Using Data Assimilation Technique and Epidemic Model to Predict TB Epidemic

Himanshu Gupta
Deptt of CSE
Uttranchal University

Kamal Kant Verma
Deptt of CSE,
Quantum School of Technology

Punit Sharma
Deptt of CSE,
Uttranchal University

ABSTRACT

People of India are very susceptible to many infectious diseases like malaria, TB, HIV etc. There are many epidemic models that are used to predict new cases of disease. Some of the popular epidemic models are SI (Susceptible-Infectious), SIR (Susceptible-Infectious-Recovered), SIRS, SIS etc.

In this research quarterly data of TB disease in Uttarakhand (India) for 7 years is collected and on the basis of this data new infected population in the next quarter is predicted using SIR epidemic model and data assimilation technique (Ensemble Kalman Filter). Analysis and implementation is done in MATLAB. Results show good agreement to measured values.

Keywords

Epidemics, Infectious Disease, Disease Dynamics, spatial-temporal SIR model & equations, Data Assimilation, Ensemble Kalman Filter, Matlab, Kalman gain Matrix

1. INTRODUCTION

People of India are very susceptible to many infectious diseases like HIV, malaria, TB etc. Various mathematical models have been formulated to study disease dynamics to describe the spreading of disease in a population. This branch of science is popularly known as Epidemiology, and the models used to study disease dynamics are known as epidemiological models. Rise and fall of an epidemics depends upon the consideration of various parameters such as: biological, disease parameter (vaccination, infection parameters, population immigration, birth and death rates, gain of permanent, temporary or no immunity, latent period, dispersal, quarantine, treatment etc. A number of epidemiological model such as SI (Susceptible-Infectious), SIR (Susceptible-Infectious-Recovered), SIRS, SIS etc. are available to study and forecast the spread of a disease and predict an epidemic.[1-6]

This paper is divided into five sections. Section 2 describe about the data assimilation method used in the implementation ie. Ensemble Kalman filter. The section 3 focus on the study of epidemic model while mainly concentrated on SIR model. The fourth section is a discussion about the tool and the fifth section gives the Result and Analysis.

1.1 Kalman filter as data assimilation technique

We use a data assimilation method for statistical tracking of epidemics caused by infectious disease. This involves two basic components: a dynamic model to forecast the state of the epidemic between arrivals of new data, and observations that are used to update an ensemble of state estimates. Data assimilation requires estimating the uncertainty both for model and observations forecasts. The goal in this paper is to incorporate sparse and noisy observational epidemic data over space and time (for Uttrakhand, India) into a dynamic statistical model so as to produce an estimate of the current state of the infected population, and to forecast the progress of the real epidemic. The data for simulation is taken from the website tbcindia.nic.in. There are a number of variants of kalman filter: Basic Kalman filter, Extended Kalman filter and Ensemble kalman filter. In this paper, Ensemble Kalman filter is used for implementation.

1.2 Ensemble Kalman Filter (EnKF) as the Data Assimilation Method.

The Ensemble Kalman Filter (EnKF) belongs to a broader category of filters known as particle filters [7, 8]. Unlike Extended Kalman Filter (XKF) estimation and SDRE estimation, particle filters use neither the Jacobian of the dynamics nor frozen linear dynamics. The starting point for particle filters is choosing a set of sample points, that is, an ensemble of state estimates that captures the initial probability distribution of the state. These sample points are then propagated through the true nonlinear system and the probability density function of the actual state is approximated by the ensemble of the estimates. A brief overview of the technique is given in [9-12]. In the weather prediction literature, there exist a large numbers of papers that employ EnKF [11, 12].

The Ensemble Kalman filter (EnKF) was introduced by Evensen. The Kalman filter formula operates directly on the mean and covariance of the model state to produce the exact filtering distribution. For completeness, major points in the development of the Kalman filter are derived here [8] [13].

The EnKF algorithm can tackle the initial state uncertainties in the model. In the following, we account for this state-dependent uncertainty by taking an ensemble approach to data assimilation. The EnKF is a popular sequential Bayesian data assimilation technique that uses a collection of almost-independent simulations (known as an ensemble) to solve the covariance problem of Kalman filtering for systems with very high-dimensional state vectors. It does this using a mainly two-step process: estimate of the covariance matrix, followed by an ensemble update. The covariance of a single state estimate in the KF is replaced by the sample covariance computed from the ensemble members. This sample covariance of ensemble forecasts is then used to calculate the Kalman gain matrix.

The ensemble Kalman filter (EnKF) is a suboptimal estimator, where the error statistics are predicted by using a Monte Carlo or ensemble integration to solve the Fokker-Planck equation. The Ensemble Kalman Filtering method is presented in three stages

1. Error statistics in fore cast step

At time k, we have an ensemble of q forecasted state estimates with random sample errors. We denote this ensemble as $X^f_k \in R^{n\times q}$, where

$X^f_k = (x^{f1}_k,...,x^{fq}_k)$, (1)

and the superscript f_i refers to the i-th forecast ensemble member. Then, the ensemble mean $\overline{\overline{X}}_k{}^f \in R^n$ is defined by

$\overline{\overline{X}}_k{}^f=(1/q)\ \Sigma\ x^{fi}_k$.

Since the true state x_k is not known, we approximate $P^f_{xyk,}$ $P^f_{xyk,}$ by using the ensemble members.

We define the ensemble error matrix $E^f_k \in R^{n\times q}$ around the ensemble mean by

$E^f_k = [\ x^{f1}_k - \overline{\overline{X}}_k{}^f \ldots\ldots\ldots\ldots x^{fq}_k - \overline{\overline{X}}_k{}^f]$……….(2)

and the ensemble of output error $E^a_{ky} \in R^{p\times q}$

by $E^a_{ky} = [\ y^{f1}_k - \overset{=}{y}_k{}^f \ldots\ldots\ldots\ldots y^{fq}_k - \overset{=}{y}_k{}^f]$……….(3)

We then approximate P^f_k by $\hat{P}^f_k$, P^f_{xyk} by $\hat{P}^f_{xyk}$, and P^f_{yyk} by $\hat{P}^f_{yyk}$, respectively,

where $\hat{P}^f_k=(1/q-1)\ E^f_k\ (E^f_k)^T$

$\hat{P}^f_{kxy}=(1/q-1)\ E^f_k\ (E^f_{ky})^T$

$\hat{P}^f_{kyy}=(1/q-1)\ E^f_{ky}\ (E^f_{ky})^T$……………….(4)

Thus, we interpret the forecast ensemble mean as the best forecast estimate of the state, and the spread of the ensemble members around the mean as the error between the best estimate and the actual state.

The second step is the analysis step: To obtain the analysis estimates of the state, the EnKF performs an ensemble of parallel data assimilation cycles, where for i =1,...,q

$x^{ai}_k= x^{fi}_k+\hat{K}_k(y^i_k- h\ (x^{fi}_k)\ )$…………… (5)

The perturbed observations Y^i_k are given by

$Y^i_k = y_k + v^i_k$…………….. (6)

where v^i_k is a zero-mean random variable with a normal distribution and covariance R_k. The sample error covariance matrix computed from the v^i_k converges to R_k as q →∞.

We approximate the analysis error covariance P^a_k by $\hat{P}^a_k$, where

$\hat{P}^a_k=(1/q-1)\ E^a_k(E^a_k)^T$

and E^a_k is defined by (2) with x^{fi}_k replaced by x^{ai} and x^f_k replaced by the mean of the analysis estimate ensemble members. We use the classical Kalman filter gain expression and the approximations of the error covariances to determine the filter gain $\hat{K}_k$ by

$\hat{K}_k =\hat{P}^{fxy}_k(\hat{P}^{fyy}_k){}^\wedge(-\ 1)$ …... (7)

The last step is the prediction of error statistics in the forecast step:

$x^{fi}_{k+1} =\ f\ (x^{ai}_k,u_k\ )+ w^i_k$,…………………..(8)

where the values w^i_k are sampled from a normal distribution with average zero and covariance Q_k. The sample error covariance matrix computed from the w^i_k converges to Q_k as q →∞.

So, the main equations for EnKF[13] are:

Forecast to Analysis of that time after getting data Step:

$\overline{\overline{X}}_k{}^f=(1/q)\ \Sigma x^{fi}_k$. [i=1(1)q]

$E^f_k = [\ x^{f1}_k- \overline{\overline{X}}_k{}^f \ldots\ldots\ldots\ldots x^{fq}_k- \overline{\overline{X}}_k{}^f]$

$E^a_{ky} = [\ y^{f1}_k- \overset{=}{y}_k{}^f \ldots\ldots\ldots\ldots y^{fq}_k- \overset{=}{y}_k{}^f]$

$\hat{P}^f_{kxy}=(1/q-1)\ E^f_k\ (E^f_{ky})^T$

$\hat{P}^f_{kyy}=(1/q-1)\ E^f_{ky}\ (E^f_{ky})^T$

$\hat{K}_k =\hat{P}^{fxy}_k\ [(\hat{P}^{fyy}_k){}^\wedge(-\ 1)]$

$x^{ai}_k= x^{fi}_k+\hat{K}_k(y^i_k- h\ (x^{fi}_k)\ )$

Analysis to forcast for next time step:

$x^{fi}_{k+1} =\ f\ (x^{ai}_k,u_k\ )+ w^i_k$

Generally Enkf needs much little time & computational resources as it solves the storage-and-retrieval problem of covariance matrix for high dimensional state space problem by calculating the covariance from the members of the ensemble as they are needed. The result is an elegant Bayesian update algorithm with dramatically improved efficiency and storage requirements.

Though Ensemble Kalman Filter was developed for non linear system, but it can also be used for Linear system because it needs much less calculation & memory & computational resources & still giving a good estimate.Though it may not give the optimal estimate (least mean squared error) like Kalman Filter but it can give about optimal estimate using less computer resources & taking less time & thus more suitable for practical implementation.By increasing the no. of Ensembles we can increase the accuracy though it may take much time. So considering both time & accuracy we should take a suitable no. of ensembles.

1.3 Epidemic models

There are many epidemic models used in prediction of a disease spread according to geographic and physical conditions. In this research SIR epidemic model is used for Analysis and prediction of disease spread. SIR is a compartmental model initially studied in depth by Kermack and McKendrick [7]. In SIR model we divide the whole population into three compartments

1) Susceptible 2) Infected 3) Removed

Three variables are used to define the state of the epidemic [8] as

S(x,y,t)=density (per unit area) of the susceptible population

I(x,y,t)=density of the Infected population

R(x,y,t)=density of the Removed population

Thus each of these variables evolves with time. In continuous time the epidemic dynamics are defined by a system of three partial differential equations for the state variables. These equations are given by

1) dS/dt=-Bst
2) dI/dt=Bst-YI

3) dR/dt=YI

where

B=probability of disease transmission

Y=Recovery rate coefficeient

The assumptions taken for this model are:

-closed environment

-No emigration/Immigration

-No birth/death

ie. Constant total population

1.4 Tool

The code is implemented in Matlab. It is choosen for implementation because it is a language that has a wide user base and is familiar to a lot of programmers. Matlab is particularly popular for implementing numerical and scientific applications. It is an array language. Matlab is used as a programming language to write various types of simulations in the areas like image processing, communication and control engineering. Matlab is a high-level language developed by mathworks. It has grown into a diverse and vast language over the years. Matlab programs, at a basic level are similar to programs written in language like c++. Each program uses set of variables that can be manipulated using operators and function calls. Matlab supports variables of several primitive types like int, complex, string, real. In matlab all variables are matrices. Scalrs are single element matrices. In matlab, a variable does't need to be defined to be of particular type its type is determined only at runtime. Matlab provides a rich set of operators to operate on matrices. It supports most of the common control flow structures. It has support for if-else statements and while, do-until, for loops. Matlab has a wide variety of toolsets for domain specific functionality. For example:-image processing toolbox provides API s to several commonly used image processing functions.

2. RESULT ANALYSIS

In this research, data of total no of TB patients registered for treatment quarterly in Uttrakhand is taken and then using SIR epidemic model and ensemble Kalman filter new cases of TB are predicted. it can be observed from the data that there is a seasonal variation in TB infection different quarters in the study represents different seasons of the year the computational model presented in this research is able to capture this seasonal variation of TB infection very well.

The percentage variance/error value varies from 1.56% to 36.66%. It is also observed from the results that the computational model is predicting below the observed value in quarter 1 and quarter 2 and over prediction in quarter 3 and quarter 4 the reason for this discrepancy is that fixed beta is taken for all the quarters while infection rate varies season to season. No significant effect of ensemble size is observed in the study. To improve the accuracy of results the infection rate beta should vary quarter to quarter.

Table 1. Data of TB patients with their observed and predicted values

Sno	**Quarter**	**Observed value**	**Predicted value (EnsembleSize100)**	**Error(Variance in %)**
1	2005Q3	3096	3096	0
2	2005Q4	2378	3155	24.7111
3	2006Q1	2695	2438	-10.5131
4	2006Q2	3253	2753	-18.0506
5	2006Q3	3136	3318	5.3516
6	2006Q4	2569	3197	19.7642
7	2007Q1	3002	2629	-13.9802
8	2007Q2	4190	3065	-36.6671
9	2007Q3	3374	4253	20.6260
10	2007Q4	2840	3437	17.3641
11	2008Q1	3132	2902	-7.8711
12	2008Q2	3806	3193	-19.1024
13	2008Q3	3432	3866	11.2822
14	2008Q4	2961	3496	15.2360
15	2009Q1	3181	3021	-5.2856
16	2009Q2	3960	3242	-22.0942
17	2009Q3	4083	4022	-1.5657
18	2009Q4	3076	4144	25.8155
19	2010Q1	3592	3142	-14.5115
20	2010Q2	4461	3654	-22.0523
21	2010Q3	3566	4527	21.1672
22	2010Q4	3136	3629	13.5728
23	2011Q1	3453	3197	-8.1117
24	2011Q2	4338	3514	-23.4063
25	2011Q3	4014	4397	8.8155
26	2011Q4	3078	4077	24.4567

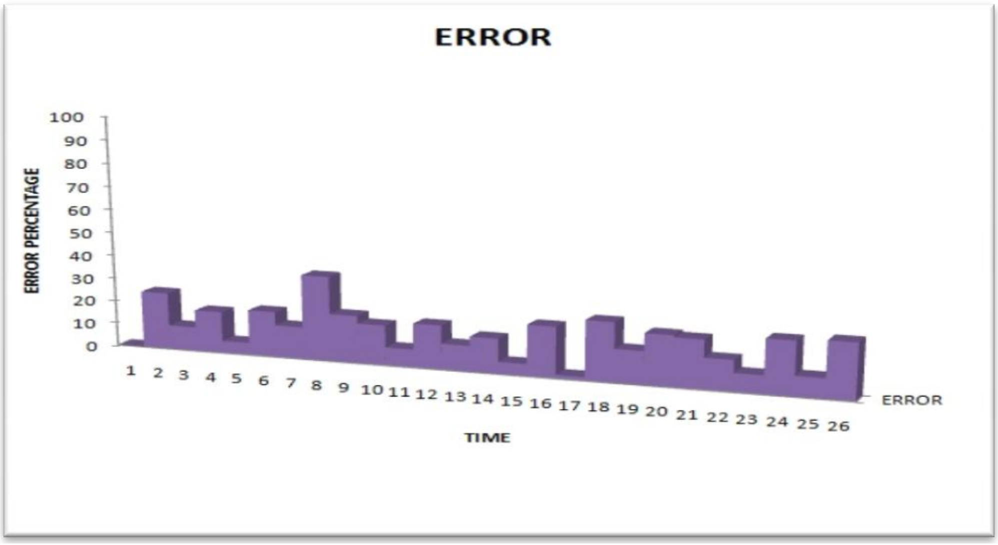

Fig 1: Percentage error/variance between predicted and observed values on each quarter

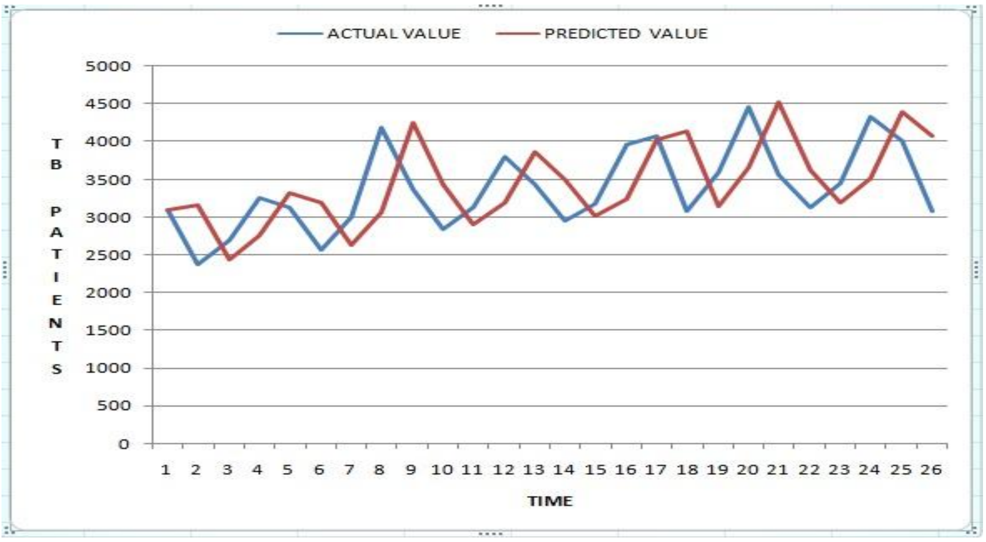

Fig 2: Predicted and observed value of patients on each quarter ensemble size 100

Table 2. Data of TB patients with their observed and predicted values for different Ensemble size

Sno	Actual	50 Ensemble	100 Ensemble	150 Ensemble	200 Ensemble
1	3096	3096	3096	3096	3096
2	2378	3156	3155	3158	3158
3	2695	2440	2438	2441	2438
4	3253	2757	2753	2757	2761
5	3136	3314	3318	3314	3316
6	2569	3197	3197	3198	3198
7	3002	2629	2629	2631	2632
8	4190	3064	3065	3063	3063
9	3374	4251	4253	4250	4251
10	2840	3435	3437	3438	3436
11	3132	2903	2902	2901	2902
12	3806	3191	3193	3197	3193
13	3432	3865	3866	3868	3868
14	2961	3497	3496	3493	3493
15	3181	3022	3021	3024	3020
16	3960	3239	3242	3244	3244
17	4083	4017	4022	4022	4020
18	3076	4146	4144	4145	4148
19	3592	3134	3142	3135	3138

20	4461	3654	3654	3654	3653
21	3566	4523	4527	4524	4523
22	3136	3632	3629	3626	3626
23	3453	3195	3197	3195	3201
24	4338	3517	3514	3515	3518
25	4014	4340	4397	4402	4399
26	3078	4074	4077	4078	4075

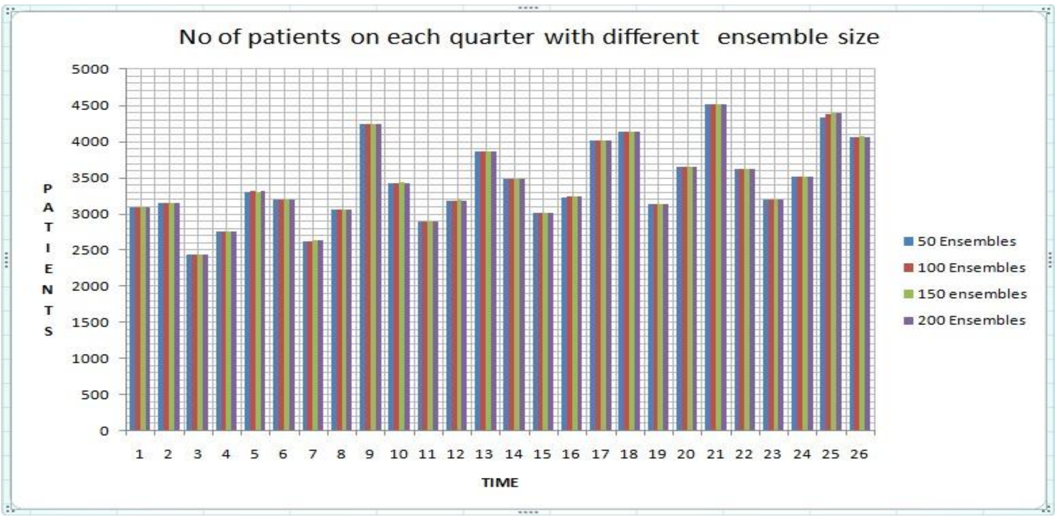

Fig 2: Predicted value of patients on each quarter with different ensemble size

3. CONCLUSION AND FUTURE SCOPE

Quarterly data of people infected from TB disease in uttrakhand (INDIA) is taken and using SIR epidemic model and ensemble Kalman filter new cases of TB are predicted. First Ensemble kalman filter with 100 ensemble size is used for prediction then different no of ensembles is tried. It is found that the predicted values are almost same for all ensemble size. There are some variations in our predicted value and observed values. In every model there are some known and unknown limitations/errors. In future more sophisticated program/model can be developed that take care of those limitations so that the system could predict the exact value.

4. REFERENCES

[1] F. Brauer. “Compartmental models in epidemiology” In Mathematical Epidemiology, volume 1945 of Lecture Notes in Mathematics, pages 19–79. Springer Berlin Heidelberg, 2008.

[2] T. L. Burr and G. Chowell.” Observation and model error effects on parameter estimates in susceptible-infected-recovered epidemiological models”. Far East Journal of Theoretical Statistics, 19(2):163–183, 2013.

[3] W.D. Flanders and D.G. Kleinbaum.” Basic models for disease occurrence in epidemiology”. International Journal of Epidemiology, 24(1):1–7, 1995

[4] L.X. Yang and X. Yang. “A new epidemic model of computer viruses”. Communications in Nonlinear Science and Numerical Simulation, 19(6):1935 – 1944, 2014.

[5] H.W. Hethcote. “The mathematics of infectious diseases””. SIAM Rev., 42(4): 599–653, December 2000.

[6] Lonela Roxana Danilla,”On- the-fly modelling and prediction of Epidemic phenomena”, Imperial College London,june 2014

[7] W. O. Kermack and A. G. McKendrick, “A contribution to the mathematical theory of epidemics”, Proceedings of the Royal Society of London. Series A, 115(772): 700–721, 1927.

[8] Ashok Krishnamurthy, “Bayesian Tracking of Emerging Epidemics Using Ensemble Optimal Statistical Interpolation (EnOSI)”, Section on Statistics in Epidemiology-JSM 2010

[9] F. E. Daum and J. Huang, “The Curse of Dimensionality for Particle Filters,” Proc. IEEE Conf. Aero., vol. 4, pp. 1979-1993, 2003.

[10] J. H. Kotecha and P. M. Djuric, “‘Gaussian particle filtering,” IEEE Trans. Sig. Proc., vol. 51, pp. 2592 - 2601, 2003.

[11] R.Daley, “Atmospheric Data Analysis”, Cambridge University Press, 1991.

[12] E. Kalnay, “Atmospheric modeling, data assimilation and predictability”, Cambridge University Press, 2003.

[13] A.Ridley, “What is the Ensemble Kalman Filter and how well does it perform?”, American control conference, 2006.