Advanced forecasting techniques

How to use advanced forecasting techniques for estimating demand of NHS services

NHS England and NHS Improvement

Contents

Introduction

This document will provide an overview of some advanced forecasting techniques that can be used as part of demand and capacity modelling for NHS services. It will also review the basic forecasting steps, how to choose and evaluate a method, links to the current suite of demand and capacity models, and software packages which can carry out these techniques, including both open-source and licensed products.

This document will not provide an in-depth tutorial on forecasting methods. If you are interested in learning about these methods in more detail, we recommend the online book *Forecasting: Principles and Practice*[1]*.* This is a freely available online textbook that goes through the main forecasting techniques in detail and is a good introduction to the topic. It is based in R (an open-source statistical program) and includes a full set of exercises and example datasets.

[1] The textbook can be accessed via the following link: http://otexts.com/fpp2

Estimating demand for NHS services using advanced forecasting techniques

Accurate forecasting of demand is an important part of demand and capacity modelling. In many cases, a simple approach (e.g. ‘next year will be similar to last year’) will work well enough for planning purposes, and this is the approach used in the suite of tools developed by the National Demand and Capacity Programme for elective care.

There are services, with more complex and unstable patterns of demand, where this simple approach will not work, and more advanced forecasting techniques are needed to better predict future demand. For instance, in an A&E setting, demand figures differ drastically between seasons, days of the week, and time of the day. It is also reasonable to assume that some aspects of the past patterns will continue into the future.

For these complex and unstable services, advanced forecasting techniques are needed to make a good estimation of future demand. These techniques improve the quality of forecasts, as they will account for additional factors such as significant variation, seasonality and historic growth trends which are not accounted for in the existing model suite.

The predictability of any variable depends on several factors including:

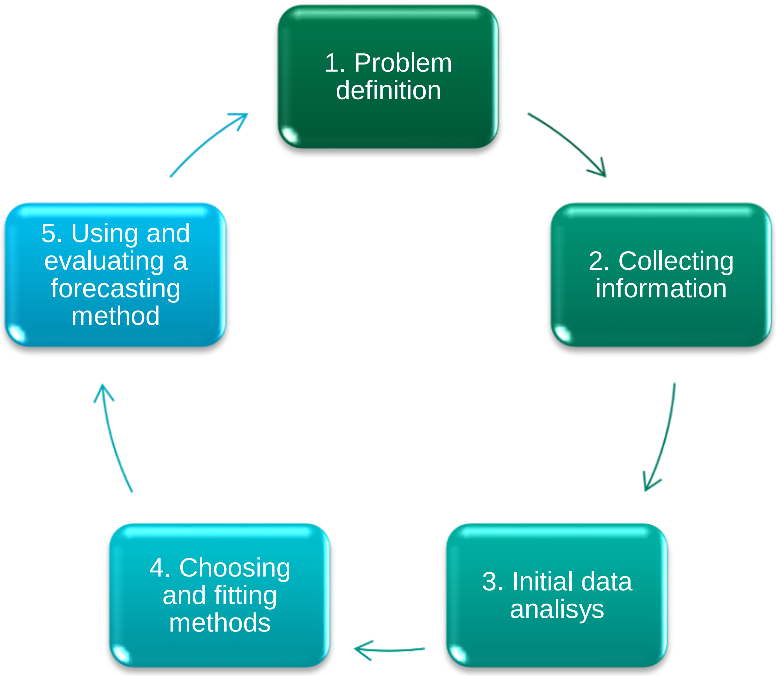
1. how well we understand the factors that contribute to it;
2. how much data is available;
3. whether the forecasts can affect the thing we are trying to forecast.

For instance, the number of attendances to Emergency Departments is well understood to depend on the time of the day, day of the week, and month of the year, among other factors. Emergency Departments are required to collect their attendance data on a regular basis and the forecasts don’t have an immediate effect on the expected number of attendances. Hence, Emergency Department attendances are highly predictable. However, some degree of error is also expected and natural.

Advanced forecasting techniques are commonly used in other industries but require specialist knowledge and training to use effectively. Therefore, this document provides an overview of some of the available advanced statistical forecasting techniques, how to select the appropriate one, existing software, how to interpret and use the output, links to the NHS demand and capacity models, and their limitations.

Forecasting: the basic steps

Forecasting is an iterative process consisting of a number of steps. The whole cycle can be repeated several times like an improvement process. The steps are as follows:



1. Problem definition: Understanding what the forecasting will be used for, who requires the forecasting, and how it fits in with the organisation. This step is perhaps the most important and not always a simple one.

2. Collecting information: Collecting the relevant information that will be used to drive the model (e.g. referrals, attendances). Beyond the basic task of collecting the information, this step also encompasses using the expertise of the individuals who use this data – operational, clinical and administrative staff – to help tell the story behind the numbers. This will sense check the validity of your source information and result in a better model.

3. Initial data analysis: Before you begin to model or forecast a given time series, an initial analysis of the data should be carried out, giving particular attention to the identification of important features such as autocorrelation[1], seasonal patterns, cyclical variations, trend, outliers, and any other significant fluctuations in the series.

Initial data analysis should also evaluate whether the time series is stationary (i.e., if basic statistical properties such as the mean and variance of the series remain constant through time). Most time series methods are based on the assumption that the data is stationary; if the time series is non-stationary, one or more data transformations may be necessary to make the time series stationary before further analysis can take place[2].

4. Choosing and fitting methods: the choice of method will depend on the question that is being asked, the available data, and the relationship between the variable and the explanatory variables.

Once a suitable method has been chosen, we then need to fit the model. This is carried out by adjusting the model parameters, so they replicate historical data.

5. Using and evaluating a forecasting method: the model can be evaluated when the data for the forecast period becomes available. The model forecasts and the real observations are measured and the error between the two are estimated. More detail on how to evaluate the accuracy of a model is provided later in this document.

[1] Autocorrelation refers to how correlated a time series is with its past values. An autocorrelation function (ACF) plot can be used to identify the strength and time offset of these autocorrelations.
[2] Some techniques that can be used to help with this analysis can be found here: https://otexts.com/fpp2/graphics.html and https://otexts.com/fpp2/decomposition.html

Advanced forecasting techniques

Some of the most common advanced forecasting techniques are:

Box–Jenkins, or ARIMA (autoregressive integrated moving average)
ARIMA models are stochastic models that combine elements of moving average methods and autoregression methods.

ARIMA modelling has four stages: before starting, the time series needs to be differenced until it is stationary. Its Autocorrelation (ACF) and Partial Autocorrelation Functions (PACF)[4] are compared with ones from various theoretical time-series to estimate the order of the process. Second, the parameters of the model are estimated. Third, the residuals are examined to see if the model is adequate. Finally, alternative models are considered.

SARIMA
Seasonal autoregressive integrated moving average (SARIMA) models extend basic ARIMA models and allow for the incorporation of a repetitive pattern, such as the weekly pattern observed in daily ED patient volumes.

When working with time series data that display seasonal patterns, it is important to not only identify the correlation between current observations and their immediate predecessors, but also to determine whether correlation exists between current observations and their predecessors from previous seasons. This is referred to as evaluating the time series at both the nonseasonal and seasonal levels.

Exponential Smoothing
Exponential smoothing is a term that is applied to a variety of methods that generate forecast-based formula that weight recent observations more heavily than more remote observations (based on weighted moving average formula). Exponential smoothing techniques include parameters for estimating the level (ie. mean), trend, and seasonality of a dataset, which can be used for forecasting.

Prophet
Facebook released a package implementing a Bayesian forecasting approach. This method recognizes repeating patterns over weeks, months, years, and identified holidays. Prophet is set up as an automated process and can be installed as a

[4] *Partial autocorrelation* is the autocorrelation that results after removing the effects of any autocorrelations due to terms at shorter lags.

package in R or Python. The basic methodology is an iterative curve-matching routine, where Prophet will then train your data on a bigger period, then predict again and this will repeat until the end point is reached.

The development team of Prophet claim that its strengths are:

- Working with high-frequency data (hourly, daily, or weekly) with multiple seasonalities, such as hour of day, day of week and time of year;
- special events and bank holidays that are not fixed in the year;
- allowing for the presence of a reasonable number of missing values or large outliers;
- accounting for changes in the historical trends and non-linear growth curves in a dataset.

Further advantages include the ability to train from a moderate sized dataset, without the need for specialist commercial software, and fast start up times for development[5]. While more advanced models are developed, time-series–based prediction offers the possibility of improving analytical capability in the short term.

However, an important principle of forecasting in general is that these tools are best applied thoughtfully, with consideration of their strengths and limitations.
Tuning parameters can help to optimize the outputs of the models. If optimizing Prophet outputs is of interest, information regarding it is available online[6].

TBATS

The names are acronyms for key features of the models: Trigonometric seasonality, Box-Cox transformation, ARIMA errors, Trend and Seasonal components. In essence, TBATS evaluates multiple forecasting techniques against a training dataset, and picks the 'best' method based on some key metrics (this is discussed in more detail later in the document).

TBATS uses a combination of Fourier terms with an exponential smoothing state space model and a Box-Cox transformation, in a completely automated manner. In a TBATS model the seasonality is allowed to change slowly over time, while other methods force the seasonal patterns to repeat periodically without changing. A downside of TBATS model, however, is that they can be slow to estimate, especially with long time series. As TBATS is automated sometimes the prediction is not useful, due to automated parameters that do not represent the reality of the observed variable.

Artificial Neural Networks

Artificial neural networks are part of machine learning and designed to mimic the architecture of the human brain. They can be used to model complex nonlinear relationships between inputs and outputs. Artificial neural networks have been demonstrated to be highly effective in applications such as pattern recognition and

[5] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2712176 and https://kourentzes.com/forecasting/2017/07/29/benchmarking-facebooks-prophet/
[6] https://towardsdatascience.com/implementing-facebook-prophet-efficiently-c241305405a3

classification. Since forecasting is essentially the process of identifying patterns from observed data and extrapolating them into the future, one would expect artificial neural networks to perform well at forecasting tasks. An artificial neural network is formed by a network of computing units, called neurons, which are connected to each other, forming a network. The strength of each connection, or weight, is updated iteratively as the network is trained, so in effect it “learns” to recognize patterns as it is provided with data.

Choosing a method

What is the 'best' method for modelling a given time series? There will be a range of considerations you will need to take into account, beyond the basic need for a reasonable degree of accuracy – availability of software, training and knowledge required, timeframe and scale of analysis. The table below[7] outlines some of the pros and cons of the forecasting methods previously described.

Forecasting Methods	Pros	Cons
SARIMA	Theoretically appropriate methodology for most time series	Complex statistical methodology that requires a higher level of expertise and experience than linear regression.
	Capable of modelling seasonal variation, trend, autoregressive, and moving average processes.	The modelling process is less informative than linear regression.
	Univariate method—no external data necessary.	Generally provided less accurate forecasts of daily ED volumes than the linear and time series regression models.
	Statistical software widely available.	
Exponential smoothing	Fully automatic, low level of expertise required, and quick to implement.	Not based on formal statistical model or theory.
	Capable of modelling seasonal variation, trend, autoregressive, and moving average processes.	The modelling process is less informative than that of linear regression.
	Effective when the parameters describing the model are changing over time.	Generally provides less accurate forecasts if your dataset includes strongly cyclical or seasonal variations (although Holt-Winters can accommodate trend and seasonal components).
	Available in most standard statistical software packages, open source and commercial.	
ARIMA	Capable of modelling variation, trend, autoregressive, and moving average processes.	Complex statistical methodology that requires a higher level of expertise and experience than linear regression.

[7] Adapted from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1553-2712.2007.00032.x (accessed on 17 Dic 2019)

	Easily interpretable results.	Multiple variables, which requires additional data collection and parameter estimation.
	Informative modelling process.	
	Statistical software widely available.	
	Consistently provided more accurate forecasts of daily ED patient volumes than the linear regression models.	
Prophet	Fully automated routine - recognizes repeating patterns over weeks, months, years, and identified holidays.	Often requires tuning the parameters to get a more accurate estimation.
	Simple syntax allows you to generate results with a single command.	Can take a long time to run for moderate sized datasets in R – could be problematic if you need to run this across multiple datasets.
	Available in R and Python.	
	Can be configured to take account of irregular holidays (eg. Easter).	
TBATS	The seasonality is not static, it can change slowly.	It can take a long time to run for large data series to be predicted.
	Automated - tests several methods (eg. ARIMA, exponential smoothing), and chooses the best fit.	
	Available in Python and R.	
Artificial neural network	Capable of modelling complex, nonlinear systems.	Black box modelling procedure means that the final model is not transparent.
	Allows for rapid adjustment to changes in the time series.	Statistical software packages provide fewer, less mature procedures to estimate artificial neural network models.
		Generally provided less accurate forecasts of daily ED volumes than the linear and time series regression models.
		Requires specialist knowledge to run and train models.

Evaluating the methods

When evaluating forecasting methods, there is no general ‘best’ method. Typically, we seek to find the method that has the least ‘error’ (ie. difference between predicted and actual value) overall. There are a number of error indicators which can be used to assess a model.

Error indicators that are commonly used are:

- MAPE: Mean Absolute Percentage Error
- MAD: Mean Absolute Deviation
- MSD: Mean Square Deviation

MAPE is the most commonly used technique. We will look at MAPE and the Akaike Information Criterion (AIC), which is another method of evaluating statistical models that is used, especially for techniques that use a combination of methods.

MAPE

MAPE calculates the mean absolute percentage error of a forecasted variable against the occurrence. If X(1), X(2), . . . denote observations in a time series, then we denote the observation at time t by X(t). For example, X(t) might denote the number of actual attendances at the ED in month t. If a forecasting method is used to predict the value of the time series at some time t, then we will denote the predicted value by X*(t). Hence, the error in our prediction would be X(t) – X*(t), the difference between the actual value and the predicted value. The ultimate test of any forecasting method is the size of these errors.

Akaike Information Criterion (AIC)

The AIC is a method of estimating the relative quality of a group of statistical models for a given set of data. The algorithm used to calculate AIC will measure two key metrics: how well the model fits the data, and how complicated the model is. A low AIC score is considered ‘better’. The final score is a useful way of selecting the ‘best’ of several models. For example, TBATS integrates the AIC into its model selection process.

TBATS will consider various models, such us:

- with Box-Cox transformation and without it.
- with and without Trend
- with and without Trend Damping
- with and without ARIMA process used to model residuals
- non-seasonal model
- various amounts of harmonics used to model seasonal effects

The model with the lowest AIC score will be selected as the final method.

It is worth noting that the AIC score only considers how ‘good’ a model is compared to a group of other models and is not an assessment by itself of how accurate a model is.

Available software

A list of available software packages is accessible with the link
https://en.wikipedia.org/wiki/Comparison_of_statistical_packages

We have highlighted here two common, and freely available packages which are commonly used for data science.

R

R is open-source software, available on almost every operating system, and there are thousands of add-on packages to do almost anything with regards to forecasting techniques. We would recommend using R together with RStudio. Instructions for downloading, installing and running R and RStudio is available online[8]. Documentation for R is extensive and freely available online, and additional modules can be easily installed, particularly when using RStudio.

R is an excellent platform for prototyping statistical programs, however it does run slowly on large datasets, and scaling up programs for large scale production of reports can be difficult.

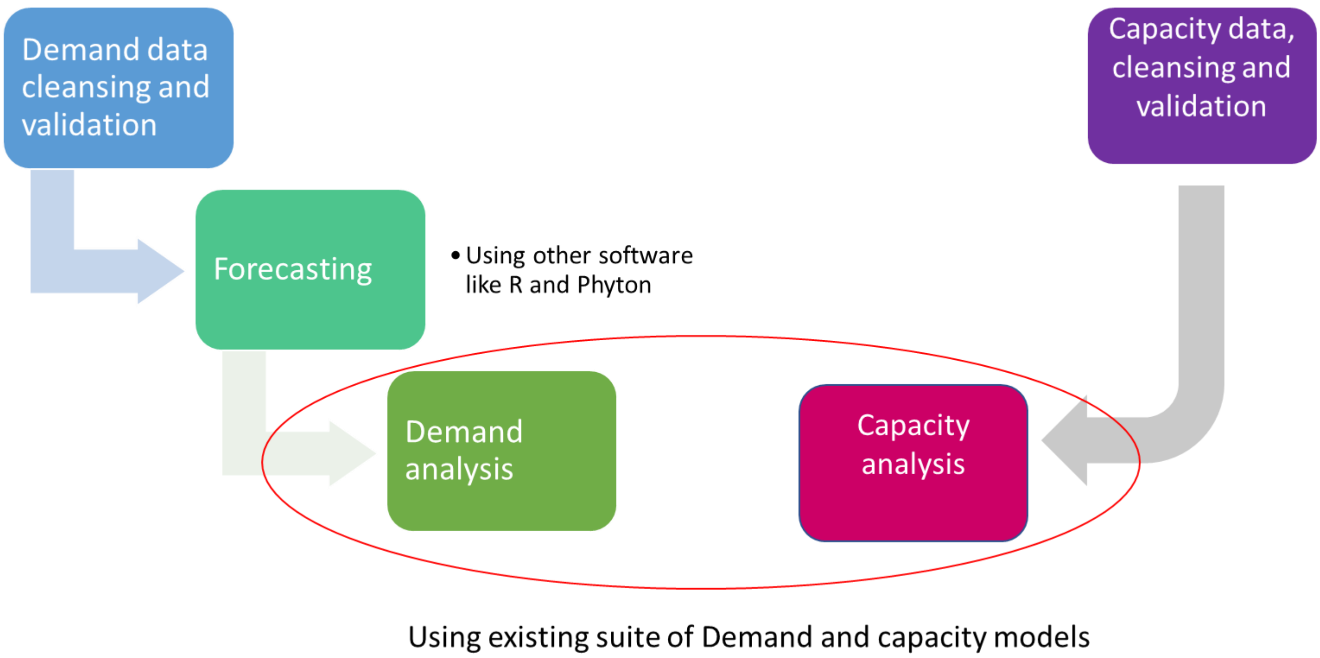
Python

Python is a programming language known for its relatively simple syntax and vast library of available modules and packages which can be used by relatively novice users to achieve results quickly. Full information on the Python programming language can be found online.

There is a vast array of information and tutorials available online for completing time series forecasting in Python, with one example given below.[9] Depending on the packages and datasets used, there may be slight improvements in the speed of data analysis compared to R.

[8] https://otexts.com/fpp2/appendix-using-r.html
[9] https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b

Link to Demand and Capacity models



Forecasting packages within R or Python can be used to provide an estimation of the expected demand for the next 52 weeks, based on the historical data. This output can be used as an input for the current set of NHS demand and capacity models which will estimate the required capacity for NHS services at an operational level. There is an important distinction between estimated demand and required capacity. Other factors need to be considered when completing a demand and capacity analysis for NHS services, such as the level of DNAs (Did Not Attend), slots lost and re-bookings. They will have a large impact and generate additional demand.

References

Champion, R., Kinsman, L., & others. Australian Health Review February 2007 Vol 31 No 1. http://www.publish.csiro.au/ah/pdf/AH070083

Jones, S.A., Joy, M.P. & Pearson, J. Health Care Management Science (2002) 5: 297. https://doi.org/10.1023/A:1020390425029

Kasapoglu OA. Selection of the Forecasting Model in Health Care. J Hosp Med Manage. 2016, 2:2. https://hospital-medical-management.imedpub.com/selection-of-the-forecasting-model-in-health-care.pdf

Skorupa, G. Forecasting Time Series with Multiple Seasonalities using TBATS in Python.Online. (Accessed November 2019) https://medium.com/intive-developers/forecasting-time-series-with-multiple-seasonalities-using-tbats-in-python-398a00ac0e8a

Spencer, J., Alun, T. & others. Forecasting Daily Patient Volumes in the Emergency Deparment. Academic Emergency Medicine. February 2008. Vol 15 N. 2. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1553-2712.2007.00032.x

Van der Merwe, R. Implementing Facebook Prophet efficiently. Online. (Accessed September 2019) https://towardsdatascience.com/implementing-facebook-prophet-efficiently-c241305405a3