PoP: Prediction on Predictions with Ensemble Method

Hideo Hirose, and Yuki Koyanagi

Abstract—In observing the widely spread of patients caused by infectious diseases or the increase of the number of failures of equipments, it is crucial to predict the final number of infected patients or failures at earlier stages. To estimate the number of infected patients, the SIR model, the ordinary differential equation model, statistical truncated model are useful. The predicted value for the final number of patients using data until truncation time \( T \) becomes a function (trend) of \( T \). To grasp the prediction trend with truncation time, the L-plot is developed here, which is to plot the predicted final value at the truncation time. We consider the use of the L-plot to predict the final number of patients. For example, we have shown to use the decay function. Applying the multiple methodologies to the same data, we can expect better predicted values. This is called the PoP, the prediction on predictions. As one of the PoP method, we propose to use the ensemble method. By applying these methodologies to the SARS case, we have found that the ensemble method works well as a PoP method.

Index Terms—PoP, Pandemic, SIR model, ordinary differential equation model, statistical truncated model, generalized logistic distribution, ensemble method, L-plot, decay function, restricted RMSE, pandemic, SIR model, ordinary differential equation model, statistical truncated model, ensemble method.

I. INTRODUCTION

In observing the increase of the number of patients caused by an infectious disease, it is crucial to predict the final number of infected patients. To determine whether the spread could be an outbreak or not is a great concern to everyone because a possible pandemic may affect the huge economical effect as well as the social damages. To estimate the number of infected patients, the SIR model [1], [11], [14], [4], the ordinary differential equation (ODE) model [3], [7], and the statistical truncated model [2], [5], [6], [8], [9] are considered to be useful to estimate the number of infected patients.

The predicted value for the final number of patients using data until time \( T \) becomes a function (trend) of \( T \). We here consider the use of this trend to predict the final number of patients. So far, we have been discussing about the better predictor in the sense that the newly proposed method is superior to other conventional methods. However, in this paper, we try to use all the methods already proposed, and to make a better result than that by using a single method. That is, we will make a prediction using the predicted values already obtained. We call this methodology the PoP, the prediction on predictions.

It seems that the prediction accuracy will not increase by this method because we use the same data. However, we may expect the better predicted values if we apply the multiple methods to the same data [10]. In this paper, we show this by applying the results of the SARS case using the proposed method.

II. PRIMARY PREDICTION METHODS

We have made the predictions for disease spread by using three primary prediction models: 1) the SIR model, 2) the ordinary differential equation model, 3) the statistical truncated model.

A. SIR Model

The SIR model is described by simultaneous ordinary differential equations to perform pandemic simulations [1], [11], [14], [4], where \( S, I, \) and \( R \) are susceptible, infectious, and removed populations, and the parameters \( \lambda \) and \( \gamma \) the infection rate and the removal rate (recovery rate), respectively.

\[
\begin{align*}
S'(t) &= -\lambda S(t) I(t), \\
I'(t) &= \lambda S(t) I(t) - \gamma I(t), \\
R'(t) &= \gamma I(t).
\end{align*}
\]

The parameters \( \lambda \) and \( \gamma \) can be computed by using the best-backward solution method, BBS ([3], [7]), when we estimate the parameters \( \lambda \) and \( \gamma \) using the observed data.

B. Ordinary Differential Equation (ODE) Model

The ordinary differential equation (ODE) model [3] uses the generalized logistic distribution such that

\[
G'(t) = \frac{\beta G(t) \exp(-(t - \mu)/\sigma)}{\sigma + \exp(-(t - \mu)/\sigma)}
\]

where, \( G(t) \) corresponds to the number of infected patients at time \( t \).

\[
G(t; \mu, \sigma, \beta) = \frac{N}{\{1 + \exp(-[(t - \mu)/\sigma])\}^\beta},
\]

Here, \( N \) is the final number of infected patients. The parameters are estimated by using the method of least squares, and the optimization is performed by the simplex method [13].

C. Statistical Truncated Model

Although we use the same probability distribution as shown above, the method is different from that. The log-likelihood function

\[
\log L(\theta) = \sum_{i=1}^{r} n_i \log \left\{ \frac{F(t_{i+1}; \theta) - F(t_{i}; \theta)}{F(t_T; \theta)} \right\},
\]

is used [2], [5], [6], [8], [9], where \( t_T \) denotes the truncation time, \( t_i \) the \( i \)th day from the beginning, and \( n_i \) the number of patients on the \( i \)th day.

Manuscript received March 31; revised May 03, 2014. This work was supported in part by JSPS KAKENHI Grant Number 24310121.

H. Hirose, Y. Koyanagi: Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Kawazu 680-8502 Japan, e-mail: hirose@ces.kyutech.ac.jp, URL: http://hirose.ces.kyutech.ac.jp

(Advance online publication: 27 May 2014)
III. BEST-BACKWARD SOLUTION METHOD, BBS

In estimating the parameters using the observed data, we use the best-backward solution method, BBS ([7]). This is basically a method of least squares, but some extension is included. The procedure for this method is as follows:

1) We obtain initial estimates for parameters using the simple forward/backward difference method.
2) Using these initial values, we solve differential equations (5) from \( t_T \) to 0 backward, where \( t_T \) is the last time of observation. We next compute \( Z_0 \) as shown below,

\[
Z_0 = \sum_{j=1}^{n} (\hat{Y}(t_j) - \tilde{Y}(t_j))^2,
\]

where, \( Y(t_j) = R(t_j) + I(t_j) \) in the SIR model or \( Y(t_j) = NF(t_j) \) in the single distribution model having a cumulative distribution function \( F(t) \); \( \hat{Y}(t_j) \) is the estimated value for \( Y(t_j) \); \( \tilde{Y}(t_j) \) is the estimated value for \( Y(t_j) \). Here, observed data \( \tilde{Y}(t_j) \), \( (j = 1, \ldots, n) \) were assumed to be available, where \( t_T = t_n \). We find parameters so that we minimize \( Z_0 \) using the downhill simplex method by [13] by iterating backward-solution until convergence. We have applied this method to the SIR model and the ODE model.

IV. UNDERESTIMATION ISSUE IN THE TRUNCATED MODEL.

Figure 1 shows a typical case of the observed and predicted number of cumulative patients using the SARS case data in Hong Kong in 2003, where the truncation date is set to the 20th day (April 6, 2003) from the beginning; the prediction method is the statistical truncated model. It seems that the predicted number of cumulative patients is underestimated. Such a tendency is often observed when we use the statistical truncated models.

V. L- PLOT

Figure 3 shows the observed and predicted number of cumulative patients using the SARS case data in Hong Kong in 2003 by the various days of truncation time; the prediction method is the statistical truncated model. It is very difficult to grasp the whole prediction trend in the figure.

The predicted value for the final number of patients using data until time \( T \) becomes a function (trend) of \( T \). To show this final values, we tried to put the prediction points at the truncation days together with the cumulative number of observed values. This is new. We call this trend plot “L-plot” here. The L-plot shows us how early the prediction method predicts the final number of patients; see Figure 4, which demonstrates the SARS case. It would be beneficial if we can consider to use the L-plot in predicting the final number of patients easily.

Figure 5 shows the predicted results by using the SIR model in L-plot form. We can find the clear difference between the two models, the statistical truncated model (Figure 4) and the SIR model (Figure 5). In the SIR model, we can see that the final number of patients are predicted rather in early stages. The stable prediction can be attained from the 20th day from the beginning. On the contrary, the stable prediction may be obtained from the 35th day from the beginning in the truncated model.
VI. PREDICTION ON PREDICTIONS, PoP

So far, we are apt to select the best model from many models in an accuracy sense. For example, we often explain that the newly proposed method is superior to the conventional methods. If this tendency is always true, then this makes sense. However, we sometimes encounter cases that method A and B produce the similar results, but method C does not; in one case, A is better than B, but in another case, B is better than A. The results vary according to the situations. We cannot simply accept which is better deterministically.

Let us take a new look at the prediction method. That is, we consider to use the combination method of these methods. In other words, we will make a prediction using the predicted values already obtained. We call this the PoP, the prediction on predictions. One idea for this is to use the trend of the predicted final values (the use of the decay function shown later), and the other is to select the better candidates for predicting the final value (the use of the ensemble method shown later and the use of mean value also shown later).

A. Use of the Decay Function

Looking at a trend itself by each prediction method (SIR, single ODE, or statistical truncated) in Figure 6, we may imagine a continuous curve fitted to the trend and its limiting value will converge to a constant value as days go on. Then, we assume the function

\[ d_i(t) = c_i - b_i e^{-a_i t} \]

where \( i \) means the prediction method id; \( a, b, c \) are constants to be fitted. The limiting value is \( c_i \). Figure 7 shows this conceptual idea to use the decay function in the case of the statistical truncated model. We may fit a curve decaying to predicted trend using the observed values until the truncation time \( T \).

B. Use of the Ensemble Method

When the observed data includes the randomness, a much more accurate estimation method may be applicable; that is, two heads are better than one. The idea is similar to the ensemble methods [15].

If each individual has the same probability \( p \) for success, then the value of the majority votes \( P \) can be expressed as

\[ P = \sum_{i=n+1}^{2n+1} \binom{2n+1}{i} p^i q^{2n+1-i} . \]

Figure 8 shows the relationship between \( p \) and \( P \). We can see that \( P > p \) whenever \( p > 0.5 \). For example, the values
of $P$ are
\[
P = 0.844, \quad (n = 1, \ p = \frac{3}{4}) \quad (8)
\]
\[
P = 0.790, \quad (n = 2, \ p = \frac{2}{3}) \quad (9)
\]
\[
P = 0.896, \quad (n = 2, \ p = \frac{3}{4}) \quad (10)
\]
This shows the effectiveness of the use of ensemble method.

C. Use of the Mean Value

Another method to use the predicted values is to take a mean value. This is simply to take a mean value of the three for the new prediction.

VII. THE CONDITIONS IN THE SIMULATION STUDY

We are dealing with the SARS case here. The literature [12] shows the mean incubation period of the disease is estimated to be 6.4 days (95% confidence interval is [5.2, 7.7]). If we apply the SIR method to the real data case, we refer to this information. However, in the simulation study mimicked to the SARS case, we assume that the incubation period is just one day because we try to compare the results with those obtained by other methods which may not require the value of the incubation period.

In the simulation, the final number of patients is set to 1,000, and $S_0$ is set to 5,000. The data generation is followed by the generalized logistic distribution function with parameter values, $\mu = 3.99$, $\sigma = 12.56$, $\beta = 3.27$, which came from the maximum likelihood estimates in the real SARS case in Hong Kong, 2003 [9]. In this paper, we have performed 100 simulation cases for the purpose of comparison.

VIII. PREDICTED RESULTS BY USING THE PoP

To show the trend to each prediction method, we made box-plots using 100 simulation cases as shown in Figure 10. The SIR shows the high bias to the final value in earlier stages. The ODE and the statistical truncated methods show the similar results, revealing the low bias in earlier stages. We may expect that a simple use of the mean value from the three may provide a better value because the single ODE and truncated results show the lower bias contrary to the SIR results.

A. Restricted Root Mean Square Error, $rRMSE$

To determine the accuracy for the prediction method, we can use the root mean square error. However, we introduce, here, the restricted root mean square error, $rRMSE$; since the predicted values for the final number of patients sometimes may have very large values or may not converge, we will locate these values to the boundary of the window (see dotted box in Figure 11 on the top). Here, $rRMSE$ is defined.
as
\[
\text{rRMSE}(j) = \sqrt{\frac{1}{|\Delta_j|} \sum_{k \in \Delta_j} (\hat{W}_r(\infty|t_T = k) - W(\infty))^2},
\]
(11)
where, \(\hat{W}_r(\infty|t_T = k)\) means the estimate of \(W(\infty)\) when using the data from the beginning to the truncation time \(k\) such as
\[
\hat{W}_r(\infty|t_T = k) = \min(\hat{W}(\infty|t_T = k), 2W(\infty)).
\]
\(\Delta_j\) expresses the days in the target area, and \(|\Delta_j|\) denotes the number of days in \(\Delta_j\). \(\hat{W}_r(\infty|t_T = k)\) attracts \(\hat{W}(\infty|t_T = k)\) at the boundary \(2W(\infty)\) if \(\hat{W}(\infty|t_T = k) > 2W(\infty)\).

\(\Delta_j\), for the rRMSE. On the top of the figure, the L-plot is shown, where we use the values in the window.

Figure 12 the rRMSE for L-plot of the SIR, single ODE, statistical truncated models, PoP methods (the ensemble method and taking the mean value) for the SARS Case. We can see that the ensemble method provides a good result. The ensemble method could remove the noisy estimates by the SIR method although the mean value was affected by this noise.

Figure 13 shows the rRMSE for L-plot of the SIR, single ODE, statistical truncated models, ensemble method, and the mean value after decaying process for the SARS Case. The figure reveals that the decaying process works and that the SIR and the ensemble methods show lower rRMSE values.

We may use the ensemble method as a PoP method.

IX. CONCLUSION

In observing the widely spread of patients caused by infectious diseases or the increase of the number of failures...
of equipments, it is crucial to predict the final number of infected patients or failures at earlier stages. To estimate the number of infected patients, the SIR model is commonly used even when the size of observed data is small. Other methods, such as the ordinary differential equation model, statistical truncated model are also useful to estimate the number of infected patients. These methods are also applicable to the case of fragile population such as the ordinary differential equation model, statistical truncated model with applications to lifetime analysis: unified censored and truncated models. IEEE Transactions on Reliability 54, 11-21, 2005.


Hirose, Models, Ensemble Method. By applying these methods to the SARS case, we have found that the ensemble method works well as one of the PoP methods.

REFERENCES

Hideo HIROSE was born in December 1951 in Japan. He obtained the Baccalaurate degree in mathematics from Kyushu University and the Dr. Eng. from Nagoya University in 1977 and 1988, respectively. He worked for Takaoka Electric Manufacturing Co., Ltd. from 1977 to 1995, and served as Vice Research Director from 1988 to 1995. He was a Professor at Hiroshima City University from 1995 to 1998. Since 1998, he has been a Professor at Kyushu Institute of Technology.

His interests include a variety of numerical computations such as finite element analysis, computational fluid dynamics, transient analysis of electrical networks, reliability engineering, and statistical data analysis.

He is a member of IEEE (Computer Society, Power Engineering, Reliability, Dylectic and Insulation societies), ASA (American Statistical Association), IMS (Institute of Mathematical Statistics), MPS (Mathematical Programming Society), ISI (International Statistical Institute), SIAM (Society for Industrial and Applied Mathematics, ACM (Association for Computing Machinery), AMS (American Mathematical Society) and 12 academic society members in Japan.

Yuki KOYANAGI was born on 2 February 1992 in Japan. He is currently a master student in Kyushu Institute of Technology.