Use of Information Criteria for finding Box-Jenkins Time Series Models for Patent Filings Counts Forecasts

Peter Hingley and Gerhard Dikta

***Abstract*—An automatic model selection procedure for Box-Jenkins time series models is used to forecast future levels of annual incoming patent filings at the European Patent Office. Selection of the best model uses either the Akaike's Information Criterion (AIC) or a novel data based version of the Exact Information Criterion called $|EIC_w|$ that is developed here.**

The methods are applied separately to autoregressive integrated moving average (ARIMA) and to autoregressive distributed lag (ADL) models. For ADL, explanatory variables include terms that are based on world Gross Domestic Product and world Research and Development Expenditures series.

We take a multi-year approach to study a sequence of overlapping 20 year data windows. Selections based on the two information criteria are made on each window. Different models can be selected by the two criteria. We compare the forecasting accuracy of the models for five years beyond the windows. Generally it appears that ADL models have lower forecast errors than ARIMA models. On the whole, the selections based on AIC have lower forecast errors for ARIMA models, while selections based on $|EIC_w|$ have lower forecast errors for ADL models. There is some serial correlation between the successive windows for the AIC values of its selected models that is not shown for the $|EIC_w|$ values of its selected models. There is also an unexpected negative correlation between AIC values and forecast errors.

$|EIC_w|$ can be recommended as a complement to AIC for selecting models in forecasting contexts, to give a fuller picture of the interactions between models and data.

***Index Terms*—ADL, Akaike´s information criterion, ARIMA, exact information criterion, Kullback-Leibler information, model selection, patent forecasts, time series.**

I. Introduction

Box-Jenkins methodology is often used to make forecasts in a business setting ([1]). Here an annualised set of historical data on patent filings is forecasted. Self determining autoregressive integrated moving average (ARIMA) models are used. Autoregressive distributed lag (ADL) models are also used, that make use of two additional explanatory variables:- Research and Development expenditures growth rates and Gross Domestic Product growth rates.

Tests are used to determine the appropriate degree of differencing for the series. All possible models up to a certain degree of complexity are studied. Then, at the accepted degree of differencing, a standard approach is to select the model that gives a minimum value for Akaike´s Information Criterion (AIC) ([2]). This penalises the goodness of fit of each model to the data by the number of estimated parameters. Forecasts up to five years ahead are generated by the selected models. Similar techniques are currently available in the "auto.arima" package that is under R [3], although our approach has some features that are not available in that package.

The European Patent Office (EPO) does an annual forecasting exercise. When this happens, each time series remains essentially the same as in the previous exercise, but with one new point added. A problem with the selection process is that different models may be selected in successive exercises. Thus it is difficult to propose that any one such selected model is correct. This is unfortunate because it would be good to ascribe an assumed mechanism to the underlying process. There can also be difficulties when dealing with a selected model that provides unrealistically flat or decreasing forecasts. In fact there have been almost continuous increases from year to year in the historical patent filings series.

As to the practical background, the EPO forecasts patent filings for budgetary planning purposes. Total Filings (TFs) are studied, that are the sum of direct filings at the EPO and world filings under the international phase of the Patent Cooperation Treaty [4] [5]. An annual exercise is carried out to fit the Box-Jenkins models. The results are considered together with the results of a customer survey [6] and another model [7], in order to come up with a consensus forecast. After the EPO opened in 1977, it took some time before clients fully exploited the system to complement the services that were already offered by the national patent offices in Europe. Here the TFs are considered from 1987 to 2018, without further breakdowns by countries of origin, technical areas etc.

There are two main aims of this study. Firstly a new alternative information criterion to AIC is proposed for the model selection process. This is a variant of the previously described Exact Information Criterion (EIC) ([8]). To our knowledge, this kind of information criterion has not previously been used in time series modelling.

Secondly, the practical forecasting problem for patent filings is examined by carrying out a retrospective modelling exercise. With this, two questions are addressed. What is the best model for each class that can be fitted to the time series from 1987 up to 2018? Then, which models are most appropriate for subsets of the data after breaking them down into overlapping 20 year windows during the period? Macroeconomic disruptions were caused by events such as the great recession (circa 2009). What then is the advantage, if at all, in finding the best fitting model to each window by minimising an information criterion? The results can help

Manuscript received May 26, 2021; revised October 30, 2021.

P. Hingley has retired and lives in the United Kingdom.
(e-mail: Peterhingley05@gmail.com)

G. Dikta is a Professor in Mathematics and Applied Mathematics at the Aachen University of Applied Sciences (FH), Germany.
(e-mail: Dikta@fh-aachen.de)

to decide whether to impose a given model or whether to accept the variations of the model specifications that may be suggested in successive forecasting exercises.

The study examines whether either or both of ARIMA and ADL model types are adequate for the required forecasting job, by measuring the forecast errors from the series of overlapping windows. The two alternative measured information criteria are compared in terms of their forecasting accuracies over the five years of data beyond the data used to fit the models in each window. The information criteria are second order effects that relate indirectly to the residual sample variance of the fitted model. Therefore it was not clear a-priori how big the effect of choosing a different criterion to the usual one would have on the forecasts. We also report on the serial correlations of the forecast errors and of the information criteria as well as on the correlations between them.

The next section describes the data that are to be analysed in more detail. Then follow sections that describe the methodology and the results. The final discussion section makes recommendations about how to use such information criteria for this particular suite of forecasting models, for these data and for other systems.

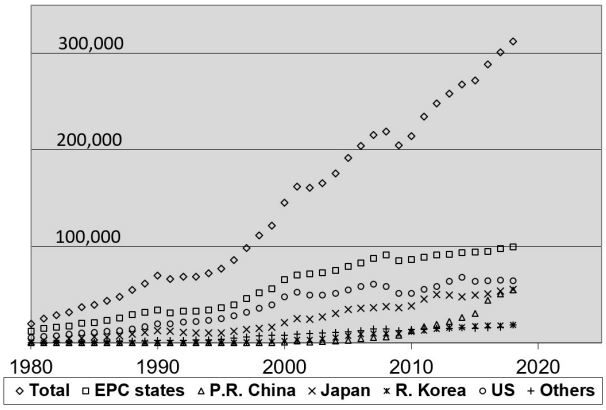
II. Data to be analysed

The sources for the time series are as follows. TFs are from the EPO epasys production database. R&D expenditures are business enterprise research and development expenditures from the OECD Main Science and Technology (MSTI) database [9], as available in early 2019, at constant prices and purchasing power parity in 2010 US dollars. GDP expenditures are from the World Bank [10], as available in early 2019, at constant prices and purchasing power parity in 2005 US dollars, with data for 2018 from estimates of GDP growth at the website of The Economist magazine [11].

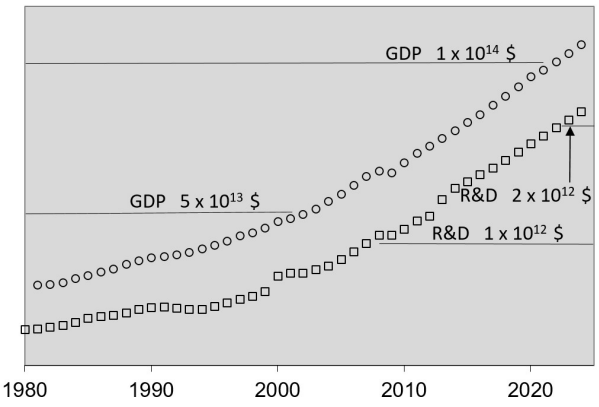
Fig. 1 shows time series for TFs, world Gross Domestic Product (GDP) and world Research and Development Expenditures (R&D). The TFs were available up to 2018 at the time of this study. The GDP and R&D series are accumulated over countries and are presented as if they are known out to 2024. The latest year to be forecasted for TFs in this exercise is 2023. Using the data that were known in January 2019, R&D growth and GDP growth were projected over the further years by means of straight line trends based on the last ten years of known data. But since the data on R&D growth were only available up to 2017 for most countries, now-casts for 2017 (where necessary) and later years for R&D growth are based on straight line trends on the previous ten years of known data. GDP growth was assumed to be known up to 2018.

In this paper, the data on TFs from 1987 to 2018 are modelled and then analysis is carried out over the set of 20 year time windows within that period. For each window, it is not presumed that the levels of TFs beyond the end of each window are known. Estimated values of GDP growth and R&D growth beyond the window are based only on data up to the end of the window.[1]

[1]Forecasts for 2020 and later may be influenced by the effects of the COVID 19 pandemic, that were not included in the models.



(a) Counts of world Total Filings (TFs) at the EPO. TFs are also split by main origins. (EPC states are EPO member countries [4])



(b) Amounts of world Total Gross Domestic Product (GDP, circles, 2005 US dollars at constant prices and purchasing power parity). Amounts of world Total Research and Development Expenditures (R&D, squares, 2010 US dollars at constant prices and purchasing power parity).

Fig. 1: The data time series of macroeconomic variables that are studied.

III. Model fitting

ARIMA and ADL models were used, with programs written in R. The formulations are simple linear models in terms of a few parameters, with normally distributed errors of constant variance within the time series context. More elaborate Box-Jenkins models (like ARIMAX, VAR [12], regression models with ARIMA error structure and dummy variables) were not used. This is because there are practical constraints on the complexity of the model when working with such short time series.

Models were accepted for consideration only if at least one autoregressive or moving average type parameter was included, in order to avoid processing a null model. ADL models without any GDP growth or R&D growth terms at

all were also excluded.

A. ARIMA models

Consider ARIMA(r,d,s) models for TFs with independently distributed homoscedastic error terms (r autoregressive terms, d differences and s moving average terms). After differencing, these are autoregressive moving average ARMA(r,s) models of the differences of TFs ([1]). The model formulation contains an error variance parameter σ^2, that is in practice estimated from the data at the same time as the estimation of the other parameters. Assume n data points in the time series w, after differencing if necessary.

Let $N(f,h)$ indicate a univariate normal distribution with mean f and variance h. The ARMA model is written as follows.

$$w_t = \gamma_1 w_{t-1} + \ldots + \gamma_r w_{t-r} + \epsilon_t + \delta_1 \epsilon_{t-1} + \ldots + \delta_s \epsilon_{t-s} \quad (1)$$

where w_t are the differenced TFs at times t and the innovations are $\epsilon_t \sim N(0,\sigma^2)$. The covariances $\mathrm{Cov}(\epsilon_t, \epsilon_{t-j}) = 0$, for all $j > 0$ and t.

The appropriate degree of differencing d is first determined by the KPSS test [13]. For this, it is assumed that the observed time series can be decomposed into a sum of a random walk, a stationary time series and a deterministic trend. The null hypothesis "variance of the random walk is 0" versus "variance of the random walk is greater than 0" is then tested under constant trend (level stationarity). If this test rejects the null hypothesis, there may be a significant contradiction to stationarity and so the observed time series is transformed by differencing.

At the smallest degree of differencing of TFs that is allowed by the KPSS test, an attempt is made to fit all possible ARMA models up to $r = 4$ and $s = 4$. The residuals of the fitted models are tested for departures from normality with the Shapiro-Wilk test [14]. If this test does not reject the normality assumption, the Ljung-Box test is applied to check for a departure of the residuals from being uncorrelated [15].

A record that is called the "Trace" is constructed of the information criteria values associated with the models that pass these tests. An automatic process selects the model from the set that has the lowest value of the information criterion in the Trace and outputs the estimated parameters. The fitted model is used to construct forecasts w_{t+i} beyond the data set. These forecasts are transformed back to give forecasts for TFs by de-differencing.

B. ADL models

In this type of model, the appropriately differenced TFs are assumed to be generated by autoregressive (AR) terms and by regression terms that are based on the explanatory series. Firstly the explanatory variables are differenced according to the same procedure that was explained for TFs in the previous subsection. The explanatory series are world R&D growth y and world GDP growth z.

The ADL model is written as follows.

$$w_t = \gamma_1 w_{t-1} + \ldots + \gamma_r w_{t-r} + \alpha_0 y_t + \ldots + \alpha_v y_{t-v} + \beta_0 z_t + \ldots + \beta_u z_{t-u} + \epsilon_t \quad (2)$$

where t, w_t and ϵ_t are as in equation (1).

The nomenclature that is used here is ADL($r, d_w; v, d_y; u, d_z$) - for r lags of w after differencing d_w times, v lags of y after differencing d_y times and u lags of z after differencing d_z times. It is allowed under equation (2) to include several such terms for each of the contributing series, r_1, r_2 ..., etc. The nomenclature is then expanded accordingly to ADL($r_1/r_2/..., d_w; v_1/v_2/..., d_y; u_1/u_2/..., d_u$).

An assumed lack of covariance between the three series in ADL is not tested for, as it would be for example in an ARIMAX approach [12]. There is the practical problem that (lagged) future values of the explanatory variables have to be assumed when making forecasts for the TFs. As mentioned in section II, linear projections are made for all forecast years for R&D growth and GDP growth values, based on the last ten years of the available data (or the available data for each window).

In the same way as for the ARIMA models, the Trace is constructed. An automatic process selects the model from the set that has the lowest value of the information criterion, from which the estimated parameters and forecasts are provided.

C. Information criteria

In the following, statistical models are specified in terms of the densities of data that are generated by them. Consider a $p \times 1$ parameter vector θ and a $q \times 1$ parameter vector ϕ. The number of parameters in the data generating model with density $g_0(w|\theta_0)$ is p and the number of parameters to be estimated in the presumed model with density $g_1(w|\phi)$ is q. The models can also contain explanatory variables. The log likelihood corresponding to $g_1(w|\phi)$ is $l(\phi, w)$. g_0 and g_1 can be the same, in which case $g_0(w) = g_1(w|\theta_0)$. The maximum likelihood estimate (MLE) of ϕ is $\hat{\phi}$.

The situation is restricted here to cases where data sets follow multivariate normal distributions. It is assumed that the true model for the data $g_0(w|\theta_0)$ is multivariate normal with design matrix X_0 and covariance matrix V_0. Call this distribution $MN_w(X_0\theta_0, V_0)$, which means a multivariate normal with mean $X_0\theta_0$ and covariance matrix V_0. The estimation model is $MN_w(X_1\phi_1, V_1)$. The symmetric covariance matrices are general, but will be specialised in subsection III.C.3 below for the present context of Box-Jenkins time series models.

The models, including the dimensions of their constituents, are written as follows.

$$g_0(w|\theta_0) = MN_w(X_{0(n\times p)}\theta_{0(p\times 1)}, V_{0(n\times n)})$$

$$g_1(w|\phi) = MN_w(X_{1(n\times q)}\phi_{(q\times 1)}, V_{1(n\times n)})$$

The dimensions q and p can be unequal, with either bigger than the other, and it is not necessary for the models to be nested. Pairwise comparisons are envisaged, but a larger set of candidate models can be used with comparisons between all pairs of models in the set.

1) Akaike's Information Criterion (AIC): This criterion is a goodness of fit statistic that is penalised by the number of parameters in the model [2]. The log likelihood of the estimation model is multiplied by -2, to which is added twice the number of parameters.

$$AIC = -2[l(\phi, w)_{|\phi=\hat{\phi}}] + 2q$$

Under the assumed multivariate normal estimation model $g_1(w|\phi)$, this is as follows.

$$AIC = \frac{n}{2}log(2\pi) + [(\frac{n-q}{2})(1 + log(\hat{\sigma}^2))] + 2q \quad (3)$$

where

$$\hat{\sigma}^2 = \frac{1}{n-q}(w - X_1\hat{\phi})^\top (w - X_1\hat{\phi}),$$

which is an estimate of the variance σ^2 and where $\top$ indicates transposition.

2) Exact Information Criterion (EIC): This criterion compares densities of the MLE between the data generating model and the estimation model. It takes account of the situation where the data generating model $g_0(w)$ differs from the estimation model $g_1(w|\phi)$. The derivation uses a technique for estimator densities (TED) that was developed previously [16] [8] [17]. An approximate approach will be developed here, where the data generating model is assumed to be an abstraction of the observed data themselves.

When models differ, the estimate obtained by maximising the log likelihood of the data under the estimation model $l(\phi, w)$ is technically a quasi maximum likelihood estimate [18]. Nevertheless it will be called a MLE (or $\hat{\phi}$) here and is given by $l'(\phi, w)|_{\phi=\hat{\phi}} = 0$. The space of $\hat{\phi}$ is $\hat{\Phi}$, a subspace of the space Φ of ϕ. The space of the data w is W.

Consider a $(q \times 1)$ vector T,

$$T(\phi, \phi^*, w) = l'(\phi^*, w) - l'(\phi, w) \quad (4)$$

where ϕ^* is fixed at an arbitrary value. Under a simple set of regularity conditions, the exact density for $\hat{\phi}$ is given as follows [16].

$$g(\hat{\phi}) = E_w[|j(\phi, w)|\big|_{\phi=\hat{\phi}}] \times g_{[T(\hat{\phi},\phi^*=\hat{\phi},w)]}(0) \quad (5)$$

where $j(\phi, w) = -l''(\phi, w)$ is the observed information under the estimation model $g_1(w|\phi)$, and $|j(\phi, w)|$ is its determinant.

The second term in (5) represents the value of the density $g_{[T(\hat{\phi},\phi^*,w)]}(t)$, for which $\phi^* = \hat{\phi}$, and hence T is zero in (4). The density is to be derived as the density of a transform $T(\hat{\phi}, \phi^*, w)$ of the data w on the data generating model $g_0(w)$. The term $E_w[|j(\phi, w)|\big|_{\phi=\hat{\phi}}]$ describes a conditional expectation, that is conditional on $\phi = \hat{\phi}$ and is taken with respect to w over $g_1(w|\phi)$. $E_w[|j(\phi, w)|\big|_{\phi=\hat{\phi}}]$ is given by the following expression.

$$\frac{\int_{W_{\widehat{\phi(v)}}} |j(\phi, w(v))||_{\phi=\hat{\phi}}\, g_1(w(v)|_{\phi=\hat{\phi}})\, ||w'(v)||\, dv}{\int_{W_{\widehat{\phi(v)}}} g_1(w(v)|_{\phi=\hat{\phi}})\, ||w'(v)||\, dv}. \quad (6)$$

The integrations in (6) are carried out on a manifold $W_{\widehat{\phi(v)}}$, that runs over an $(n-q)$ dimensional subset of W. The term $||w'(v)||$ indicates the magnitude of the Jacobian from co-ordinates v that index the manifold to w. The evaluation of expectation (6) involves taking conditional expectations over data sets, that can usually be done without evaluation of the integrals in the expression. This is because in practice terms in w can be replaced by $E_w[w|_{\phi=\hat{\phi}}]$, terms in w^2 by $E_w[w^2|_{\phi=\hat{\phi}}]$, etc.

The EIC is based on the idea of Kullback-Leibler Information and is given by

$$EIC = H(\hat{\theta}, \hat{\phi}) = \log\left(\frac{m_A(\hat{\theta})}{m_B(\hat{\phi})}\right) \quad (7)$$

where $m_A(\hat{\theta})$ is the analytic density of the maximum likelihood estimate $\hat{\theta}$ under the data generating model. $m_B(\hat{\phi})$ is the analytic density of the maximum likelihood estimate $\hat{\phi}$ under the estimation model.

Now we will find the densities m_A and m_B for the models in this study. Under the estimation model $g_1(w|\phi)$,

$$l'(\phi^*, w) = X_1^\top {V_1}^{-1}(w - X_1\phi^*),$$
$$l''(\phi, w) = X_1^\top {V_1}^{-1} X_1,$$

and

$$E_w[|j(\phi, w)||_{\phi=\hat{\phi}}] = |X_1^\top {V_1}^{-1} X_1|.$$

The density of $l'(\phi^*, w)$ is $MN(X_1^\top {V_1}^{-1}(X_0\theta_0 - X_1\phi^*), (X_1^\top {V_1}^{-1} V_0 {V_1}^{-1} X_1))$. And so,

$$g_{[T(\hat{\phi},\phi^*,w)]}(t) = \frac{1}{\sqrt{(2\pi)^q |X_1^\top {V_1}^{-1} V_0 {V_1}^{-1} X_1|}} \times \exp\Big[-\frac{1}{2}((t - X_1^\top {V_1}^{-1}(X_0\theta_0 - X_1\phi^*))^\top \big(X_1^\top {V_1}^{-1} V_0 {V_1}^{-1} X_1\big)^{-1} (t - X_1^\top {V_1}^{-1}(X_0\theta_0 - X_1\phi^*)))\Big].$$

Setting T to 0, substituting $\hat{\phi}$ for ϕ^* and multiplying by $E_w[|j(\phi, w)||_{\phi=\hat{\phi}}]$, according to equation (5), gives the density $g(\hat{\phi})$, which is $m_B(\hat{\phi})$.

$$m_B(\hat{\phi}) = \frac{|X_1^\top {V_1}^{-1} X_1|}{\sqrt{(2\pi)^q |X_1^\top {V_1}^{-1} V_0 {V_1}^{-1} X_1|}} \times \exp\Big[\frac{-1}{2}((X_1^\top {V_1}^{-1}(X_1\hat{\phi} - X_0\theta_0))^\top \big(X_1^\top {V_1}^{-1} V_0 {V_1}^{-1} X_1\big)^{-1} (X_1^\top {V_1}^{-1}(X_1\hat{\phi} - X_0\theta_0)))\Big] \quad (8)$$

If the data generating model is also used for estimation, then $\hat{\phi} = \hat{\theta}$ and the density becomes $m_A(\hat{\theta})$. This simplifies to the multivariate normal distribution $MN_{\hat{\theta}}(\theta_0, (X_0^\top {V_0}^{-1} X_0)^{-1})$ for $\hat{\theta}$.

$$m_A(\hat{\theta}) = \sqrt{\frac{|X_0^\top {V_0}^{-1} X_0|}{(2\pi)^p}} \times \exp\left[\frac{-1}{2}\ (\hat{\theta} - \theta_0)^\top\ X_0{}^\top {V_0}^{-1}\ X_0(\hat{\theta} - \theta_0)\right] \quad (9)$$

The evaluation of equations (8) and (9) requires knowledge of the data generating model $g_0(w, \theta_0)$. If the form of this is known, it can be used to estimate $\hat{\theta}_0$ from the data and this is substituted for θ_0.

But here we want to look at alternative estimation models for w without knowing what g_0 is. A modified EIC will be used, termed EIC_w. This measures the estimation model against the observed data, rather than measuring against a specific model that is assumed to be generating the data. It is then analogous to AIC for assessing models against the observed data set that is to be forecasted.

In equations (8) and (9), set $X_0\theta_0$ to w (the observed time series data set of TFs after differencing). Also set V_0 to $\hat{\sigma}^2 I$, where I is the $(n \times n)$ identity matrix and $\hat{\sigma}^2$ is the variance estimated from the residuals of the fit of the estimation model to the data. This usage of the variance estimate from g_1 in the data generating model g_0 is an approximation. So this version of the EIC is no longer "exact".

With these assumptions, and recalling that the determinant $|\hat{\sigma}^2 I| = (\hat{\sigma}^2)^n$, equation (8) is as follows.

$$\begin{aligned} m_B(\hat{\phi}) &= \frac{X_1^\top V_1^{-1} X_1}{\sqrt{(2\pi)^q \hat{\sigma}^{2q} \left| X_1^\top V_1^{-1} V_1^{-1} X_1 \right|}} \\ &\quad \times \exp\left[\frac{-1}{2\hat{\sigma}^2} ((X_1^\top V_1^{-1} (X_1\hat{\phi} - w))^\top \right. \\ &\quad \left. \left(X_1^\top V_1^{-1} V_1^{-1} X_1\right)^{-1} \right. \\ &\quad \left. (X_1^\top V_1^{-1} (X_1\hat{\phi} - w))) \right] \end{aligned} \tag{10}$$

A version of (7) is used where $m_A(\hat{\theta})$ refers to a true model given by the data and $m_B(\hat{\phi})$ is given by (10). For $m_A(\hat{\theta})$ under (9), consider $\hat{w}$ to have a multivariate normal distribution with mean as w. That is $MN(X_0\theta_0, \hat{\sigma}^2 I)$ with $\theta_0 = w_{(n\times 1)}$, $X_0 = I_{(n\times n)}$, where I is the $n \times n$ identity matrix. This model is not realisable in general but, when $\hat{w} = w$, (9) shows that

$$m_A(\hat{\theta}) = (2\pi\hat{\sigma}^2)^{\frac{-n}{2}} \quad .$$

Equation (7) then gives

$$\begin{aligned} EIC_w &= H(w, \hat{\phi}) \\ &= \frac{-n}{2} \log(2\pi\hat{\sigma}^2) - \log(m_B(\hat{\phi})), \end{aligned} \tag{11}$$

The first term in this equation differs between estimation models because the $\hat{\sigma}^2$ terms themselves differ under the various models $g_1(w|\phi)$.

The design matrix X_1 is taken from the linear formulation of the parameters in equations (1) or (2), with w_{t-1} to w_{t-r} substituted by their estimates $\hat{w}_{t-1}$ to $\hat{w}_{t-r}$ after the model has been fitted. The terms ϵ_t to ϵ_{t+1-s} are estimated by $(w_t - \hat{w}_t)$ to $(w_{t+1-s} - \hat{w}_{t+1-s})$.

For an ARMA estimation model, after differencing of the data, the design matrix X_1 has n rows and $r + s$ columns that represent the γ and δ parameters that are to be estimated in equation (1). For an ADL model, after differencing of the dependent and explanatory data, the design matrix X_1 has n rows and $r+(v+1)+(u+1)$ columns. These rows represent the autoregressive terms (the γ terms) and the regression terms that depend on the explanatory series (the α and β terms) in equation (2).

3) Covariance structure of the time series models for EIC: In order to use equations (10) and (11), an explicit form for the covariance matrix V_1 is required that incorporates the time series structure for the fitted model g_1. V_1 is composed of covariance terms $Cov(w_t, w_{t-k})$ for the differenced observations at the various lags k.

For ARMA models, the terms of V_1 are built up according to a formulation by [1] (Chapter 3.3). The models considered here are limited to having AR or MA terms up to order 4.

Using estimates of the parameters from equation (1), $\hat{\gamma}_i$ is the i-th AR parameter (or 0 if this does not exist), $\hat{\delta}_i$ is the i-th MA parameter (or 0 if this does not exist) and $\hat{\sigma}^2$ is the error variance.

Terms S_0 to S_4 are set up as follows.

$$\begin{aligned} S_0 &= 1 \\ S_1 &= \hat{\delta}_1 + \hat{\gamma}_1 \\ S_2 &= \hat{\delta}_2 + \hat{\gamma}_2 + (\hat{\delta}_1 \times \hat{\gamma}_1) + (\hat{\gamma}_1 \times \hat{\gamma}_1) \\ S_3 &= \hat{\delta}_3 + (\hat{\gamma}_1 \times S_2) + (\hat{\gamma}_2 \times S_1) + \hat{\gamma}_3 \\ S_4 &= \hat{\delta}_4 + (\hat{\gamma}_1 \times S_3) + (\hat{\gamma}_2 \times S_2) + (\hat{\gamma}_3 \times S_1) + \hat{\gamma}_4 \end{aligned}$$

Then, for $i > 4$, S_i terms are given sequentially as

$$S_i = \sum_{0<k<5} \hat{\gamma}_k \times S_{i-k}.$$

The covariance terms are built from the S_i terms.

$$\begin{aligned} Cov(w_t, w_{t-k}) &= Cov(w_t, w_{t+k}) \\ &= \hat{\sigma}^2 \times \sum_{i=0}^{\infty} S_i \times S_{i+k} \end{aligned} \tag{12}$$

It is not possible to calculate this infinite series in practice. So an experiment was done that found stable results by summing i in equation (12) from 0 to 30. The EIC results that are reported here for ARIMA and ADL models use this approach.

For ADL models, a conditional approach is taken. The scheme to build up V_1 is also based on the above description for ARMA, by disregarding the $\hat{\delta}_i$ terms but retaining the non-zero $\hat{\gamma}_i$ terms for the autoregressive lags of w that are active in the model.

IV. Results

A. Fits to the data set

The data were analysed by using ARIMA and ADL models. The selection of the models was done from the minimum values of either AIC by (3) or $|EIC_w|$, which is the absolute value of EIC_w from (11). In all cases the degrees of differencing that the KPSS tests identified were 1 for TFs, 0 for R&D growth and 0 for GDP growth. In conformity with the stationarity assumption, the absolute values of the estimates of the autoregressive and moving average parameters were usually less than 1. This was the case for all the selected models.

Table 1 shows results for fits to the TFs data from 1987 to 2018. ARIMA(2,1,1) was selected by both information criteria. This model forecasts a modest progression of 1.1 per cent compound annual growth in TFs from 2019 to 2023.

TABLE 1: ARIMA AND ADL MODELS FITTED TO TOTAL FILINGS DATA FROM 1987 TO 2018. THE SAME ARIMA AND ADL MODELS WERE SELECTED IN THESE CASES BY BOTH INFORMATION CRITERIA.

	Total Filings (TFs)	
Year	ARIMA model (2,1,1)	ADL model (3,1;2,0;1/3,0)
Actual 2019	323,525	323,525
Forecast 2019	321,050	319,773
Forecast 2020	326,294	326,129
Forecast 2021	330,555	333,544
Forecast 2022	333,046	340,419
Forecast 2023	335,054	347,171
SE of Forecast 2019	7,619	3,797
AIC	650.9	550.0
EICw	-264.6	-219.7
Fitted parameters	$\hat{\gamma}_1$ AR lag 1: 0.0169 $\hat{\gamma}_2$ AR lag 2: 0.4614 $\hat{\delta}_1$ MA lag 1: 0.7771	$\hat{\gamma}_3$ w lag 3: -0.2512 $\hat{\alpha}_2$ y lag 2: 80,221 $\hat{\beta}_1$ z lag 1: 372,602 $\hat{\beta}_3$ z lag 3: -164,171

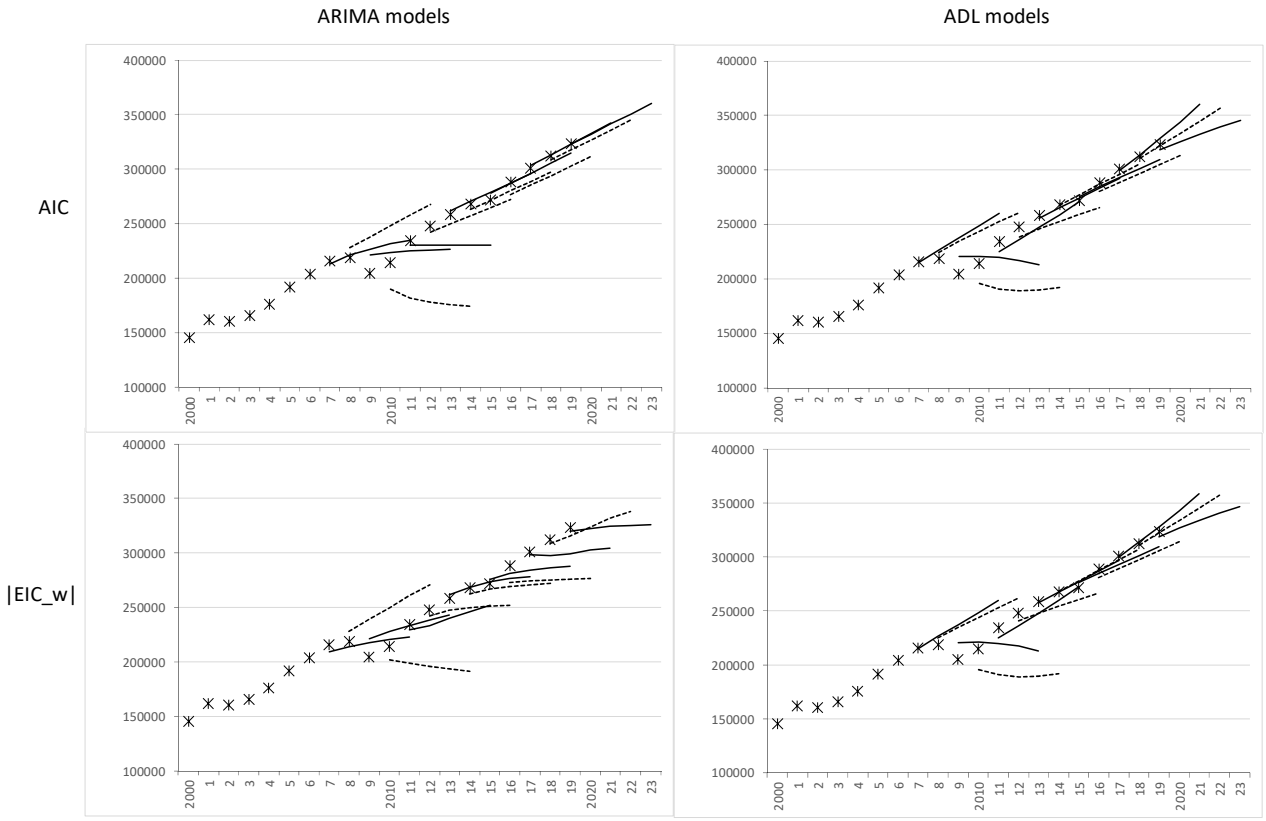


Fig. 2: Total Filings forecasts by models fitted to 20 year windows. Results are shown from 1987-2006 (Window 1) to 1999-2018 (Window 13). A comparison of ARIMA and ADL models, with either AIC or $|EIC_w|$ as the information criterion that was used for the model selections.

ADL(3,1;2,0;1/3,0) was selected by both information criteria, with a compound annual growth of 2.1 per cent on the same period. The standard error of the first forecast year is lower for ADL than it is for ARIMA. This suggests that the ADL model fits the data better. We might expect better forecasts via ADL than via ARIMA.

Next, a set of 13 successive 20 year windows of TFs data were taken, from 1987-2006 to 1999-2018. ARIMA and ADL models were selected by minimising AIC or $|EIC_w|$, after fitting the models to the data for each window. Fig. 2 shows the forecasts by the four combinations. Results in terms of models that were selected and forecast statistics are shown in Table 2 for ARIMA models and in Table 3 for ADL models. The statistics shown are the means and median absolute percentage errors (MAPE and Median APE), the mean percentage errors (MPE) of the forecasts and the standard errors of MPE.

Tables 2 and 3 suggest that $|EIC_w|$ often selects models with more parameters than AIC does. Usually the forecasting accuracies are better for ADL models than for ARIMA models. This is seen with the values derived from both AIC and $|EIC_w|$. For the ARIMA model, the forecasting accuracy is in most cases better for the models selected using AIC rather than $|EIC_w|$. But, for one year ahead, $|EIC_w|$ is marginally better for the MAPE statistic. For the ADL model, on the whole the forecasting accuracy is in most cases slightly better for the models selected using $|EIC_w|$ rather than AIC. But, for five years ahead, AIC is marginally better for the Median APE statistic. For the ADL model, the differences in the forecasts themselves between the two criteria are slight and are hardly visible in Fig. 2.

B. Comparisons of information criteria

Table 4 shows additional metrics for the selected models, and those closest to them, for several examples. $\log[m_A(\hat{\theta})]$ and $\log[m_B(\hat{\phi})]$ are shown with the EIC_w and AIC values. The examples are the whole data set (1987-2018) and four of the 20 year windows (1: 1987-2006, 4:1990-2009, 7:1993-2012 and 13:1999-2018). These periods represent the beginning, middle and end of the data, as well as the data set ending with the 2009 recession. In each case, results for up to five of the closest models including the selected model

TABLE 2: ARIMA MODELS FITTED TO 20 YEAR WINDOWS. THE UPPER PART OF THE TABLE SHOWS FORECASTING STATISTICS. THE LOWER PART OF THE TABLE SHOWS THE MODELS SELECTED FOR EACH WINDOW.

	ARIMA models Selection criterion: AIC				ARIMA models Selection criterion: |EICw|			
	Percentage error				Percentage error			
Years ahead	MAPE	Median APE	MPE	S.E.(MPE)	MAPE	Median APE	MPE	S.E.(MPE)
1	3.2	1.7	-0.5	4.6	3.0	2.0	-0.6	3.8
2	5.3	2.6	-1.5	8.8	5.9	4.5	-2.0	7.9
3	7.9	4.1	-2.6	11.4	8.2	6.7	-3.9	9.8
4	8.9	7.2	-5.3	11.9	9.6	8.2	-6.7	9.7
5	9.5	5.7	-7.7	12.3	11.1	9.1	-9.0	9.8

Window	Start/End	Model	AIC	Start/End	Model	EICw
1	1987/2006	(2,1,1)	393.8	1987/2006	(2,1,2)	-136.7
2	1988/2007	(2,1,2)	393.5	1988/2007	(3,1,2)	-96.8
3	1989/2008	(1,1,0)	391.2	1989/2008	(2,1,2)	-149.2
4	1990/2009	(1,1,1)	395.4	1990/2009	(3,1,2)	-110.0
5	1991/2010	(0,1,1)	401.0	1991/2010	(2,1,2)	-135.1
6	1992/2011	(1,1,2)	401.1	1992/2011	(2,1,1)	-140.6
7	1993/2012	(1,1,2)	401.7	1993/2012	(2,1,1)	-138.3
8	1994/2013	(1,1,2)	401.4	1994/2013	(2,1,1)	-143.1
9	1995/2014	(1,1,2)	401.8	1995/2014	(2,1,1)	-148.4
10	1996/2015	(1,1,2)	402.5	1996/2015	(2,1,2)	-105.1
11	1997/2016	(1,1,2)	404.7	1997/2016	(3,1,2)	-133.3
12	1998/2017	(1,1,2)	404.4	1998/2017	(4,1,0)	-157.6
13	1999/2018	(1,1,2)	404.5	1999/2018	(2,1,1)	-150.9

TABLE 3: ADL MODELS FITTED TO 20 YEAR WINDOWS. THE UPPER PART OF THE TABLE SHOWS FORECASTING STATISTICS. THE LOWER PART OF THE TABLE SHOWS THE MODELS SELECTED FOR EACH WINDOW.

	ADL models Selection criterion: AIC				ADL models Selection criterion: |EICw|			
	Percentage error				Percentage error			
Years ahead	MAPE	Median APE	MPE	S.E.(MPE)	MAPE	Median APE	MPE	S.E.(MPE)
1	2.7	1.7	-0.8	3.9	2.6	1.7	-0.7	3.8
2	4.9	3.2	-0.9	7.6	4.8	3.2	-0.8	7.6
3	7.3	4.8	-1.4	10.6	7.3	4.7	-1.3	10.5
4	8.3	5.1	-3.5	11.2	8.1	4.8	-3.3	11.2
5	8.8	5.3	-5.2	11.7	8.6	5.4	-5.0	11.8

Window	Start/End	Model	AIC	Start/End	Model	EICw
1	1987/2006	(1,1;0/1/2/3,0;0/1/2/3,0)	302.7	1987/2006	(1/3,1;0/1/2/3,0;0/1/2/3,0)	-22.6
2	1988/2007	(1/2,1;0/1/2/3,0;0/1/2,0)	297.2	1988/2007	(1/2,1;0/1/2/3,0;0/1/2/3,0)	-14.5
3	1989/2008	(1,1;1,0;1/2/3,0)	293.3	1989/2008	(1,1;0/1,0;1/2/3,0)	-12.6
4	1990/2009	(1,1;0/1,0;1/2/3,0)	294.0	1990/2009	(1,1;0/1,0;1/2/3,0)	12.4
5	1991/2010	(3,1;0/1/2/3,0;0/1/2/3,0)	305.8	1991/2010	(1/2/3,1;0/1/2/3,0;0/1/2/3,0)	-39.5
6	1992/2011	(3,1;1/2,0;1/3,0)	308.9	1992/2011	(1/2/3,1;0/1/2/3,0;0/1/2/3,0)	-32.6
7	1993/2012	(3,1;2,0;1/3,0)	308.4	1993/2012	(2/3,1;0/1/2/3,0;1/2/3,0)	-33.3
8	1994/2013	(3,1;1/2,0;1/3,0)	309.9	1994/2013	(1/2/3,1;0/1/2/3,0;0/1/2/3,0)	-2.0
9	1995/2014	(2/3,1;1/2,0;1/3,0)	301.5	1995/2014	(1/2/3,1;0/1/2/3,0;0/1/2/3,0)	-6.4
10	1996/2015	(2/3,1;1/2/3,0;1/3,0)	301.9	1996/2015	(2/3,1;0/1/2/3,0;0/1/3,0)	20.0
11	1997/2016	(2/3,1;0/2/3,0;0/1,0)	318.6	1997/2016	(2/3,1;0/2/3,0;0/1/3,0)	-37.5
12	1998/2017	(3,1; ;1,0)	318.3	1998/2017	(3,1; ;1,0)	-137.7
13	1999/2018	(3,1;0,0 ;1,0)	312.0	1999/2018	(1/3,1;0/1,0 ;0/1,0)	-80.0

are shown. Where there are less than five, this indicates that there were fewer than five candidate models allowed by the normality [14] and auto-correlation tests [15] at the degrees of differencing that were specified by the KPSS tests [13].

Often both AIC and $|EIC_w|$ select the same set of best fitting models for a combination of technique and data set. There is more variation between windows for the selected sets of five models for $|EIC_w|$ than there is for AIC. When using $|EIC_w|$, the values of $\log[m_B(\hat{\phi})]$ are closer to zero under the ARIMA models than under the ADL models. This means that the density under the estimation model (that is represented by the log density $\log[m_B]$) is lower for the ADL models than it is for the ARIMA models. It is not surprising, because ADL models typically contain more parameters than the corresponding ARIMA models.

In order to examine the extent that forecasting ability

TABLE 4: ARIMA AND ADL MODELS FITTED TO ALL DATA AND TO FOUR OF THE THIRTEEN WINDOWS. FOR EACH WINDOW, THE CRITERIA AIC AND $|EIC_W|$ FOR UP TO FIVE OF THE BEST MODELS ARE SHOWN.

	ARIMA models						ADL models					
	Selection criterion: AIC		Selection criterion: |EICw|				Selection criterion: AIC		Selection criterion: |EICw|			
All data: 1987-2018	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw
1	(0,1,1)	653.2	(0,1,1)	-279.6	**-2.6**	-277.0	(3,1;2,0;1/3,0)	**550.0**	(3,1;2,0;1/3,0)	-257.0	-37.4	**-219.7**
2	(2,1,1)	**650.9**	(2,1,1)	**-276.7**	-12.1	**-264.6**						
Window 1 1987-2006	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw
1	(0,1,1)	396.7	(0,1,1)	-158.1	**-2.1**	-156.0	(1/2/3,1;0/1/2/3,0;0/1/2/3,0)	306.3	(1/2/3,1;0/1/2/3,0;0/1/2/3,0)	-130.5	**-106.3**	-24.1
2	(2,1,1)	**393.8**	(1,1,2)	**-149.1**	-8.0	-141.1	(1/3,1;0/1/2/3,0;0/1/2/3,0)	304.3	(1/3,1;0/1/2/3,0;0/1/2/3,0)	-130.5	-107.9	**-22.5**
3	(2,1,2)	394.4	(2,1,2)	-149.5	-12.9	**-136.7**	(1,1;0/1/2/3,0;0/1/2/3,0)	**302.7**	(1,1;0/1/2/3,0;0/1/2/3,0)	**-129.8**	-106.6	-23.2
Window 4 1990-2009	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw
1	(0,1,1)	396.2	(1,1,1)	-156.3	**-4.5**	-151.8	(1/2/3,1;0/1/2/3,0;0/1/2/3,0)	300.3	(1/2/3,1;0/1/2/3,0;0/1/2/3,0)	**-129.7**	-176.3	46.6
2	(1,1,1)	**395.4**	(2,1,1)	**-155.6**	-11.8	-143.8	(1/3,1;0/1/2/3,0;0/1/2/3,0)	298.4	(1/3,1;0/1/2/3,0;0/1/2/3,0)	-129.7	-179.8	50.0
3	(1,1,2)	396.7	(4,1,1)	-156.5	-8.3	-148.2	(1/3,1;0/1/2,0;0/1/2/3,0)	296.7	(1/3,1;0/1/2,0;0/1/2/3,0)	-129.8	-193.6	63.8
4	(3,1,1)	398.1	(1,1,2)	-155.7	-15.7	-140.0	(1,1;0/1,0;1/2/3,0)	**294.0**	(1,1;0/1,0;1/2/3,0)	-131.9	-143.8	**11.8**
5	(2,1,2)	398.1	(3,1,2)	-150.5	-40.5	**-110.0**	(1,1;1,0;1/2/3,0)	299.1	(1,1;1,0;1/2/3,0)	-136.6	**-93.5**	-43.1
Window 7 1993-2012	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw
1	(1,1,0)	403.8	(1,1,0)	-160.9	-0.2	-160.8	(2/3,1;0/1/2/3,0;1/3,0)	309.7	(2/3,1;0/1/2/3,0;1/2/3,0)	-134.0	-100.6	**-33.4**
2	(0,1,1)	402.8	(2,1,0)	-160.8	**0.1**	-160.9	(2/3,1;0/1/2,0;1/3,0)	308.6	(2/3,1;0/1/2,0;1/3,0)	**-133.7**	-89.8	-43.9
3	(1,1,2)	**401.7**	(1,1,1)	-160.3	-4.9	-155.4	(3,1;0/1/2,0;1/3,0)	308.6	(2/3,1;0/1/2,0;1/3,0)	-134.9	-69.8	-65.1
4	(2,1,2)	402.6	(2,1,1)	**-159.4**	-21.1	**-138.3**	(3,1;1/2,0;1/3,0)	**308.4**	(3,1;0/1/2,0;1/3,0)	-136.9	-60.8	-76.1
5	(3,1,2)	404.0	(3,1,1)	-160.2	-15.3	-144.9	(3,1;2,0;1/3,0)	311.5	(3,1;1/2,0;1/3,0)	-137.5	**-52.5**	-85.1
Window 13 1999-2018	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw	Candidate models	AIC	Candidate models	log[mA]	log[mB]	EICw
1	(1,1,0)	405.7	(1,1,0)	-160.4	**-0.4**	-160.0	(1/3,1;0/1,0;0/1,0)	315.6	(1/3,1;0/1,0;0/1,0)	**-144.3**	-64.3	**-80.0**
2	(2,1,0)	407.6	(4,1,0)	**-158.1**	-0.5	-157.6	(3,1;0/1,0;0/1,0)	314.1	(3,1;0/1,0;0/1,0)	-144.6	-56.3	-88.3
3	(1,1,1)	407.5	(0,1,1)	-160.7	-1.5	-159.1	(3,1;0,0;1,0)	**312.0**	(3,1;0,0;1,0)	-145.5	-26.8	-118.7
4	(1,1,2)	**404.5**	(2,1,1)	-160.1	-9.2	**-150.9**	(3,1;,;1,0)	315.0	(3,1;,;1,0)	-148.0	**-10.9**	-137.1
5	(2,1,2)	406.3	(0,1,2)	-160.4	-1.5	-158.8						

correlates with the measured information criteria over the same set of data windows, Fig. 3 shows the timecourses for forecasting accuracy (PE, the absolute percentage error of the forecast at one year ahead) and for information criteria, over the four combinations. The serial correlation coefficients across time are shown beneath the charts.

Using significance tests for the correlation coefficients [19], cut-offs of $P < 0.05$ give indications of significant discrepancies from a null hypothesis of no correlation. The realised values of AIC show significant positive serial correlation over the timecourse for both ARIMA and ADL. On the other hand, for $|EIC_w|$ there is no significant serial correlation over the timecourse for either ARIMA or for ADL. For PE, there are significant positive serial correlations for ADL models, but not for ARIMA models.

Fig. 4 plots the absolute percentage errors against the information criterion values, over the four combinations. There are significant negative correlations between AIC and PE for both ARIMA and ADL models. But there are no significant correlations between $|EIC_w|$ and PE for either the ARIMA or the ADL models.

These significance tests for correlations are only indicative. Firstly, the power of the tests is low because only 13 data points are used in each case. Secondly, it is not unreasonable to have significant serial correlations. In each window there are 19 values that are identical to, and only one value that is different from, the immediately neighbouring windows.

V. DISCUSSION

An advantage of the information criterion AIC is that it is rather simple to calculate. The application of equation (3) requires only the estimated residual variance $\hat{\sigma}^2$, the number of parameters q and the number of data points n.

For the information criterion $|EIC_w|$, the term $m_B(\hat{\phi})$ in equation (10) takes some effort to calculate. But this can be programed readily from the data together with the model specification and parameter estimates. As has been seen in the results tables, the forecast selections by $|EIC_w|$ provide additional information to those given by AIC. This may be helpful to make improved business decisions.

Other information criteria can be used to select the models [2]. For example, the Bayesian Information Criterion (BIC) is a variant of AIC with a different penalty term. When using BIC with these data, we found no differences in the model selections to those made by using AIC.

$|EIC_w|$ has some characteristics that differ from AIC and other closely related information criteria. AIC is essentially equivalent to the log-likelihood of the model after fitting to the data, but with an additional penalty term due to the length of the vector of parameters $\hat{\phi}$. $|EIC_w|$ measures the difference between true and estimated models of the logarithm of the densities of the estimated parameters at the MLE. The presumption is that, if these densities at the same central point are similar, then the densities as a whole will be similar. This means that the estimated parameters should be trustworthy. The penalty for additional parameters in $|EIC_w|$ is implicit, because there is a tendency for estimator densities from longer parameter vectors to be lower than for shorter ones.

The derivation of AIC follows an asymptotic argument regarding the distribution of the expected value of estimated Kullback-Liebler information under repeated sampling [2]. This is to some degree analogous to the asymptotic derivation of the normal distribution for parameter estimates under the central limit theorem. But the derivation of EIC_w, in

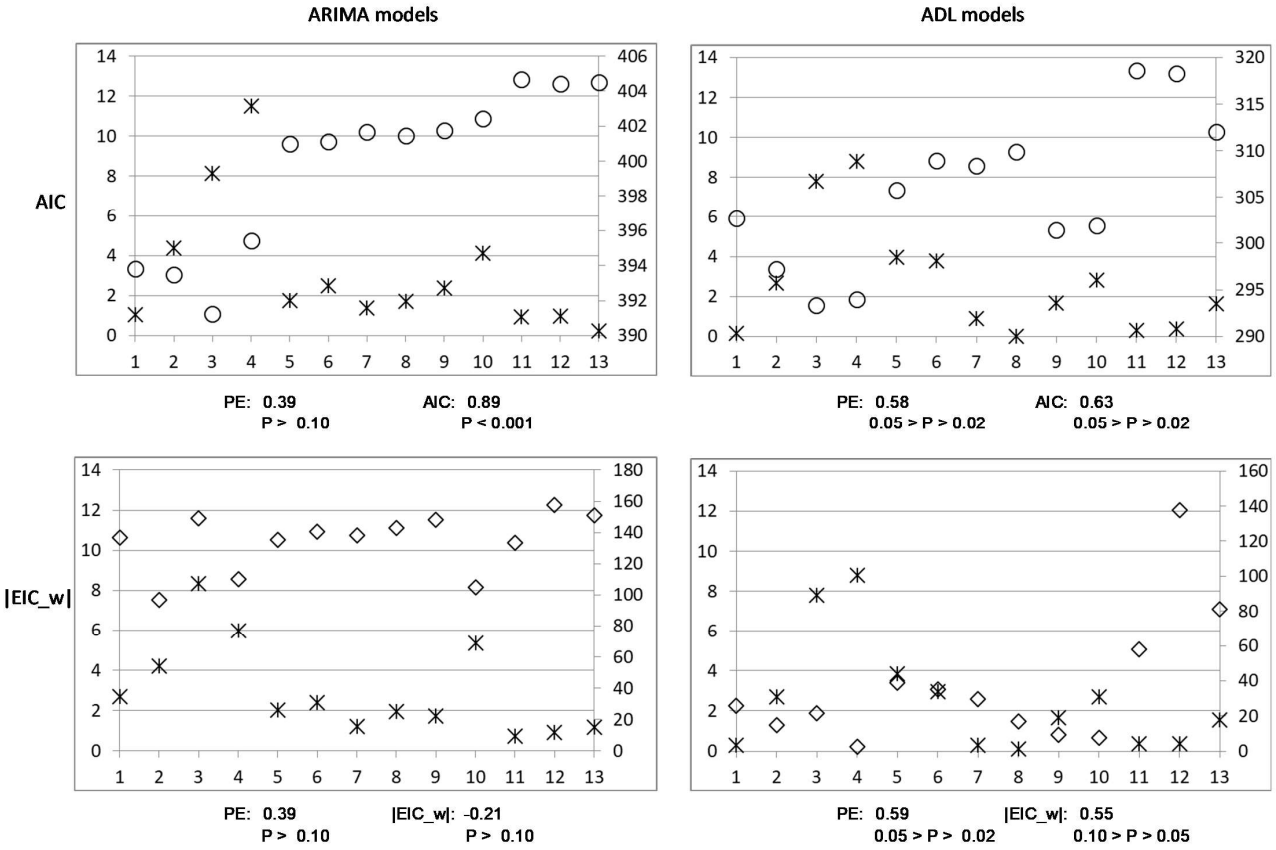


Fig. 3: One year ahead forecasts for the 20 windows. Timecourses for absolute percentage forecasting error one year ahead (PE: stars, scale on the left side of each chart) and for information criteria (AIC: circles, $|EIC_w|$: diamonds, scale on the right side of each chart). Results are shown for the selected models for the windows from 1987-2006 (Window 1) to 1999-2018 (Window 13), on the horizontal scale. Under each chart the serial correlation coefficient for PE and the serial correlation coefficient for the information criterion are shown, with P values for two sided significance tests of no serial correlation [19].

subsection III.C.2 above, follows an exact argument up to equation (9). An approximation is applied only at the last stage of asserting the data set as the true data generating model, in order to give EIC_w in equations (10) and (11).

When applying the technique for estimator densities that underlies EIC in other situations, a practical difficulty can be lack of tractability in calculating the distribution of the term T in equation (5). This is caused by the form of the error distribution that is assumed for the data, even with some simple error distributions (EG see [20]). But for the normal error distributions that were studied here, there is no such difficulty if the requirements for the handling of the time series structure are taken into account.

It was chosen to use the minimum absolute value $|EIC_w|$ as a model selection criterion. The assumption is that the model whose estimator density lies as close as possible to the true one is the best one to take. The accuracy of the MLE estimate is directly targeted. This is important because the MLE is being used to construct point forecasts outside the data set.

To give some further insight into model selections using the information criteria, an experiment was done to compare simple linear models (straight line model vs quadratic model), on a design with six data points and independent homoscedastic normal errors. (The details are not shown here). In this setup, an F test is often carried out to determine the appropriate model [21]. Its purpose is to prove the statistical significance of a candidate model with more parameters before accepting it rather than a simpler model. The information criteria may be able to make a more balanced choice between models, because statistical significance does not have to be established.

In this experiment, when simulated data sets were generated from a quadratic model, $|EIC_w|$ provided more correct model assignments of the quadratic model than AIC did, and both criteria performed better than the F tests did. But when simulated data sets were generated from a straight line model, AIC provided more correct model assignments of

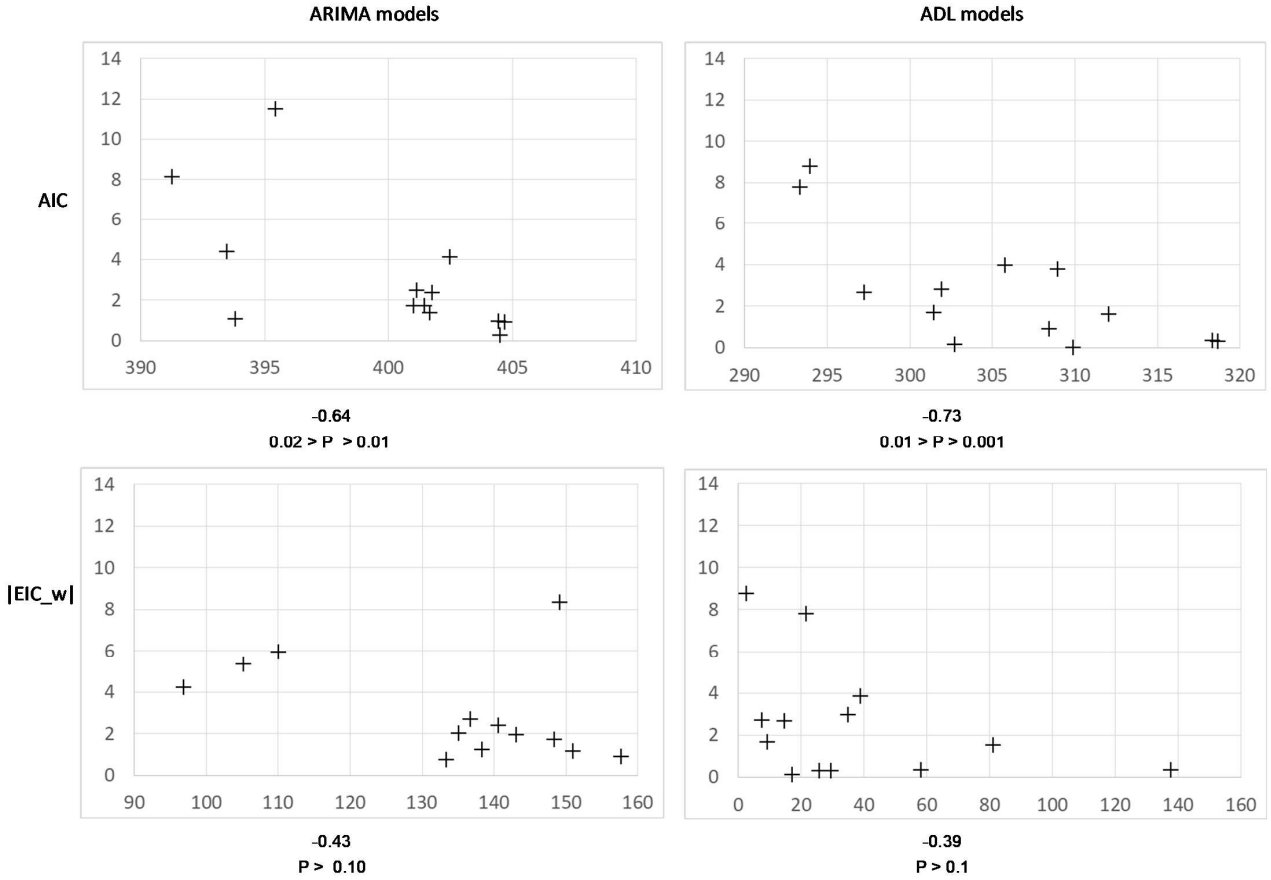


Fig. 4: One year ahead forecasts for the 20 windows. Scatter plots of absolute percentage forecasting errors (PE, vertical scales) with information criteria (horizontal scales), for the selected models. Results are shown from 1987-2006 (Window 1) to 1999-2018 (Window 13). The number under each chart is the correlation coefficient between PE and the information criterion, with the P value for a two sided significance test of no serial correlation [19].

the straight line model than $|EIC_w|$ did, and the F tests performed better than both of the criteria.

For the approach to model discrimination of this paper, that used the densities $m_A(\hat{\theta})$ and $m_B(\hat{\phi})$, other ways could be used to make the selections. For example, if interval forecasting had been the goal, then estimator variances or other higher moments could be considered. In that case, more difficulties would be found with the calculations, because there is no obvious way to directly calculate the variance of $m_B(\hat{\phi})$ as $g(\hat{\phi})$ in equation (5). This is a formula for a probability density function whose second moment might need to be found by integration.

Since an estimation model gives a worse fit than the observed data do to themselves, in most cases this study generated negative EIC_w values. (See Tables 2 and 3.) The exceptions were under ADL for Window 4 (for the years 1999 to 2009) and for Window 10 (for the years 1996 to 2015). In these cases, the model with the lowest value of $|EIC_w|$ was given by a positive EIC_w. It can be noted that both these windows ended in years where there were small corrections in the upward progress of EPO Total Filings (see Fig.1).

The exercise to find free fitting models to the whole TFs data set (for the years 1987-2008) gives a lower standard error of the first year forecast under the ADL model than under the ARIMA model. (See Table 1.) So one could recommend usage of the ADL model, with the ARIMA model as only a confirmatory tool to look at short term effects. Here AIC and $|EIC_w|$ led to the same models being selected under ARIMA and ADL. This brings up the possibility that differences in selected models for the 20 year windows exercise might be related to the shorter data sets that were considered.

It does not seem advisable to apply the specific ADL model of Table 1 without further checking as new data points appear. It is more appropriate to apply the models to a variety of lengths of the historical data. Then, using both information criteria, see whether the forecasts agree well enough under all permutations to give confidence in the results. This could be reconfirmed each year that the forecasting is done.

With the 20 year rolling windows and the results that are presented in Tables 2 and 3, the best results were

obtained using $|EIC_w|$ as the selection criterion under the ADL model. It should be noted that, for ADL models using $|EIC_w|$, the conditional approach assumes that no contribution to the covariances is made by the regression terms relating y to w and z to w in equation (2). An experiment was done to accommodate the regression terms into the covariance expression, by adding in terms from the straight line regressions that had been used to project the explanatory variables. This led to no change in the models that were selected by the modified $|EIC_w|$ values that were created. Therefore it can be considered that the conditional approach to ADL models is adequate for these data.

The results in Fig. 2 can be compared with forecast projections for TFs using windows under a dynamic log-linear regression model setup (see Fig. 3 in [7]). That approach tends to forecast stronger growth than do the Box-Jenkins models that were studied here.

The patent filings series is a macroeconomic series that is affected by major international events. The clearest event in Fig. 2 was the great recession in 2009. So stories can be told about forecasting failures for some windows. All techniques forecasted drops in the years after the 1990-2009 Window 4. For 2010 to 2014, a steady decline was forecasted by ARIMA and a decline followed by a partial recovery was forecasted by ADL. But in fact there was a strong and immediate recovery in TFs from 2010 onwards.

For the later windows (1992-2011 Window 6 to 1999-2018 Window 13), Table 2 shows that AIC under the ARIMA model consistently selects the ARIMA(1,1,2) model. This model may be dominated by its MA terms, so that forecasts are selected that are essentially straight lines. These happen to fit the out-turns rather well compared to the ARIMA models selected by $|EIC_w|$, which have at least two AR terms and tend to prescribe positive but plateauing forecast trends. Perhaps this relates to the higher $|EIC_w|$ values found for ARIMA models than were found for ADL models.

It may also be noted that the AIC values for ARIMA models are higher than for ADL models. This is consistent with a worse fit by ARIMA than by ADL. For ADL itself, $|EIC_w|$ has slightly better forecasting accuracy than AIC. But all the accuracy statistics may be influenced by the bad forecasts from Window 3 (1989-2008) and Window 4 (1990-2009).

Fig. 4 explored the correlations of AIC values and $|EIC_w|$ values with one year ahead absolute forecasting error PE values, over the windows. Surprisingly enough, a significant negative correlation was found between AIC and PE both for ARIMA and ADL models. This means that larger values of AIC correlate with greater degrees of forecasting accuracy. Prima facie this is an argument against using minimum values of AIC as a way to selecting a model for forecasting. Reasons for this await further research. But it could be due to sampling effects because of the high degree of data overlap between successive windows. The correlations of AIC and absolute forecasting error at two to five years beyond the data sets gave similar effects to the one year effects in most cases.

The information criteria that were used for model selection reflect aspects of the goodness-of-fit of the model to the data set rather than forecasting ability per se. So it is unsurprising that positive relationships between the criteria and PE were not established. But this study suggests that $|EIC_w|$ is a forecasting assessment tool that is at least as good as AIC.

ACKNOWLEDGMENT

The authors would like to thank the European Patent Office for providing the patent filings data. The forecasts that are reported are not to be considered as official forecasts.

REFERENCES

[1] P. Brockwell and R. Davis, "Time Series: Theory and Methods," 2^{nd} edition, *Springer*, 2006.
[2] K. Burnham and D. Anderson, "Model Selection and Multimodel Inference," 2^{nd} Edition, *Springer*, 2002.
[3] R. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The Forecast Package for R," *Journal of Statistical Software*, vol. 27, no. 3, pp1-22, 2008.
[4] P. Hingley and M. Nicolas, "Background," In "Forecasting Innovations," *Springer*, pp1-8, 2006.
[5] P. Hingley and M. Nicolas, "Methods for Forecasting Numbers of Patent Applications at the European Patent Office," *World Patent Information*, vol. 26, no. 3, pp191-204, 2004.
[6] European Patent Office, "Patent Filings Survey 2018 - Intentions of Applicants regarding Patent Applications at the European Patent Office and Other Offices," 2019. https://www.epo.org/service-support/contact-us/surveys/patent-filings/archive.html
[7] P. Hingley and W. Park, "A Dynamic Log-linear Regression Model to Forecast Numbers of Future Filings at the European Patent Office," *World Patent Information*, vol. 42, pp19-27, 2015.
[8] P. Hingley, "Distributions of Maximum Likelihood Estimators and Model Comparisons," *Current Themes in Engineering Science 2007*, pp111-122, 2008.
[9] Organisation for Economic Cooperation and Development, "Main Science and Technology Indicators," 2021. http://www.oecd.org/sti/msti.htm
[10] World Bank, "World Development Indicators," 2021. http://data.worldbank.org/data-catalog/world-development-indicators
[11] The Economist Markets and data, 2021. https://www.economist.com/markets-data
[12] G. Dikta, "Time Series Methods to Forecast Patent Filings," In "Forecasting Innovations", *Springer*, pp95-124, 2006.
[13] D. Kwiatkowski, P. Phillips, P. Schmidt and Y. Shin, "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root," *Journal of Econometrics*, vol. 54, nos. 1-3, pp159-178, 1992.
[14] S. Shapiro and M. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, nos. 3-4, pp591-611, 1965.
[15] G. Ljung and G. Box, "On a Measure of a Lack of Fit in Time Series Models," *Biometrika*, vol. 65, no. 2, pp297-303, 1978.
[16] P. Hingley, "Analytic Estimator Densities for Common Parameters under Misspecified Models," In "Theory and applications of recent robust methods", *Statistics for Industry and Technology*, *Birkhauser*, pp119-130, 2004.
[17] P. Hingley, "Applications and Extensions of a Technique for Estimator Densities," *IAENG International Journal of Applied Mathematics*, vol. 39, no.1, pp61-70, 2009.
[18] H. White, "Estimation, Inference and Specification Analysis,", *Cambridge*, 1994.
[19] J. Murdoch and J. Barnes, "Statistical Tables for Science, Engineering, Management and Business Studies," 2^{nd} Edition, *Macmillan*, 1970.
[20] P. Hingley, "Estimating the Mean of a Small Sample under the Two Parameter Lognormal Distribution," In "Mathematical Methods and Models in Biosciences, Texts in Biomathematics," *Biomath Forum*, pp100-121, 2018.
[21] N. Draper and H. Smith, "Applied Regression Analysis," 2^{nd} Edition, *Wiley*, 1981.

Peter Hingley worked at the European Patent Office in Germany and retired in 2020. His main research interests are in statistical modelling and robustness.

Gerhard Dikta is a Professor in Mathematics and Applied Mathematics at the Aachen University of Applied Sciences (FH), Germany. His main research interests are in modelling, simulation, optimisation and time series, with applications in science and business.