}|y_{1:t}) = \int f\,(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx$$

$$p(x_t|y_{1:t-1}) = \frac{g(y_t|\,x_t)p(\,x_t|\,y_{1:t-1})}{\int g(\,y_t|\,x_t)p(x_t|\,y_{1:t-1})dx}$$

Or as a discrete representation,

$$\pi_k(x_{0:t+1}|y_{1:t})\ \alpha\ \Pi_{i=1}^{t} g(y_i|x_i)\Pi_{i=1}^{t} f(x_i|x_{i-1})p(*)$$

Depending on the types of functions in **Eq. (1)** and **Eq. (2)** and the noise, KF are the approach for the linear/Gaussian case.

As is known, because there is no general solution, numerical approximations are necessary. In addition, our problem can express a multi-modal distribution that adds complications for linear approximations (in KF). The observations of several histograms from clinic service time support this reasoning for the case. As it is, PF have a potential success.

The goal will be to draw a set of i.i.d samples (or N particles: $x_t^i, i = 1..N$), and a set of weights $w_t^i, i = 1..N$ that approximates $p(x_{0:t}|y_{1:t})$, and updates them when needed at a time “$t+m$”, $(m > 1)$, with a new “y_{t+m}”. By means of an (auxiliary) importance distribution, $h_k(x_{0:t}|y_{1:t})$ and the set of weights chosen as, $(x_{0:t}) = p(x_{0:t}|y_{1:t})/\pi_k(x_{0:t}|y_{1:t})$, we approximate,

$$\pi_k(x_{0:t}|y_{1:t}) \approx \sum_{i=1}^{N} w_t^i\,\delta\big(x_t^i\big)\,dx_{0:t}.$$

The PF estimates the sequence of states $x_{1:k}$, rather than x_k. Particles can adapt positions (re-generation) and weights based on consultation observations. However, there is a tendency to particle accumulation where the support of the posterior density is bigger. This is in detriment of other subspaces, “modes” or extremes, that can collapse under badly distributed observation sequences. For some applications, this behaviour is highly desirable (e.g., localization, sensor fusion), but not for our case, where consultation time has a wide span, and some “modes” repeat too often. Particle Mass filter (PM) in contrast to PF, selects deterministically the particles and keeps them fixed. Particles are selected to conform to a grid in the state space. The density of the grid (the number of particles) determines the posterior distribution quality estimation, e.g. posterior distribution tails’ are well represented, but some “modes” can still disappear if they are too narrow for the grid. In stochastic programming, both approaches have been used, and some hybrids have attracted attention (Ballantyne et al., 2003)). Many hybrid filters consider combining deterministic grid particles with stochastic particle ones in a hybrid PMF-PF filter, emphasizing prediction and measurement updates as similar characteristics. Alternatively, we generate a new particle when it is needed, at specific discreet values. It is also important to note that most industrial models are suitable to make several updates when in use.

2.2 Mixture Models and Cluster Weighted Modelling (CWM)

Among probabilistic models, mixture models can represent subpopulations within a population, without explicating to which data aggrupation (or their observed samples) a point belongs (Marin, Mengersen, & Robert, 2005; Sun, Deng, & Han, 2012). These models represent the population‘s pdf and they are useful to make estimations. Most paradigms, assume attributes on the sub-populations, as they are called “clusters”, using past observations or simply measuring their influence as “weights”.

CWM is an approach to model the joint probability of data coming from different sources using a mixture. It can be considered as a time-series’ non-recursive approach that relates an input “z” to a consultation time observation “y” in the future (Bröcker, Engster, & Parlitz, 2009; N. A. Gershenfeld, 1999; N. Gershenfeld et al., 1999; Schoner & Gershenfeld, 2001). CWM needs “unsupervised training” before use with the data set and generates the output using only points lying in the neighbourhood of the interesting point. The search for good weighting values and a small model size helps to reduce approximation error (with low overfitting). Following Gershenfeld, it is also possible to do “functional approximation” with high quantity of data and optimization in an iterative stratified training for the same purpose. In addition, it is also feasible to approximate pdf around “local models” of the data.

As mixtures models are very flexible, we will use the CWM setup to explain the derived model as a useful concept, but not as specific tool. Let us consider the available data as a set of “k” pairs of points $T = \{(z_1, y_1), (z_2, y_2), \dots, (z_k, y_k)\}$, with scalar inputs “$z_i$” and scalar observed outputs “y_i” of the unknown system (the clinic process). When “z” and “y” are drawn from a joint probability $p(z, y)$, the (z_i, y_i) tupples are their realizations and we can try to find an approximation for “$E(\hat{y}|\hat{z})$”. This distribution will sub-divided over clusters “c_m” as,

$$p(y, z) = \sum_{m=1}^{M} p(y, z, c_m),$$

$$p(y, z) = \sum_{m=1}^{M} p(y|z, c_m)\, p(z|c_m)\, p(c_m),$$

$$p(y, z) = \sum_{m=1}^{M} w_m\, p(y|z, c_m)\, p(z|c_m),$$

with $w_m = p(c_m)$.

The neighbourhood of influence of the cluster c_m, for an input vector “z”, is given by $p(z|c_m)$. Some clusters have null (or very small) participation related to a given vector “z_i”. The influence of this cluster in the output for others vectors “$z_j \neq z_i$” is given by parameter “w_m”. Some fdp definitions considers multivariate separable Gaussians but also functionals. In general, Gaussians assist in the interpolation of points and functionals help with linear expressions inside the cluster, $y = Az, A \in R^{n_1 \times n_2}$. Estimations of the joint density $p(\hat{y}, \hat{z})$, by means of the Bayes formulation, also estimates $p(\hat{y}|\hat{z})$. Variants of CWM can arbitrarily split the data to model the time-series relationship among past observations and current observations using as using a “Tree-based Cluster Weighted Model” (Boyden III, 1997). This algorithm reduces the interactions among several points and clusters, determining a fixed number of clusters, assigning the observations to the most probable. Considering arbitrarily predefined clusters, and small data sets by cluster, along with the understanding of the problem’s structure, it is possible to avoid “unsupervised classification” which is a very expensive computationally.

3 Mathematical model

3.1 *Introduction*

Our objective is to maintain a pdf that models the random variable which represents consultation time *q(*)* for an ordinary day, a sample from this *q(*)*. The first step consists in implementing a PF that models consultation time making “corrections” to a sequentially updated filtered distribution $q(x_t|x_{t-1}, y_{1:t})$. Usually it is considered that $q(x_t|x_{t-1}, y_{1:t})$ can be simplified or restrained to $f(x_t|x_{t-1})$ for many PF applications. However, in our case, the “state-0” ($x_t = 0$) is much too frequent in sequence, and is a consequence of particular dynamics: system errors, human behaviour or queues with multiple doctors. The effect of these errors is an unbalance for the conditional probability for this state with respect to the “others” continuous states. In practice, it is very difficult to filter misleading observations for “state-0”, because it is also “tied” with other states (in the longer tail).

As the transition $x_{t-1} = 0 \rightarrow x_t = 0$ is too frequent, it will also force frequent regeneration steps for a PF. To avoid re-generation, transitions are treated in two ways: using the common PF algorithm but moderating the influence using an anti-saturation step, and outside the common operation using a simplified Finite Mixture algorithm loosely based on the CWM model to enhance/reduce the transition probability for these observations. These “approximations or corrections” are not limited to being Markovian and they are external to the PF algorithm. We want irrelevant transitions (patterns) not to condition the future forecast for consultation time, as a correction should not introduce an important “deformation” in the posteriori distribution. In the presence of patterns (like very long service times followed by several “state-0”, or mean-long-mean sequences when doctors pause), the corrections must virtually re-shape the posterior density function to consider the most probable sequence of observations, but “state estimations” must remain untouched by these patterns (or receive a very small correction). Because of the small corrections, the collapse should not happen and regeneration steps become unnecessary. For interpreting sequences, we will use mixtures, with predefined clusters, and use small data sets in them to avoid training, which consumes computational resources. The algorithm uses the long observations history to generate a cluster-based model as “frequency histograms”. A few works reference the use of mixture to improve computational efficiency in Markovian chains and sequential Monte-Carlo (Niemi & West, 2009)

3.2 *Mathematical Problem definition*

Given a series of observations, $O = \{y_1, \dots, y_k\}$, $y_i \in R$. We would like to forecast the next steps, $\{y_{k+1}, \dots, y_{k+s}\}, s \in N, s > 1$, with $s \in \{3,5,8\}$. The $O = \{y_1, \dots, y_k\}$, values are the indirect observations of patient consultations. It is not possible to have precise knowledge of the underlying system. This does not allow us to easily forecast the “s” steps as we do not have the initial conditions and observations are corrupted by noise or by other imprecisions such as the number of doctors in the queue, etc.. We could try to model this system using time series but we would need to make approximations with human supervision.

I

We can consider a subset of the past observations $\{y_{i-1}, \dots, y_{i-d}\}, d \in N$, as an input vector for the forecast model. It is possible to do a forecasting for y_{k+s} but an iterative forecasting in single steps for y_{k+s} is desirable as deviations (or errors) could emerge in the model output; thus an iterated forecasting can add additional observations (corrections) when they are available. Nevertheless, notification given in advance to patients will keep these deviations. For forecasting, Finite Mixture model complements PF model considering $= ([y_{k-d}], \dots, [y_{k-1}])$. Of course, we must subdivide historical data in "days" and "consultations-per-day" to keep cross-consultation timing in the mixture.

3.3 *Algorithm*

First, a PF is set up. As for the PF, we want to avoid excessive attention to "state-0" in the distribution. This helps to initially fit canonical distributions like Pareto, Weibull, Lognormal, Gamma and Negative Exponential, among other distributions that easily fit as a priori distribution and which are also easy to sample. Particles at "state-0" can be added to "match" the experimental distribution. The excessive number of "long service time" values (long tails) is "discarded" in the fitting process too. This behaviour is desirable to simplify the generation of particles to those with high support. This approach is different from the PF literature for non-linear models. **Figure 1,** presents an example from real data. If new observations are longer than particle filter support, then this particle is created.

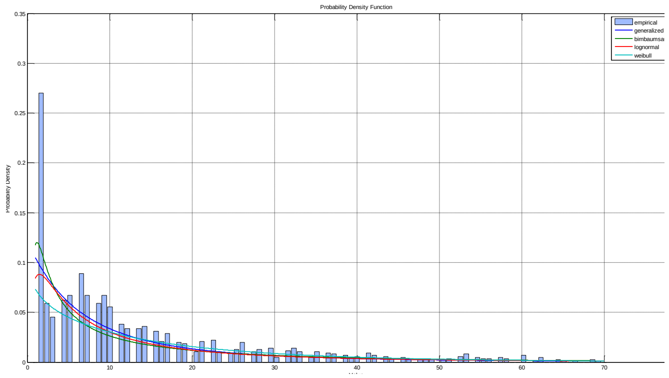


Figure 1. Histogram "Consultation service time". Coloured lines indicate distribution fits. Excessive "state-0" representation is visible.

Then, given historical data observations $y_i, I \in N$ from the clinical consultation process, a simplistic way to model the distribution is by "histogramming y_i". In practice the interval is restricted to [0, 70] minutes. We suppose that the distribution "disappears" outside the boundaries. We would like to generalize the behaviour for very long and very short medical consultations However, it is still needed a function to draw smoothed samples (or approximations) from intervals without observations. An example of histogram for "y_{i+1}vs y_i", is shown in **Figure 2**.

Gaussians kernel density estimators have problems with positive random variables with big probability mass near zero. Usually values close to "zero" ("state-0") can be represented with truncated Gaussians, which is not well covered in CWM literature. Another problem is that Gaussian-overlapping leaves bumps in the results. That is gaining in the local modelling, leaves artifacts in the functional dependence at generalizing. Any non-linear behaviour will appear as an overlapping, because Gaussians capture proximity to hidden states. As traditional CWM considers Gaussians, their direct use is not easy. To control the interpolation among dedicated functions, we prefer to use histogramning versions of $p(y|z, c_m) * p(z|c_m)$. In many cases, the result is very close to a bell-shaped form, or to a Gamma-shaped form as needed. The behaviour remains Gaussian for values close to the mean, but differs on the extremes. In this application of finite mixtures to time series, we restrict observations to a discrete value set. In this case, the joint density distribution is factored on distinct "m" clusters, for each delay as joint distribution, $p([y_i], [y_{i-k}]), k \in N, 1 \leq k \leq d = M$, where "[ ]" is round operator, and the output of each of them is a histogram of the frequency of seen each "state" (without expliciting the dependency). The cluster weights have identical, $p(c_m) = w_m = 1/M$. The resulting weighted histogram is a probability mass function too.

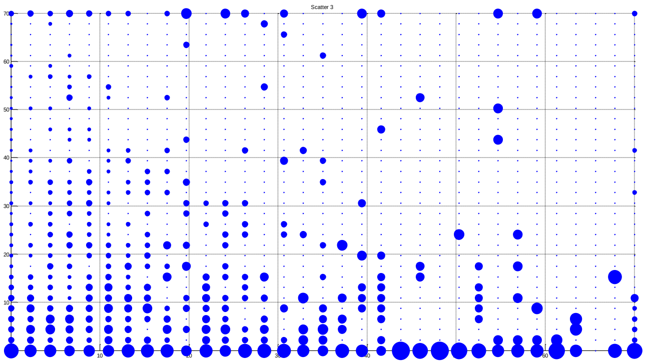


Figure 2. 2D Histogram for y_{i+1} vs y_i. Blue dots represent high relative frequency for discretized observations.

We replace the delta function by "narrow" Gaussian kernels and we use these "histograms" to represent (resume) data of each specialized predefined cluster. These narrow Gaussians act as the likelihood for hidden states. In this way, we can predict quasi-deterministically on (multiple) neighbour observations (more than distributions). From CWM, we also reduce the big number of clusters, the expectation maximization and the slow convergence in name of speed and computational limitations. This version of finite mixture density, close to CWM, is effective to model static patterns (multi-modes, values close to zero, long-tails tied outputs, etc.) with known outputs, as it can replace a huge lookup table for related outputs and other rules. We note that in the implementation, when there is no multiple cluster "agreement" for a future state, it will re-appear as the "mixture density characteristics" of the CWM estimations that loosely approximates a uniform distribution of overlapping Gaussians.

We use the quasi "deterministic" character of the mixture implementation to do importance sampling on the Particle Filter posterior distributions for the clinic's process. We note that the PF operates in a Markovian way, with continuous states that do not coincide with discrete observations in the mixture, but the approximation should suffice. We can do this without altering the PF, using a “copy” of the distribution estimates as the unbiased proposal distribution for important sampling. When there is no particle for the IS to work on, it is necessary to “create” this particle. The finite mixture "important" neighbour observations emphasized (bias) related particles by sampling them more frequently. The variance related to this sequence reduces. This method "encourages" important states but PF updates follow the common procedure as the anti-saturation step works.

We note that with the IS, past consecutive doctor's consultations could affect estimation of future consultations. But when the process contains interleaved patterns that span all the output space and repeat evenly, the finite mixture has captured them, and we could infer that IS distribution would remain close to a flat distribution over and over. The restriction to discrete observations and the effect of doing IS with a bumpy density adds modes and noise to the output, so the model's variance is still bigger than ideal, but very useful for a few samples (very few a day) without degeneration. The deterministic forecasting model delays PF “degeneration” because consecutive “bumps” remain in the mixture (when they have appeared in the pass). The finite mixture is re-initialized once a day to include new data. Many calculations can be done in advance for the finite mixtures but not for the PF calculations. Another way to implement this algorithm is by using Bayesian Networks, but the problem's structure, the noisy historical data, the dimensionality and the number of states is critical to fast adaptation with low computational power machines or thin clients. The simplified algorithm is presented as **Algorithm 1**.

4 CASE STUDY

ChronoMetriq has been offering their services to multiple clinics in the province of Quebec. They attempt to optimize access to health services without changing clinical processes. They offer paid automatic follow-up service, sending notifications to patients (text message, phone call, web page). These notifications are scheduled when there are 8, 5 and 3 patients remaining in queue (8-P, 5-P, 3-P, respectively) before getting their turn. All patients are responsible for attending their consultation on time. Patients are encouraged to be in the consultation room at the last notification (3-P). The service includes a small risk to patients that decide to wait elsewhere than the waiting room. The company developed their system for queues with 1 doctor, but the system is flexible enough to operate with more than 1 doctor. Consultation time is very sensitive information but the company estimates it is very improvable for short and very long consultations. A mean time bigger than 10 [min] is reasonable for a single doctor queue.

The code was implemented in Matlab V. 7.12 (R2011a) in a machine with OS MS Windows 7, I7, 8 GB). For this development, the density estimation and the sampling from known fdp, uses special libraries from Matlab. The rest of the code does not rely on them and taking this into consideration, it will be easy to implement for exploitation. The running time for the test runs is less than 30 sec. The number of particles is set up to be bigger than 400 in all the examples. Parameters were fixed in advance with several tests.

Considering a single working day, we present some examples for timing notifications based on three patients in advance to forecast the consultation time. All the data is real (it was not modified), and comes from walk-in clinics in operation during the last semester of 2014. For these results, we have considered only 40 consultations for standardization. In all the examples, the number (ratio) of “regular” and “private” patients, is unknown, but the company estimates that the number of private patients is considerably fewer than 50% of the total. As the clinical process continues uninterrupted, in this condition, the simulation presented do not consider delays cascading on following consultations. We did not consider no-shows, as patients that receive notifications have a high probability to honour their consultations.

```
Algoritm 1: Algorithm Particle Filter & Finite Mixture

Step-1:   Choose a priori distribution (proposal):

              $p(x_k|x_{1:k-1}, y_{1:k-1},) = p_x(*)$ ,
          A fitted distribution from historical filtered
          data (e.g. LogNormal, Gamma, Pareto, etc.)

Step-2:   Generate N particles as:
          $x_i \sim p_x(*), i = 1..N$
          Set PF weights, $w^i_{k=0} = {}^1/_N$
          $p(y_i|x_i)$ Estimation (Gaussian kernel)
Step-3: Loop :
          If $y_k > x_i$ , then create new particle with
              $x_k = [y_i]$
              $w_k^{N+1} = \min(w_k^i)$ and normalization

          Measurement Update
              $w^i_{k+1} = \frac{w^i_k * p(y_k|x_k)}{\sum_{i=1}^{N} w^i_k * p(y_k|x_k)}$
              Anti-saturation, $w^i_{k+1} \in [w_{\min}, w_{\max}]$
              and normalization

          Finite mixture particle weights estimation:
              $M_i(*)$ represents the finite clustered
              mixture weight estimation
              $\widetilde{w^i_{k+1}} = M_i(([y_{k-d}], \dots, [y_{k-1}]) \times [y_k])$
          If there is not enough particle support, then
          fail to the finite mixture forecast value and
          create a particle.
              $x_k = [y_i]$
              $w_k^{N+1} = \min(w_k^i)$ and normalization

          Sample from,
              $\tilde{x} \sim p(x) = \sum_{i=1}^{N} \widetilde{w^i_{k+1}} \, w^i_{k+1} \delta(x^i_k)$
          to simulate incoming consultations and
          forecast aggregated waiting time.
End.
```

4.1 Example

In this example, where we have a mean consultation time in the interval [5.0, 9.0] min, we can infer that there is more than one doctor working, but the precise information is unknown. **Table 1**, presents the results of using the algorithm on this real data.

I

The first 3 consultations do not have a good forecast, as they are very variable. As is possible to see from the **Table 1.**, "**Mean Error**" and "**St. Dev"** reduce when more information is available. From previous test with other time series, parameters were chosen to have a positive error equivalent to a "Consultation Time" and deviation equivalent to another consultation for "3-P" in advance

First consultation variability has a big impact on the forecast. "Consultation 34", is very short, followed by an excessively long "Consultation 35", very disruptive in the sequence. The accumulated error remains for 3 more consultations.

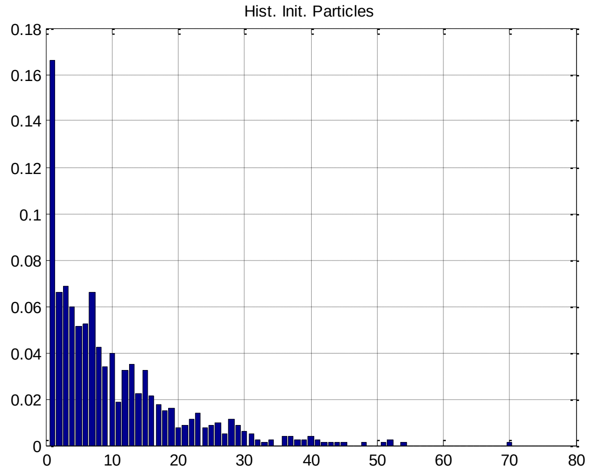
Table 1: "Consultation Time" Forecast Error

3-P: 3 Patients in advance, 5-P: 5 Patients in advance, 8-P: 8 Patients in advance.

# Cons.	Init. Time	PF Mean	Cons. Time	3-P Error	6-P Error	8-P Error
1	0	9,0	12			
2	12	9,4	2			
3	14	7,4	5			
4	19	6,9	2	-7		
5	21	5,6	9	-20		
6	30	6,0	3	1		
7	33	5,3	15	-6	-24	
8	48	6,2	5	13	-24	
9	53	5,9	19	4	8	-22
10	72	6,6	20	26	9	-21
11	92	7,2	0	22	42	31
12	92	5,4	7	23	24	11
13	99	5,7	10	2	33	39
14	109	6,2	13	4	14	25
15	122	7,0	0	9	36	47
16	122	5,3	16	9	0	11
17	138	6,2	1	6	13	42
18	139	4,7	4	1	4	2
19	143	4,5	10	0	11	6
20	153	5,2	0	-2	-4	3
21	153	4,0	11	-4	-5	11
22	164	4,9	17	5	-2	-7
23	181	5,7	7	8	5	9
24	188	5,9	5	23	17	11
25	193	5,7	12	8	21	5
26	205	6,5	9	9	10	16
27	214	6,8	7	6	32	33
28	221	6,8	19	13	15	12
29	240	7,4	4	11	25	50
30	244	6,6	1	15	14	21
31	245	5,0	5	1	19	20
32	250	5,0	9	-7	-8	2
33	259	5,5	4	-9	10	17
34	263	5,2	0	5	-9	-13
35	263	4,0	41	-8	-14	-1
36	304	4,3	4	32	12	15
37	308	4,2	9	24	31	18
38	317	4,9	25	45	25	9
39	342	5,6	0	21	48	52
40	342	4,3	7	20	38	32
Mean		**5,9**	**8,7**	**8,2**	**12,5**	**15,2**
St. Dev.		**0,9**	**5,9**	**10,0**	**13,7**	**14,3**

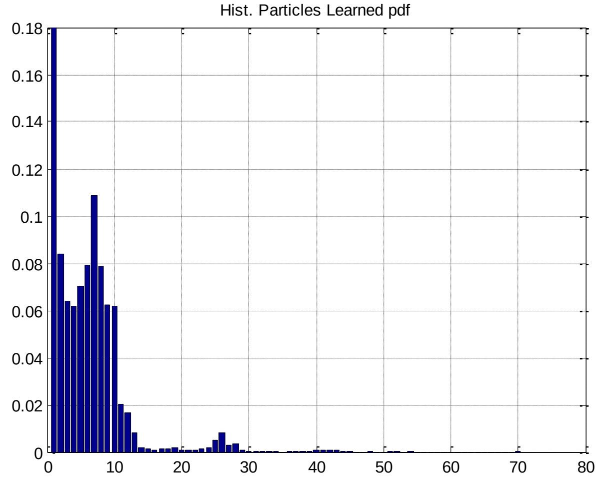
In column "PF Mean", it is possible to appreciate how the PF adapts to new information. As can be seen, the "PF Mean" is lower than the consultation time, as it tries to follow many short consultations in sequence, and "Std. Dev." is low. This is a sign of slow ongoing degeneration (particle collapses for long consultations), as can be seen in **Figure 3**, compared to **Figure 4**. The combination with the mixture seems to be working, as was conceived. Comparing **Figure 4** and **Figure 5** shows that there are too few particles in "State 41".

Figure 3: Histogram for Initial Particles



Note: Particles for "State 0", follow Clinics historic data.

Figure 4: Histogram for Last Particles

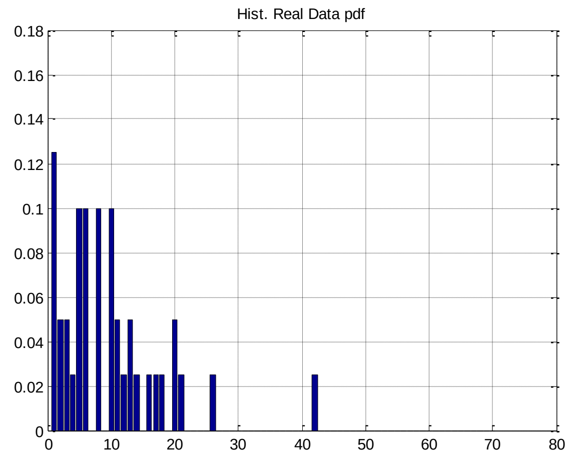


Particles in the range [20, 30], are well represented, longest consultations do not.

4.1 Analysis for the Industry

Using the last available information is usually most important for forecasting. As the number of doctors can be more than one in a queue, a forecast based on "3-P" and "5-P" in advance could seem unpractical. Considering the data presented in the paper, the advice for patients: "To be in the waiting room for "3-P" in advance (the last notification)" seems good.

Figure 5: Histogram for Consultation's Time.



It is known (informally) that patients can save almost 3.5 hours of their time with this system, even considering this advice. So sending three forecasts in advance is reasonable as people can adjust to the possibility that the queue could advance earlier than thought (but rarely the inverse). As we show in the tables, considering the second forecast to be "6-patients" in advance, could be a better tradeoff for queues with reduced "consultation time", as more than one doctor could be present working. First consultation's big errors are difficult to manage, as the PF (mainly) attempt to follow the pace of the queue. This behaviour seems acceptable, as first consultation happens in the first opening hour, and people are reluctant to subscribe for the paid service.

We have already made suggestions to the company about increasing the precision for recording time from minutes to seconds during the study. In addition, we advise considering the re-schedule for the second forecast to "6-patients" in advance (as in the table). We also recommend allowing the system to record delayed patients when possible, as we do not know how much time in advance/late patients arrive at the waiting room, or how they deal with the staff if any other problem arises. At present, ChronoMetriq is considering ameliorating the system to do "time forecast for consultation", still based on the number of patients in advance. Other improvements would consider more training to clinic staff and ways to know the number of doctors in the queues, their break times, and considering other means to measure consultation time.

5 Conclusion

We have developed a practical forecasting tool for consultation service time estimation and patients' waiting time estimation. The tool is intended to have a practical implementation after the analysis from the company. The tool is intended to be used without supervision, as the system can auto-adapt easily. We have also proposed some improvements to the system to facilitate forecasting. The tool is based on a simplified Particle Filter and Finite Cluster Mixtures working collaboratively. The proposed algorithm avoids PF regeneration and deterministically intensifies to forecast states based on sequences of past observations. The algorithm also introduces the generation of particles for long consultation times without support (if needed). Future research can include a complete CWM approach and hybridization based on the discrete states in the mixture. One axe for research is to include new information sources, and that PF work considering the number of doctors in the queues, as this is more challenging for the finite mixture part of the forecasting. Another axe for research is to consider several clinics working in a network. These clinics will use a centralized web service for the inscription of patients. The service will suggest clinics based on availability (using current forecasts) and the time to travel from patients' homes.

6 Acknowledgements

Funding for this research has been given by CRSNG/NSERC (Conseil de Recherches en Sciences Naturelles et en Génie du Canada) and Chronometriq Inc. We are also immensely grateful to Pierre L'Ecuyer for his valuable comments and suggestion for this project.

7 References

Ballantyne, D. J., Hailes, J., Kouritizin, M. A., Long, H., Wiersma, J. H., Ballatyne, D. J., … Wiersma, J. H. (2003). Hybrid weighted interacting particle filter for multitarget tracking. In I. Kadar (Ed.), *Proceedings of SPIE (Signal Processing, Sensor Fusion, and Target Recognition)* (Vol. 5096, pp. 244–255). International Society for Optics and Photonics. doi:10.1117/12.488522

Barron, W. M. (1980). Failed appointments. Who misses them, why they are missed, and what can be done. *Primary Care*, *7*(4), 563–74. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7010402

Bastani, P. (2009). *A queueing model of hospital congestion*. Retrieved from http://summit.sfu.ca/item/9568

Boyden III, E. S. (1997). *Tree-based Cluster Weighted Modeling: Towards A Massively Parallel Real-Time Digital Stradivarius*. *Citeseer*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.9725&rep=rep1&type=pdf

Bröcker, J., Engster, D., & Parlitz, U. (2009). Probabilistic evaluation of time series models: A comparison of several approaches. *Chaos*, *19*(4), 1–14. doi:10.1063/1.3271343

Cappe, O., Godsill, S. J., & Moulines, E. (2007). An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo. *Proceedings of the IEEE*, *95*(5), 899–924. doi:10.1109/JPROC.2007.893250

Carpenter, J., Clifford, P., & Fearnhead, P. (1999). An improved particle filter for non-linear problems. *IEE Proceedings Radar Sonar and Navigation*, *146*(1), 2. doi:10.1049/ip-rsn:19990255

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, *12*(4), 519–549. doi:10.1111/j.1937-5956.2003.tb00218.x

Channouf, N., L'Ecuyer, P., Ingolfsson, A., & Avramidis, A. N. (2007). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science*, *10*(1), 25–45. doi:10.1007/s10729-006-9006-3

I

Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics*, *182*(1), 1–69. Retrieved from http://www2.ee.kuas.edu.tw/~lwang/WWW/BayesianFilteringFromKalmanFiltersToParticleFiltersAndBeyond.pdf

Côté, M. J., & Stein, W. E. (2007). A stochastic model for a visit to the doctor’s office. *Mathematical and Computer Modelling*, *45*(3), 309–323. doi:10.1016/j.mcm.2006.03.022

Gershenfeld, N. A. (1999). *The Nature of Mathematical Modeling*. Cambridge University Press. Retrieved from https://books.google.ca/books?id=lSTOh8U7NkkC

Gershenfeld, N., Schoner, B., & Metois, E. (1999). Cluster-weighted modelling for time-series analysis. *Nature*, *397*(6717), 329–332. doi:10.1038/16873

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation.

Green, L. V., & Savin, S. (2008). Reducing Delays for Medical Appointments: A Queueing Approach. *Operations Research*, *56*(6), 1526–1538. doi:10.1287/opre.1080.0575

Gustafsson, F. (2010). Particle filter theory and practice with positioning applications. *Aerospace and Electronic Systems Magazine, IEEE*, *25*(7), 53–82. doi:10.1109/MAES.2010.5546308

Huang, Y., & Hanauer, D. A. (2014). Patient No-Show Predictive Model Development using Multiple Data Sources for an Effective Overbooking Approach. *Applied Clinical Informatics*, *5*(3), 836–860. doi:10.4338/ACI-2014-04-RA-0026

Hulshof, P. J. H. (2013). *Integrated decision making in healthcare: an operations research and management science perspective*. Universiteit Twente. Retrieved from http://books.google.com/books?id=OSW7oAEACAAJ&pgis=1

Jerbi, B., & Kamoun, H. (2011). Multiobjective study to implement outpatient appointment system at Hedi Chaker Hospital. *Simulation Modelling Practice and Theory*, *19*(5), 1363–1370. doi:10.1016/j.simpat.2011.02.003

Klassen, K. K. J., & Yoogalingam, R. (2009). Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management*, *18*(4), 447–458. doi:10.1111/j.1937-5956.2009.01021.x

Li, H. (2014). A Brief Tutorial On Recursive Estimation With Examples From Intelligent Vehicle Applications (Part IV): Sampling Based Methods And The Particle Filter. *Hal.archives-Ouvertes.fr*, (Part IV). Retrieved from https://hal.archives-ouvertes.fr/hal-01054713/

Marin, J.-M., Mengersen, K. L., & Robert, C. P. (2005). Bayesian Modelling and Inference on Mixtures of Distributions. In D. D.K. & C. R. Rao (Eds.), *Handbook of Statistics* (Vol. 25, pp. 459–507). Elsevier. doi:10.1016/S0169-7161(05)25016-2

Mukherjee, A., & Sengupta, A. (2010). Likelihood function modeling of particle filter in presence of non-stationary non-gaussian measurement noise. *Signal Processing*, *90*(6), 1873–1885.

Niemi, J. B., & West, M. (2009). *Bayesian Analysis and Computational Methods for Dynamic Modeling*.

Patrick, J., Puterman, M. L., & Queyranne, M. (2008). Dynamic Multipriority Patient Scheduling for a Diagnostic Resource. *Operations Research*, *56*(6), 1507–1525. doi:10.1287/opre.1080.0590

Qu, X., Rardin, R. L., Williams, J. A. S., & Willis, D. R. (2007). Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, *183*(2), 812–826. doi:10.1016/j.ejor.2006.10.003

Qu, X., & Shi, J. (2011). Modeling the effect of patient choice on the performance of open access scheduling. *International Journal of Production Economics*, *129*(2), 314–327. doi:10.1016/j.ijpe.2010.11.006

Ristic, B., Arulampalam, S., & Gordon, N. (2004). Beyond the Kalman filter. *IEEE AEROSPACE AND …*, *19*(7), 37–38. doi:10.1109/MAES.2004.1346848

Robinson, L. W., & Chen, R. R. (2010). A Comparison of Traditional and Open-Access Policies for Appointment Scheduling. *Manufacturing & Service Operations Management*, *12*(2), 330–346. doi:10.1287/msom.1090.0270

Rosenbluth, M. N., & Rosenbluth, A. W. (1955). Monte Carlo Calculation of the Average Extension of Molecular Chains. *The Journal of Chemical Physics*, *23*(2), 356. doi:10.1063/1.1741967

Salzarulo, P. a., Bretthauer, K. M., Côté, M. J., & Schultz, K. L. (2011). The Impact of Variability and Patient Information on Health Care System Performance. *Production and Operations Management*, *20*(6), 848–859. doi:10.1111/j.1937-5956.2010.01210.x

Sampson, R., O’Rourke, J., Hendry, R., Heaney, D., Holden, S., Thain, A., & MacVicar, R. (2013). Sharing control of appointment length with patients in general practice: a qualitative study. *The British Journal of General Practice : The Journal of the Royal College of General Practitioners*, *63*(608), e185–91. doi:10.3399/bjgp13X664234

Schoner, B., & Gershenfeld, N. (2001). Cluster-Weighted Modeling: Probabilistic Time Series Prediction Characterization and Synthesis. In *Nonlinear Dynamics and Statistics* (pp. 365–385). doi:10.1007/978-1-4612-0177-9_15

Shi, Y., & Han, C. (2007). The divided difference particle filter. *FUSION 2007 - 2007 10th International Conference on Information Fusion*. doi:10.1109/ICIF.2007.4408063

Silverman, J., Kinnersley, P., Round, T., Irving, G., Holden, J., & Hartshorn, C. (2012). Calling time on the 10-minute consultation. *The British Journal of General Practice : The Journal of the Royal College of General Practitioners*, *62*(596), 118–119. doi:10.3399/bjgp12X625102

Sun, Y., Deng, H., & Han, J. (2012). Probabilistic models for text mining. In *Mining Text Data* (pp. 259–295). Springer.

Swisher, J. R., Jacobson, S. H., Jun, J. B., & Balci, O. (2001). Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & Operations Research*, *28*(2), 105–125. doi:10.1016/S0305-0548(99)00093-3

Tai, G., & Williams, P. (2012). Optimization of scheduling patient appointments in clinics using a novel modelling technique of patient arrival. *Computer Methods and Programs in Biomedicine*, *108*(2), 467–76. doi:10.1016/j.cmpb.2011.02.010

I

Zacharias, C., & Pinedo, M. (2014). Appointment Scheduling with No-Shows and Overbooking. *Production and Operations Management*, *23*(5), 788–801. doi:10.1111/poms.12065

Zonderland, M. E. (2012, January 27). *Curing the queue*. University of Twente. Retrieved from http://books.google.com/books?id=IzI7MwEACAAJ&pgis=1