Commentary on "Transparent modelling of influenza incidence": the need to justify complexity

The naïve forecasting method tested by Konstantinos Katsikopoulos and his fellow researchers (Katsikopoulos et al., this issue) has five key attributes. It is simple, transparent, inexpensive to apply, underpinned by theory and, most importantly, relatively accurate. Yet often researchers are tempted to overlook the possibility that reliable forecasts can be obtained by applying simple accessible rules. The world can appear to be complicated and messy. Such a world, it is often assumed, must require the application of advanced intricate models to have any chance of providing a faithful representation of its features -and the greater a model's intricacy, the greater the predictive accuracy it should yield. Besides the expectation of greater accuracy, there are other reasons why a modeller may opt for complexity. Advanced methods with names such as a neuro-fuzzy-stochastic frontier analysis or genetic fuzzy systems with data clustering sound impressive and can bring prestige to those working with them. The challenge of building complex forecasting schemes can also be intellectually rewarding.

Whatever their motivations, it is reasonable to expect that any researcher who suggests the use of a complex forecasting method in a particular situation should ensure that their recommendation satisfies a number of criteria. First, they should consider whether their method and its output are likely to be acceptable and transparent to its intended users, who may not be mathematicians or statisticians. Back in the 1980s, Taylor and Thomas (1982) found that staff at British Gas were sceptical about the use of an advanced - but to them mysterious - method that was designed to forecast the daily demand for natural gas. This was despite the general accuracy of the method -the intended users tended to focus on, and recall, its rare large errors. It is clearly pointless to devote resources to the development of a method that is never used or constantly overridden, however elegant and accurate it may be.

Where models lack transparency or are not understood, they also deny users the opportunity to learn about the structure and dynamics that apply to the domain they are forecasting. For example, marketeers may be prevented from gaining insights into the key drivers of a product's demand. Understanding a method's rationale may also serve to alert users to its limitations and the circumstances where its forecasts do, and do not, require adjustment. If forecasts are generated by a black box, there is a danger that people will make judgmental adjustments for the effects of forthcoming events that have already been incorporated into the algorithm, leading to double-counting.

Second, any proposed new method should be compared for accuracy against both established and simpler benchmark methods -in the latter case to establish if the increased complexity is justified both in terms of enhanced accuracy and additional cost. As Katsikopoulos et al. point out, this testing was absent in the case of Google Flu trends. Even when benchmarks are used, they are often confined to comparing one complex method with another (recent examples include Rezaee et al., 2019, Wang and Wang, 2020, and Zhang et al., 2020). The use of naïve forecasts as a benchmark is likely to be particularly appropriate in a wide range of circumstances akin to those found in flu forecasting where there is little or no underlying stability. In many of these circumstances, such as in stock market forecasting, naïve forecasts may be difficult to beat. When time series *are* characterised by stable underlying structures with noise superimposed, one would expect the use of appropriate models to lead to the gains in accuracy over naive forecasts. However, in practice, these gains may not always be realised because of errors in estimating models. Alternatively, the gains may be insufficient to justify the extra costs of complexity. Goodwin et al. (2017) provide equations for calculating the theoretical upper limits of gains in accuracy than can be achieved over naïve forecasts for several types of stationary time series. For example, for an ARIMA(1,0,0) series, with a first-order autocorrelation of 0.7, even an optimal model can only reduce the mean squared error (MSE) of naive one-period-ahead forecasts by 15 per cent in the long run.

Third, accuracy assessment should be based on a method's ability to forecast out-of-sample observations and a sufficient number of these observations should be employed to ensure that reliable inferences can be drawn about the method's performance (Tashman, 2000). The problem is that complex methods can be greedy in that they can require long series of past data to estimate parameter values. When there are constraints on data availability, it is tempting to use most data points in the model fitting process, leaving too few observations for reliable out-of-sample testing. Goodwin (2011) reported several instances of this. In one case, researchers claimed that the use of their 'flexible integrated meta-heuristic framework based on an artificial neural network multilayer perceptron' would provide "more reliable and precise forecasting for policymakers" concerned with electricity supply. But they tested their method on just eight unseen annual household electricity figures from Iran. Based on just two out-of-sample observations, other researchers claimed that their Polynomial Curve and Moving Average Combination Projection (PCMACP) model could "reliably and accurately be used for forecasting natural gas consumption". In the light of cases like this, Goodwin suggested the tongue-in-cheek rule that: "If the name of a method contains more words than the number of observations that were used to test it then it's wise to put any plans to adopt the method on hold."

All of this is not intended to dismiss the potential value of complex models in forecasting. As Katsikopoulos et al. point out, these methods may be appropriate in relatively stable situations, as long as factors such as algorithm aversion (Dietvorst et al., 2015) and the need for transparency do not preclude their use. Indeed, where the environment is stable, modern computing power provides an exciting opportunity to apply more powerful techniques in pursuit of greater forecasting accuracy. However, the findings of researchers such as Katsikopoulos et al. and others, including Green and Armstrong (2015), suggest that, in a volatile world, the range of situations where highly complex methods can be exploited may be more limited than many researchers expect.

References

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144, 114.

Goodwin, P. (2011). High on complexity, low on evidence: Are advanced forecasting methods always as good as they seem? *Foresight: The International Journal of Applied Forecasting*, (23), 10-12.

Goodwin, P., Petropoulos, F., & Hyndman, R. J. (2017). A note on upper bounds for forecast-value-added relative to naïve forecasts. *Journal of the Operational Research Society*, 68, 1082-1084.

Green, K. C. and Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, 68, 1678-1685.

Katsikopoulos, K., Şimşek, Ö., Buckmann, M., & Gigerenzer, G. (this issue). Transparent modeling of influenza incidence: Big data or a single data point from psychological theory? *International Journal of Forecasting.*

Rezaee, M. J., Dadkhah, M., & Falahinia, M. (2019). Integrating neuro-fuzzy system and evolutionary optimization algorithms for short-term power generation forecasting. *International Journal of Energy Sector Management*. 13, 828-845.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of forecasting*, 16, 437-450.

Taylor, P. F., & Thomas, M. E. (1982). Short term forecasting: horses for courses. *Journal of the Operational Research Society*, 33, 685-694.

Wang, B., & Wang, J. (2020). Deep multi-hybrid forecasting system with random EWT extraction and variational learning rate algorithm for crude oil futures. *Expert Systems with Applications*, *161*, 113686.

Zhang, J., Li, L., & Chen, W. (2020). Predicting stock price using two-stage machine learning techniques. *Computational Economics*, 1-25.