Improved Forecast with a Combination of Mechanistic and Statistical Predictive Models

Georgiy V. Bobashev [1], Stephen P. Ellner [2], Barbara A. Bailey[3]

[1]RTI International, Research Triangle Park, NC
[2]Cornell University, Ithaca, NY
[3]San Diego State University, San Diego, CA
bobashev@rti.org, spe2@cornell.edu, babailey@sciences.sdsu.edu

Abstract

Predictive models based on past data could be good predictors of the future outcomes; however, they usually don't explain the causal and feedback relationships leading to the outcome. Conversely, mechanistic models could uncover complex interaction between underlying processes, but sometimes their calibration and validation could be unrealistic. Combining the two approaches into a semi-mechanistic model can lead to a winning combination. We present examples of historic epidemic data as well as simulated data, where a combination of neural networks with a mechanistic Susceptible, Exposed, Infected and Recovered (SEIR) model produces more reliable predictions with less parameterization.

Background

Predicting future data is usually based on a premise that similar inputs will produce similar outputs under a particular model. Thus, researchers predict future outcomes using past input data and corresponding output data, and they assume that the input-output model remains valid both in the past and in the future. One of the main challenges is identifying a space of “similar inputs.” This challenge is not only related to very complex systems where one needs to choose the governing factors out of hundreds of variables. Even simple systems, especially with chaotic and noisy behavior, could complicate the prediction base. Regression models assume that the outcome is related to the value of other variables at the same (or past) times, and autoregressive models also consider past outcome values. This principle remains true for spatial and functional prediction. For example, in social sciences the concept of Blau space assumes that the decision making of individuals not only depends on individual characteristics but also group characteristics, such as those of neighbors and other individual that share characteristics such as education, marital status, and type of job. These factors define distance in the Blau space. Although the need for identifying similar inputs is ubiquitous, the approaches could differ significantly depending on the understanding of what the criterion of good prediction is. Prediction and explanation approaches are sometimes contrasted to each other. Predictive models are often focused on predicting a number with some confidence intervals, while causal models are focused on the explaining what happens, under certain assumptions, when the system is perturbed in a particular way. For example, a simple predator-prey model (e.g., foxes and rabbits) could produce cyclic population dynamics. If a researcher is given a single time series for foxes, then a simple sine wave would fit the data perfectly. However, this fit will not explain why the dynamics is as such, what determines amplitude and periodicity, and how the system could be controlled in the future. Knowing the dynamics of rabbit populations and the relationship between the rabbits and foxes will not add anything to the numeric prediction, but will add a lot to the understanding of the processes controlling the populations. Much more complex infectious disease models in populations could provide good insight on complex causal and feedback relationships between hosts and pathogens, but when asked to predict the number of sick people during an influenza epidemic season, the models would not be refined enough to the make the forecast.

In this presentation, we would like to elaborate more on an approach proposed by Ellner et al. (1998) and Bobashev et al. (2000), where mechanistic models are combined with the predictive time series models to produce better prediction and to have some explanatory power.

Predicting Historic Epidemic Data

Consider a Susceptible, Exposed, Infected, Recovered (SEIR) epidemic model that is quite standard for many infectious diseases such as influenza, measles, smallpox, mumps, etc. In this model, the population is divided into 4 SEIR compartments, and after birth, an individual sequentially passes through each of these compartments with certain transition rates. Assuming a homogeneous mixing (i.e., everyone has an equal chance to meet anyone), the model is represented as a system of differential equations where individuals "flow" between the compartments. The basic equations are:

$$\begin{aligned}&dS/dt=-\beta(t)SI+m\\&dE/dt=\beta(t)SI-\upsilon E-mE\\&dI/dt=\upsilon E-\gamma I-mI\\&dR/dt=-\gamma I-mR,\end{aligned}\qquad (1)$$

where *S,E,I,* and *R* are the proportions of SEIR populations, respectively. *β(t)* is a seasonally varying contact rate, *υ* is 1/average length of the latent period, *γ* is the 1/average length of the infected period, and *m* is the birth/mortality rate. The population size is assumed to be at an equilibrium and the mortality rate is equal to the birth rate, and both are constant.

Depending on the parameter value, this model could exhibit a steady state, periodic behavior, and the cascade of period doubling into chaotic regimes. Historic measles epidemic data (Figure 1) suggests that it could correspond to the chaotic regime of the SEIR model (Ellner et al. 1995, Grenfell et al., 1995).

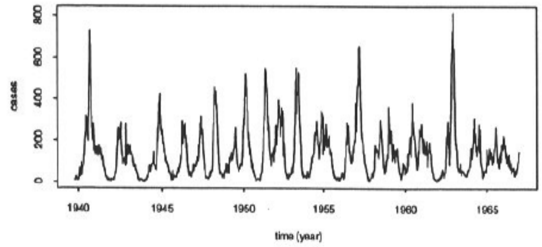

Figure 1. Measles incidence data for Liverpool before mass immunization.

A large body of literature exists where SEIR models are fitted to the data and used to produce estimates of future outbreaks. However, the question about other factors that can impact the prediction could not be adequately accounted for within SEIR framework. We suggest that once other components of the epidemic process could be obtained from even a qualitative fit (i.e. up to some linear transformation), these components can be used as additional predictors in a statistical model and, thus, improve its predictive performance. In SEIR model such additional component could be the prevalence of the susceptible population. In diseases like pandemic influenza at the time of an outbreak all population could be considered susceptible. For seasonal diseases, such as measles before mass vaccination, or common influenza, the proportion of susceptible population is not known a-priori but could be reconstructed using some basic mechanistic assumptions (Bobashev et al, 2000). The essence of the reconstruction method is that susceptible is being reduced by the epidemic process and being replenished by the newborns. Thus, knowing the birth rate and the epidemic incidence and assuming that almost everyone got measles during their lifetime before mass vaccination, it is possible to generate a complimentary time series of susceptible prevalence.

In our study we used a statistical time series model that incorporates both seasonal forcing and scaled S_t. One of the underlying assumptions is that the number of newly infected individuals depends only on the current number of infected and susceptible and the contact rate between them. Although we don't know the exact seasonal shape of the contact rate we used sine and cosine waves with annual periodicity as model inputs hoping that through the data fitting process it would partially account for the true seasonal trends. Thus, we considered a model of the form:

$$C_{t+\tau}=F(C_t,S_t,\sin(\omega t),\cos(\omega t))+\varepsilon_t\,, \qquad (2)$$

where F is an estimated function, periodic components represent seasonal forcing, and ε_t is a random exogenous noise. The simplest form of the model would be a regression, however, because we believe that the epidemic process is nonlinear, we used a neural network approach, with 1 and 2 hidden levels and 2 and 3 nodes per level. As the control model we have considered a "more phenomenological" model successfully used in Ellner et al. (1998):

$$C_{t+\tau}=F(C_t,C_{t-l},\ldots,C_{t-ml},\sin(\omega t),\cos(\omega t))+\varepsilon_t\,,$$
(3)

where function F was also implemented as a neural network fit with 1 and 2 hidden levels and combinations of 2 and 3 nodes per level.

In both models the quality of fit was measured by using Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).

We trained the models on the first half of the weekly epidemic data and tested them on the second half. In order to measure the goodness of fit, we used a so-called *pseudo* R^2 criterion (Ellner et al. 1998). The functional form of *pseudo* R^2 is

1-Mean square error(residuals)/Variance(data)

If *pseudo* R^2 is less than zero, it means that the model makes worse prediction than just using the mean.

We have used mechanistic methodology described in Bobashev et al. 2000 to reconstruct the prevalence of the susceptible population and used that prevalence as a covariate in a neural network model 2. A comparison of predictions made by models 2 and 3 shows that model 2 provides better prediction of future data. (Figure 2)

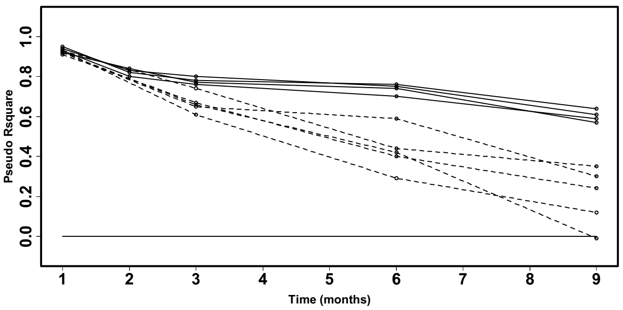

Figure 2. Prediction of Liverpool incidence data several months ahead using models 2 (solid lines) and 3 (broken lines).

As was shown before, (Ellner et al. 1998) addition of known mechanistic components improves the forecast because it "helps" the predictive model, such as neural network, to define the shape of the relationships between the set of predictors and the outcome. Long-term prediction of chaotic trajectories is known to be impossible. However the addition of a known variable that reduces the strength of chaotic properties (e.g., reducing a Lyapunov exponent) will lead to the forecast improvement. For example, the number of susceptibles at the end of an epidemic could be quite different for the same peak size (Figure 3) and the number of susceptible at the end of the epidemic dictates (together with the birth rate) when the next epidemic might start.

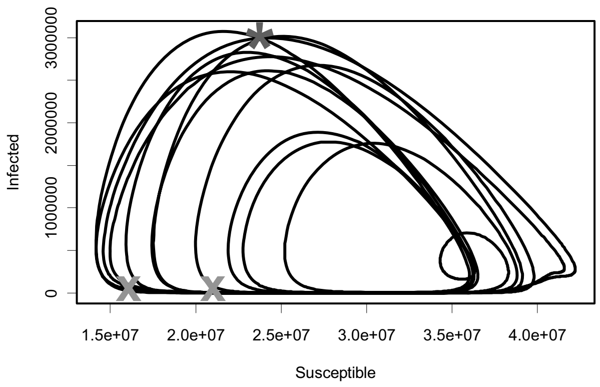

Figure 3. Shapes of relationships between the susceptible and infected populations in SEIR model. For two epidemic trajectories the value at the peak is the same (marked with the star) but the values for the susceptible at the end of the epidemic is very different (marked with an x)

Discussion

The presented approach could be used to improve prediction of not only real data but could be used to differentiate one model from another by fitting neural network models to the output of the simulation models, and using several outputs could help distinguish which model has generated the data. Although the predicted approach has been developed to fit epidemic data it could also be applied to a broad range of predictive models that are based on time series of events and have a theoretically valid mechanistic model. Such emulators, for example, could be used as simple desk-top predictive models based on more complex agent-based models that require significant computational resources.

References

Anderson R. M., and May, R. M. 1991. *Infectious Diseases of Humans: Dynamics and Control.* Oxford: Oxford Univ. Press.

Bobashev, G. V.; Ellner, S.; Nychka, D. W.; and Grenfell, B. 2000. Reconstruction of Susceptible and Recruitment Dynamics from Measles Epidemic Data." *Mathematical Population Studies* 8(1): 1–29.

Ellner, S.; Gallant, A. R.; and Theiler, J. 1995. Detecting Nonlinearity and Chaos in Epidemic Data. In *Epidemic Models: Their Structure and Relation to Data*. D. Mollison, ed. Proceedings of NATO ARW on Epidemic Models. Cambridge: Cambridge Univ. Press.

Ellner, S.; Bailey, B.; Bobashev, G. V.; Gallant, A. R.; Grenfell, B.; and Nychka, D. W. 1998. Noise and Nonlinearity in Epidemics: Combining Statistical and Mechanistic Modeling to Characterize and Forecast Population Dynamics. *American Naturalist* 151(5): 425–440.

Grenfell B. T.; Kleczkovski, A.; Ellner, S.; and Bolker B. M. 1995. Non-linear Forecasting and Chaos in Ecology and Epidemiology: Measles as Case Study. In *Nonlinear Time Series and Chaos, Vol.2, Proceedings of the Royal Society Discussion Meeting* 345–371. H. Tong, ed., Singapore: World Scientific.

Sugihara, G., and May, R. M. 1990. Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time Series. *Nature* 344: 734–741.