World Patent Information xxx (2015) 1–9

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office

Peter Hingley [a, *], Walter Park [b]

[a] European Patent Office, Munich, Germany
[b] American University, Washington, DC, USA

ARTICLE INFO

Article history:
Received 30 January 2015
Received in revised form
1 July 2015
Accepted 10 July 2015
Available online xxx

Keywords:
Business cycles
Lognormal
Gross domestic product (GDP)
Patent filings forecasts
Research and development expenditure (R&D)
Linear model

ABSTRACT

An econometric model is applied to forecast future levels of patent filings at the European Patent Office out to 2019, using historical data from 1990 to 2013 with 28 source country terms. Descriptors include Research and Development expenditures and Gross domestic product, where the latter is split into trend and business cycles components. The model is applied to logarithmically standardised data.

The effects on the forecasts of additional future positive and negative stimuli to the GDP components are considered. Reasonable forecasting accuracy is found. Using a series of shorter historical data windows may give improved accuracy for short term forecasts.

1. Introduction

The European Patent Office (EPO) forecasts future patent filings in order to plan for likely workloads in the patent granting process, such as expected numbers of searches, substantive examinations, grants and renewals. These plans have implications for the requirements for staff and infrastructure. Procedurally, there is an annual cycle that proceeds from the forecasts via a business plan to finalisation in a budget document [1]. This budget is renewed annually and covers five years beyond the year in which it is produced.

The time series that are to be forecasted are shown in Fig. 1 with data up to 2013. EPO filings are a mixture of different types. Here we will consider forecasting the sum of Euro-direct filings and Euro-PCT international phase filings (Total filings in Fig. 1), after removing divisional filings (a form of retrospective Euro-direct filing that is forecasted separately). Other types of filings and downstream workload forecasts are then usually obtained by applying ratios to the forecasts for Total filings.

A variety of approaches are available that are based on historical data [1–3] or surveys [4]. The regression method that will be considered involves a dynamic log-linear model for annualised data that has been used since 2007 [5]. This operates on transformed EPO Total filings from 28 countries or regions, with autoregressive terms as well as source country Gross Domestic Product (GDP) and Research and Development expenditures (R&D) as independent variables. The model has recently been extended to consider the effects of business cycles [6]. This paper discusses the way that the approach has been customised for the forecasting process at EPO.

Matters of particular concern include transforming the data to achieve stationarity, how to calculate confidence intervals for the filings forecasts and how to interpret the forecasts and their accuracy against the later outcomes. The paper is organised as follows. Section 2 explains the model. In Section 3.1 a panel data set from 1990 to 2013 is fitted both to a model in levels and to a model in year-to-year differences. Section 3.2 shows the forecasts and interprets them. Section 4 considers the effect of a hypothetical boom or recession for one year during the forecasted period and also a scenario that is based on assumptions about the shape of the future business cycle. Section 5 looks at stability by fitting subsets of the same data in terms of a number of overlapping time windows. Section 6 discusses further directions.

2. The model for making parameter estimates and forecasts

The following regression model is used for EPO Total filings from a source country:-

* Corresponding author.
E-mail address: phingley@epo.org (P. Hingley).

http://dx.doi.org/10.1016/j.wpi.2015.07.002
0172-2190/

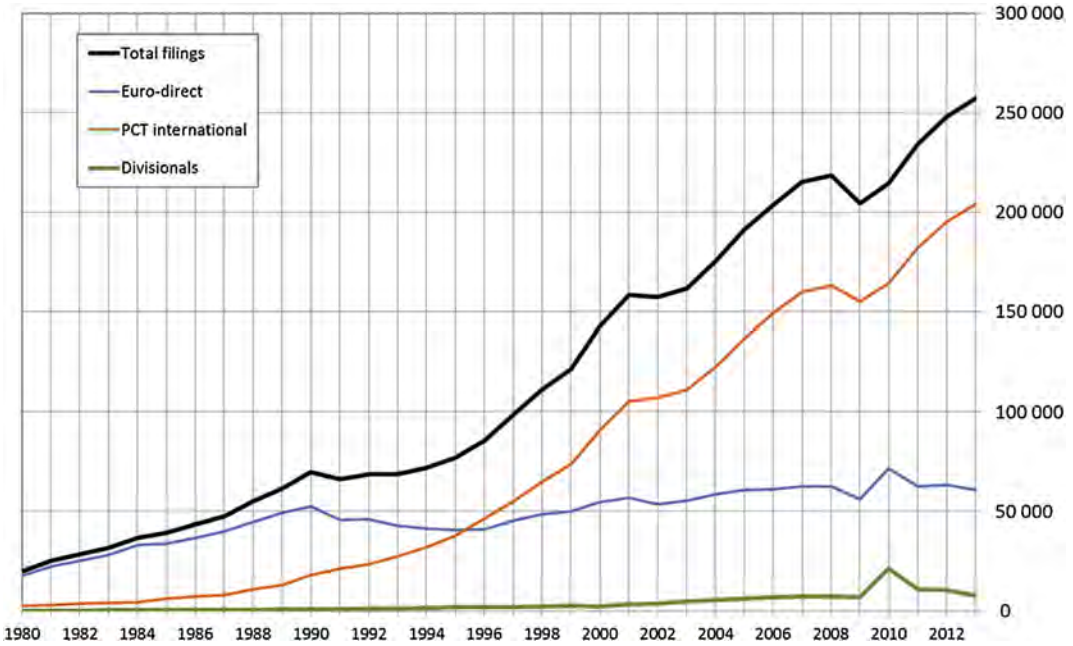

Fig. 1. The historical series of patent filings at EPO. (Adapted from Ref. [4]). Total filings are the sum of Euro-direct and PCT international phase from which divisional filings have been removed.

$$\log\left(\frac{P}{L}\right) = \alpha_0 + \alpha_1 \log\left(\frac{P}{L}\right)_{-1} + \alpha_2 \log\left(\frac{P}{L}\right)_{-2} + \alpha_3 \log\left(\frac{R}{L}\right) + \alpha_4 \log\left(\frac{Y^T}{L}\right) + \alpha_5 u + \varepsilon$$

Where P is the number of EPO Total filings from a source country[1];

L is the number of workers in the source country[2];
$_{-1}$ and $_{-2}$ indicate lags of one year and two years respectively;
R is R&D expenditures,[3] usually lagged by 5 years;
The GDP of the source country Y[4] is split into two components:-
Y^T is the "trend" level of output
u is the business cycle variable (a ratio of cyclical GDP to trend GDP);
ε is an error term, assumed to be normal with constant variance;
log( ) denotes natural logarithm.

Total filings P are transformed as indicated to log(P/L). This allows for a standardisation between countries, as L is treated as a proxy for country size, and for stabilising error by the logarithmic transformation. Based on [10], the value of R is lagged by five years in order to incorporate the concept that R&D expenditures have their effect after a delay. Most EPO filings are subsequent filings that take place up to a year after first filings, so the assumption is that R&D expenditures "cause" first filings about 4 years later on. Qualitatively similar results are obtained via a model with no lag in R [6], and in Section 5 below some comparisons are made between lags of 1, 3, 5 and 7 years. No time dummies are included, which gives better forecasting ability by assuming that the process remains stable over time.

The GDP term Y is decomposed via the Hodrick and Prescott filtering method [11] into its trend and cyclical components (Y^T and Y^C respectively) and then the business cycle variable is $u = Y^C/Y^T$. This is detailed in Ref. [6], where it is demonstrated that the usage of u and Y^T rather than Y improves the goodness of fit to the model for filings on the historical training data set, and so may also provide improved forecasting ability.

Annex 1 indicates the way that the forecasts for EPO filings from the source countries and their variabilities were calculated and combined to make the forecasts for Total filings. The authors will be prepared to share further details of the methods on request.

3. Results

3.1. Fitting the models

The analysis here reflects the data up to 2013 that were available in the second half of 2014. The model is fitted to a 28 source country-of-origin data set using annualised EPO Total filings from 1990 to 2013.[5] Data for the variables are calculated both as levels

[1] Filings refer to the sum of Euro-direct and PCT international phase filings [1], excluding divisionals, except where otherwise specified. Euro-direct are obtained from the EPO production database and PCT are as reported by WIPO.

[2] Number of workers data are provided by the World Bank [7].

[3] R&D expenditures are business enterprise research and development expenditures (BERD) from OECD MSTI 2013 edition 2 [8], at constant 2005 PPP international dollars. Comparable data are taken from UNESCO for countries that are not given by MSTI. For most countries, data were available up to 2012 at the time of analysis and have been trended out to 2013 and beyond by using linear regression on the last 10 years of available data.

[4] GDP expenditures are obtained from the World Bank's World Development Indicators [7], or Penn World Tables [9] for Chinese Taipei, standardised to real constant 2005 PPP international dollars. Agency forecasts for 2014 and 2015 are used where available and for later years have been trended by using linear regression on the last 10 years of available data.

[5] 27 individual countries were the following, together with a 28th group "ZZ" that represented the residual between the measured Total filings in a year and the sum from the 27 countries: Australia (AU), Austria (AT), Belgium (BE), Brazil (BR), Canada (CA), China & Hong Kong (CN-HK), Denmark (DK), Finland (FI), France (FR), Germany (DE), Hellas (GR), Ireland (IE), Israel (IL), Italy (IT), Japan (JP), Republic of Korea (KR), The Netherlands (NL), New Zealand (NZ), Norway (NO), Portugal (PT), Singapore (SG), Spain (ES), Sweden (SE), Switzerland (CH), Chinese Taipei (TW), United Kingdom (GB), United States of America (US), Others (ZZ).

Please cite this article in press as: P. Hingley, W. Park, A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office, World Patent Information (2015), http://dx.doi.org/10.1016/j.wpi.2015.07.002

Table 1
Model for EPO filings with training data set for Levels (1990–2013) and Differences (1991–1990 to 2013–2012). Parameter estimates and standard errors. AR1 and AR2 are one year and two year lags respectively for standardised patent filings, R is standardised R&D expenditure by business sector, Y^T is the trend level of standardised GDP, u is the business cycle variable. Also shown are Error variance and its square root, the Data point standard deviation. See footnote 5 for the key to the country names.

Parameter		Model in levels		Model in differences	
		Estimate	Standard error	Estimate	Standard error
Intercept	AU	−6.017	0.892	0.000	0.034
α_0	AT	−5.932	0.888	0.033	0.034
for	BE	−5.952	0.893	0.045	0.033
Countries:	BR	−5.965	0.861	0.123	0.034
	CA	−6.002	0.894	0.039	0.033
	CN-HK	−5.425	0.789	0.103	0.048
	DK	−5.817	0.877	0.056	0.034
	FI	−5.726	0.867	0.051	0.034
	FR	−5.923	0.884	0.018	0.033
	DE	−5.831	0.876	0.017	0.033
	GR	−6.208	0.909	0.055	0.033
	IE	−5.992	0.900	0.049	0.035
	IL	−5.828	0.878	0.054	0.034
	IT	−6.040	0.898	0.021	0.033
	JP	−5.825	0.875	0.051	0.033
	KR	−5.700	0.871	0.176	0.038
	NL	−5.830	0.879	0.031	0.033
	NZ	−5.902	0.879	0.065	0.033
	NO	−6.083	0.910	0.015	0.033
	PT	−6.105	0.902	0.171	0.036
	SG	−6.053	0.917	0.128	0.037
	ES	−6.066	0.899	0.089	0.034
	SE	−5.755	0.868	0.028	0.034
	CH	−5.751	0.871	0.035	0.033
	TW	−6.046	0.897	0.039	0.036
	GB	−5.975	0.885	−0.022	0.033
	US	−5.999	0.898	0.014	0.033
	ZZ (Others)	−5.763	0.820	0.052	0.034
α_1	AR1	0.823	0.038	−0.072	0.038
α_2	AR2	0.043	0.036	−0.045	0.037
α_3	R	−0.013	0.027	−0.048	0.103
α_4	Y^T	0.448	0.076	2.190	0.399
α_5	u	0.842	0.410	1.094	0.355
Error variance		0.0228		0.0247	
Data point standard deviation		0.151		0.157	

and as year-to-year differences in all the transformed independent and dependent variables (with one less time point in the training data series).

The parameter estimates and standard errors obtained by both approaches are shown in Table 1. Following [6], the intercepts α_0 are allowed to vary between countries in order to capture unobserved heterogeneity across countries, such as technological development or openness to international markets, while the parameters α_1 to α_5 are considered as fixed across countries so that pooled estimates are obtained. There are 33 estimated parameters, being 28 individual country intercepts and 5 slopes pertaining to α_1 to α_5. It was verified that the model with differences that included the split GDP variables Y^T and u gives a significantly better fit than the equivalent 32 parameter model that used only total GDP without splitting (F test on 1 and 639 degrees of freedom gives $0.025 > P > 0.01$).

For the model in levels, most of the fitted parameters are statistically significantly different from zero. But the high value of the AR1 term is an indication of a possible unit root that suggests non-stationarity of the data [12]. So the extent to which the set of independent variables is causal for the filings process is not guaranteed. For the model in differences, many parameters apart from the slopes Y^T and u that relate to GDP are not significant. But these significances are assessed at the level of differences and it may be that the effects are significant when the data are transformed back to levels. While it is tempting to suggest that it is good enough to model filings by using only contemporaneous GDP as a predictor, as is in fact done in some patent offices, we believe that this can lead to under-fitting. That kind of forecast is even more critically dependent on the quality of the underlying forecast used for future levels of GDP itself.

Regarding the lack of significance of the R&D expenditures variable R, anecdotal evidence from several patent offices suggests that the effect of R has become less important with the advent of more strategic patenting behaviour by applicants in recent years. It will be seen in Section 5 below that R does become significantly positive for some subsets of the data.

3.2. Interpreting the forecasts

In the following we consider only the model in differences.

As described in Annex 1, the fitted values/forecasts for Total filings for years 2014–2019 were calculated, together with estimates of their variances. Table 2 shows the forecasts for the model in differences.

Fig. 2 shows the observed data for recent years as well as the fitted values for 2013 and the forecasts for years 2014–2019. The Total filings numbers are reported first, and then the filings from the important countries of origin China, Germany, Japan and United States of America. 95% confidence intervals are calculated for each individual forecasted year and the forecasts and limits are connected over time by smoothed lines using Excel.

The model gives optimistic forecasts for Total filings, particularly towards the end of the period. A compound annual growth rate is suggested of 6.9% from about 258 000 in 2013 to about 384 000 in

Please cite this article in press as: P. Hingley, W. Park, A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office, World Patent Information (2015), http://dx.doi.org/10.1016/j.wpi.2015.07.002

Table 2
Total filings forecasts by the model in differences. The results for 2013 show the model fit to the last year in the training data set, with standard errors in brackets.

Year	Actual total filings	Model in differences	
		Total filings forecast	Standard error
2013	257 457	263 931	(3202)
2014		269 097	4848
2015		290 054	6130
2016		305 976	7301
2017		328 792	8416
2018		355 074	9568
2019		384 053	10 829

2019. The confidence intervals widen towards the end of the forecasting period.

Regarding the country forecasts, a feature to notice in Fig. 2b for Japan is that the forecasts experience a downward "kink" in 2014. There is a strong level of projected growth for China, where about 54 000 additional filings are expected by 2019 compared to 2013. This reflects strong domestic filings growth in China both in the past and as predicted for the future [13]. The growth rates that are predicted for Japan, Germany and US in Fig. 2b, 2c are more modest.

4. Effects of variations of GDP over the forecast period

Experiments were done to consider the effects on the forecasts of varying the assumptions about how future economic growth will develop, in terms of possible booms and dips in GDP during the forecast period.

Firstly, positive or negative shocks were imposed onto the assumed standardised values of Y^T and u in the year 2016 only. This is the first as-yet-unknowable year beyond the horizon for which predictions of GDP were readily available from professional forecasters in 2014 [EG see Ref. [14]]. The assumed values were taken the same as for the model in differences in Table 2, but for a boom (respectively dip), a 1% increase (respectively decrease) was imposed on Y^T. Also a 10% increase (respectively decrease) was imposed on u if it was positive, or a 10% decrease (respectively increase) was imposed on u if it was negative. The logic for a larger relative displacement to u than to Y^T is that u is essentially more variable from year-to-year because it is a cycle indicator. This simulation was taken to emulate a minor one-off effect on all relevant world economies simultaneously in 2016, such as has happened from time to time in the past, rather than a more major event like the 2009 recession. Table 3 shows the results of these perturbations.

For 2016, the boom increases Total filings by 6750 (+2.2%) and the dip decreases filings by 6660 (−2.2%). However the numbers of filings in 2017 and thereafter are hardly affected at all by the perturbations to 2016. This behaviour is a little different to another experiment (reported in Ref. [6]) where a larger single-year shock caused effects to persist for a year beyond the year directly affected. The behaviour here may be due to the fact that, in the model in differences, a one off boom (respectively dip) in the difference from 2015 to 2016 is followed by a one off dip (respectively boom) in the difference from 2016 to 2017, caused simply by the rectification in 2017 to the levels of the independent variables in 2016. Also the estimated autoregressive parameters are rather low for the model in differences.

This result does conform to the observed experience in the filings time series at EPO. For example, in 1997 there was a one-off reduction in initial filing fees that led to a temporary boom in filings for that year only. This is a somewhat smaller causative effect from within the system than a change to over all economic growth, but it does show that a single perturbing shock can change filings for a short period only.

In a second experiment, the fixed assumptions about future levels Y^T of GDP were maintained, but with a process of cycles artificially imposed to give a synthetic "boom" in 2014–2016, that is then followed by a synthetic "bust" in 2017–2019.

The resulting forecasts for Total filings are shown in the right side column in Table 3. The series of forecasts remains reasonably similar to those shown earlier in Table 2. During the early years from 2014 to 2016, the new forecasts are higher than those of Table 2, because of the imposed cyclic boom, with a peak difference in 2016 of about 6000 filings (312 026–305 976). But from 2017 onwards, the forecasts are lower due to the imposed mild bust. It remains to be seen whether this prediction will be borne out by the facts later on. However, this variation from the earlier scenario depends on the presumptions made about likely future cyclicity. Other scenarios for cycles could be chosen to perform what-if type analyses for future filings.

5. Forecasts for subsets of the data

Bearing in mind the possible heterogeneity of the patent filings process due to its susceptibility to bouts of temporal enthusiasm and lethargy among applicants, it is interesting to fit the model on training data from a series of short time windows and to look at variations of parameter estimates and forecasts over successive windows. Parameter estimation can be done on short windows because there are a large number of degrees of freedom for the error term due to the multiplication of data points over the 28 country data set. Fig. 3 shows forecasts up-to-7 years ahead that are based on successive 7-year windows of the training data within the over all 23 year data set from 1991 to 2013.[6]

It can be seen that there was apparent under-enthusiasm for the first three training data sets 1991–1997 up to 1993–1999. The next six training sets (1994–2000 up to 1999–2005) gave reasonable forecasts for at least the first three forecasted years in each case. After that, for the four training sets from 2000–2006 to 2003–2009, there was over-enthusiasm because of the inability to forecast the one-off drop in Total filings in 2009 that was due to the onset of the financial crisis. The next training set gave reasonably good forecasts as far as can be checked up to now, but after that some under-enthusiasm returned for the last training set shown for 2004 to 2010. It is easy to be wise after the event about the financial crisis (and to some extent about the earlier termination of the dot-com boom), but there was not much in the independent variables of the models that indicated the likelihood of those down-turns in advance.

Fig. 4 shows the estimates for some of the model parameters over the training data windows. Variations can be seen between years that are smoothed to some extent by the usage of overlapping data sets for the successive windows.

The patterns of variation between the windows demonstrate that there may be some correlation effects between parameter estimates. In the rest of this paragraph, numbers in square brackets are equivalent parameter estimates for the whole 23 year data set from Table 1 and represent norms to which the estimates from the windows can be compared. The autoregressive terms AR1(×10)

[6] For the data in each window, independent variables were assumed known only up to the end of the window, and were estimated by trends beyond the end of the window from 10 years previous data. Trended values of lagged R&D expenditures were used from 3 years beyond the end of the window, because the values for the last two years of each window are not usually known at the time of analysis. Forecasts from 15th (2005–2011), 16th (2006–2012) and 17th (2007–2013) sets of data are not shown in Fig. 3.

Please cite this article in press as: P. Hingley, W. Park, A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office, World Patent Information (2015), http://dx.doi.org/10.1016/j.wpi.2015.07.002

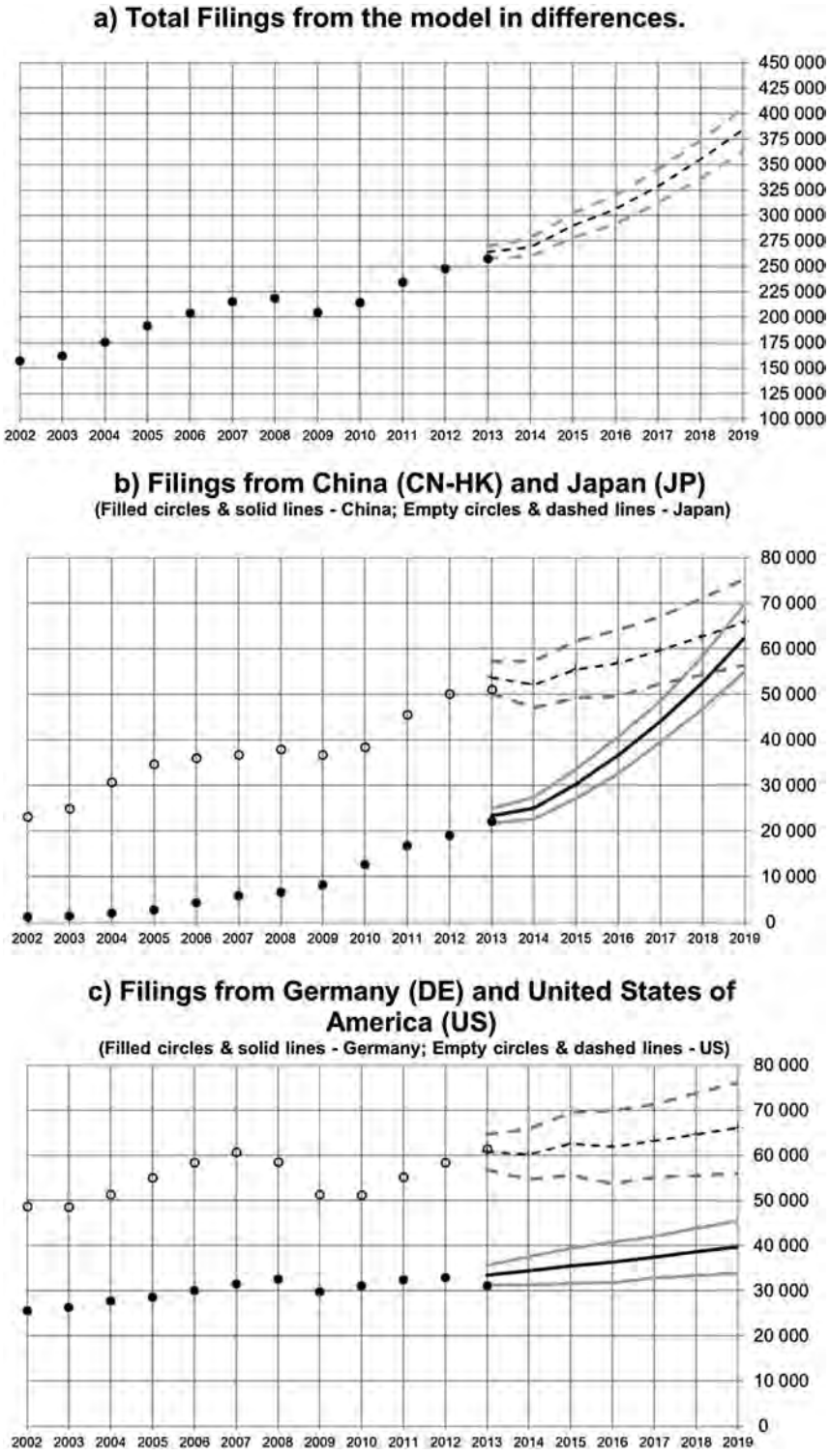

Fig. 2. Filings forecasts by the model in differences. Black lines are the forecasts and grey lines are the 95% confidence intervals for the forecasts. a) Total filings by model in differences. b) Filings from China and Japan. c) Filings from Germany and United States of America.

[−0.72] and AR2 (×10) [−0.45] are generally negative and are often quite large, which argues that the short period data accentuates the role of self-determination in the estimation. There may be complementary trends between the estimated slope coefficients for the GDP based variables u [1.09] and Y^T [2.19]. The 5 year lagged R&D expenditures variable R [−0.05] may exhibit long period cyclicity.

Although the windows approach allows for variations and heterogeneities over time to be considered, there is a statistical balance to be struck in terms of accuracy of estimation in comparison to an approach that fits the model over a longer period. This can be assessed by considering an alternative system of windows of increasing length. These cumulative windows are fitted from the same starting year 1991, and then extending from 7 years to 8, and so on, up to the full 23 year data set as in Section 3. As may be

Please cite this article in press as: P. Hingley, W. Park, A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office, World Patent Information (2015), http://dx.doi.org/10.1016/j.wpi.2015.07.002

Table 3
Total filings forecasts by the model in differences, as in Table 2. The effects of positive or negative boosts in 2016 to the GDP related variables Y^T and u are shown in the left and central and parts of the table. The effect of imposing a cycle (without altering Y^T) is shown in the right hand part of the table.

Year	Actual total filings	Positive boost to u (10%) & trend (1%) in 2016		Negative boost to u (10%) & trend (1%) in 2016		Scenario of imposed cycles	
		Total filings forecast	Standard error	Total filings forecast	Standard error	Total filings forecast	Standard error
2013	257 457	263 931	(3202)	263 931	(3202)	263 922	(3202)
2014		269 097	4848	269 097	4848	270 746	4940
2015		290 054	6130	290 054	6130	303 821	6836
2016		312 825	7195	299 216	7586	312 026	8278
2017		328 278	8686	329 316	8521	325 689	9733
2018		354 758	9808	355 393	9677	347 718	10 835
2019		384 106	11 056	384 000	10 921	382 368	12 041

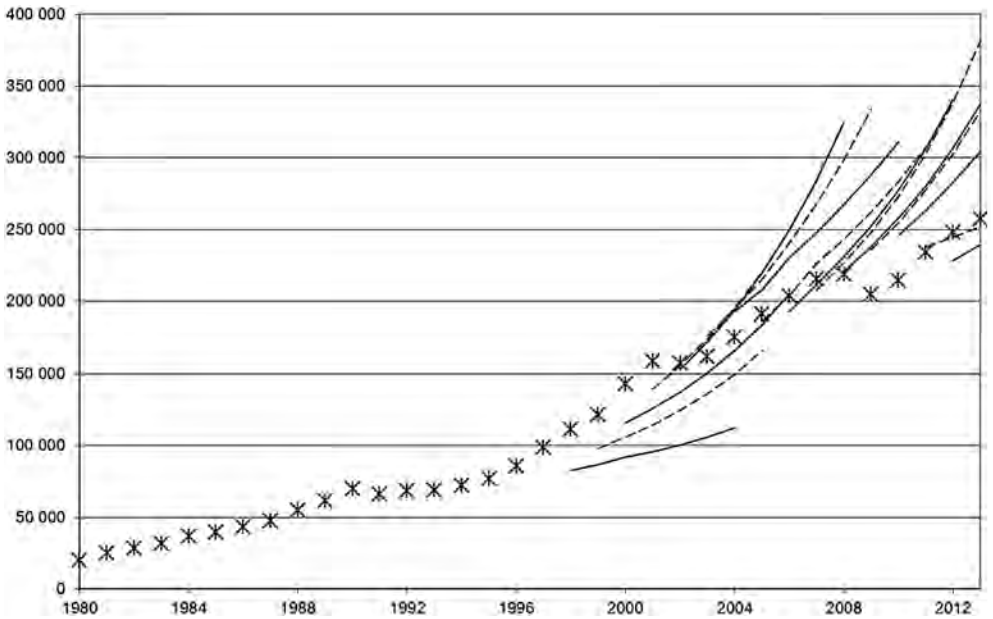

Fig. 3. Filings forecasts by the model in differences when fitted to successive seven year data windows (1991–1997, 1992–1998, …, 2004–2010). The points show the observed data/out-turns and the lines show the forecasts.

expected, the Data point standard deviation decreases from the first window (0.211) for 1991–1997 to the last one (0.157) for 1991–2013.

These stability issues are demonstrated further in Table 4, which shows three forecast error statistics for both sets of windows experiments for horizons from one to seven years ahead. Mean percentage error statistics may be of most use to planners who wish to steer the process over several years. However a measure that is used more often in the literature is the Mean absolute percentage error (MAPE) [15]. Median absolute percentage errors (Median APE) are also shown.

Mean percentage errors out to 4 years ahead seem to be acceptably low, but these reflect aspects of over-optimism cancelled by over-pessimism. The 7 year windows give lower MAPE values than cumulative windows for 1–2 years ahead, while the cumulative windows perform better for 3–7 years ahead. This is consistent with an in-house EPO aphorism that longer term forecasts should be generated by longer historical training data sets and shorter term forecasts by shorter data sets.[7] But MAPE values of 10% and over for only 1 year ahead are rather high. This may have been caused by occasional bad years near the unexpected process breaks in 2002 and 2009, which is why the median APE values behave somewhat better for the first few years ahead, particularly for the cumulative windows.

Although the overall estimate for R in Table 1 was not statistically significant, in Fig. 4 the estimate for R varies between positive and negative values that are sometimes statistically different from zero. At the extremes, the R value is −0.54 for 2003 with a standard error of 0.20, while 0.34 for 2010 has a standard error of 0.13. R was lagged by 5 years in these results and this can be justified by the following experiment.

The 5 years windows exercise was repeated at alternative lags for R of 1, 3 and 7 years. The model fit in terms of Data point standard deviation (as in Table 1) is best at 1 year lag for 13 of the 17 tested windows (3, 5 and 7 year lags are best in 1, 2 and 1 windows respectively). A lag of 1 year also looks best in terms of numbers of windows in which the fitted value of R is positive (all 17 windows at 1 year compared to 6, 8 and 7 windows at 3, 5 and 7 years respectively).

However it is also important to see which lag gives the best forecasting performance. Table 5 shows MAPE values for the various lags.

In terms of MAPE, the 5 and 7 year lags beat lags of 1 and 3 years at all the forecasting horizons. The 7 year lag beats the 5 year lag at horizons up to 4 years, while the 5 year lag beats the 7 year lag at

[7] Related to the corresponding author by Mr. Raphael de Roeck.

Please cite this article in press as: P. Hingley, W. Park, A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office, World Patent Information (2015), http://dx.doi.org/10.1016/j.wpi.2015.07.002

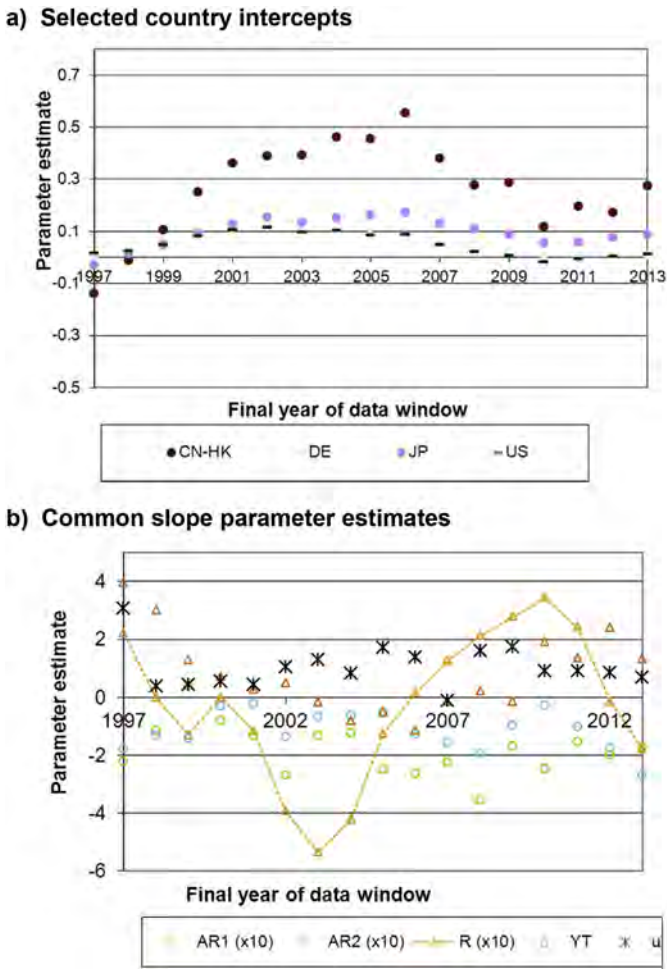

Fig. 4. Parameter estimates by the model in differences that was fitted to successive seven year data windows (1991–1997, 1992–1998, …, 2007–2013). The points show the estimated parameters from the training data window that ended in the previous year. (For keys to parameter symbols, see Table 1).

horizons of 5–7 years. The 5 year lag may be preferable to the 7 year lag because of this better longer term forecasting performance. Other reasons for this choice include that the 5 year lag is already at the high end of rational expectation, that the 7 year lag gives fewer positive estimates of R and that the 7 year lag tends to give higher Data point standard deviations.

6. Future directions

Since the windows approach in Section 5 has shown some heterogeneity over time in the behaviour of the data vis-a-vis the model, a possible enhanced technique that may show some promise is to replace the linear model with a weighted version that gives more weight to the more recent observations.

Another possibility is to create separate dynamic log-linear models for filings in distinct technologies or industries. Two or three major areas may be enough (say "pharmaceuticals/biology"; "electricity"; and "the rest"). Only a small number of industries should be selected because, apart from the increased effort involved in fitting several models instead of only one of them, there are difficulties in finding accurate enough long term time series of industry specific R&D or GDP data.

This paper has been about Total filings, but forecasts for Total applications (Euro-direct + Euro-PCT regional phase filings [1]) are also important for workload planning at EPO. Rather than using the Total filings forecasts and applying ratios, more effort is being made recently to forecast Total applications directly. However, since the PCT part of Total applications is further down the road from the first filing than in Total filings, structural models that are dependent on R&D expenditures and GDP may not fit so well in this case.

The modelling approach that was adopted here did not deal well with major events that led to disruptions of the patent filings trends, in particular the financial crisis of 2008 and 2009. However the availability of the business cycle variable u allows for 'what-if' planning against similar further disruptions. Manipulation of u makes it easier for future forecasted filings possibly to go down as well as up, a situation that is otherwise difficult to model because the history of EPO patenting has always previously been upwards, except for a small dip in 2009.

On its own no single model should necessarily be trusted. Each year the forecasted scenario from this model is compared with results of other methods, typically an applicant survey [4], simpler regression-fitting exercises, and a Box–Jenkins based Auto Regressive Integrated Moving Average model with Exogenous Input (ARIMAX type model) that also includes GDP and R&D expenditures as prognostic factors [10,3]. But these other approaches are

Table 5
Model in differences for EPO filings. Forecast errors as mean absolute percentage error (MAPE) for lags of R of 1 year, 3 years, 5 years and 7 years. Comparison of forecasting accuracy from one to seven years ahead of the training data set. 7-year windows data as in Fig. 3.

Lag for R	MAPE with 7 year windows			
	1 year	3 years	5 years	7 years
1 year ahead	10.9%	9.5%	10.0%	9.4%
2 years ahead	13.1%	11.6%	11.0%	10.6%
3 years ahead	18.5%	15.8%	14.8%	13.9%
4 years ahead	25.4%	20.7%	19.1%	18.7%
5 years ahead	33.8%	26.1%	23.4%	23.8%
6 years ahead	45.1%	33.0%	28.6%	29.6%
7 years ahead	60.9%	42.5%	35.4%	37.0%

Table 4
Model in differences for EPO filings. Forecast errors as percentages. Mean percentage error, mean absolute percentage error (MAPE), median absolute percentage error (Median APE). Comparison of forecasting accuracy from one to seven years ahead of the training data set. 7-year windows data as in Fig. 3. Cumulative windows as described in the text.

	7 year windows			Cumulative windows		
	Mean percentage error	MAPE	Median APE	Mean percentage error	MAPE	Median APE
1 year ahead	−3.5%	10.0%	9.3%	−4.0%	10.5%	6.6%
2 years ahead	−0.6%	11.0%	8.8%	−1.1%	11.4%	5.6%
3 years ahead	3.6%	14.8%	12.9%	2.7%	14.1%	10.5%
4 years ahead	8.7%	19.1%	18.1%	7.7%	18.4%	18.6%
5 years ahead	13.9%	23.4%	23.8%	12.7%	22.6%	20.6%
6 years ahead	19.0%	28.6%	31.8%	17.0%	26.8%	25.7%
7 years ahead	25.6%	35.4%	36.5%	22.4%	33.2%	37.3%

Please cite this article in press as: P. Hingley, W. Park, A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office, World Patent Information (2015), http://dx.doi.org/10.1016/j.wpi.2015.07.002

limited to considering Total filings as a single series, or else split down by a few major areas such as the IP5 regions, rather than the more extensive 28 country data set that is available for the dynamic log-linear model.

Only one new data point is added each year to the total filings series, although there can also be retrospective corrections made for earlier years. It is unlikely always to be most appropriate to select the best fitting method after inclusion of the latest data point or the latest research results, because this can lead to over-fitting and radical changes of models each year. Rather, it seems better to have a fixed type of descriptive model and adapt it mildly from time to time to suit the circumstances. In practice a weighted averaging procedure for making forecasts over scenarios is usually taken, with weights related to the known accuracy of each component forecast for the preceding year [1].

Forecasts of patenting activity are integral to patent offices around the world for purposes of planning and budgeting. They are also important for policymakers to measure trends in innovation and to multinational businesses that monitor trends in international technology diffusion to gauge market potential. The modelling approach developed here could be useful to such practitioners. One of the major difficulties about projecting future trends is the impact of business cycles. Cyclical shocks are difficult to predict but they should not be ignored. The academic literature thus far has followed two (not mutually exclusive) approaches to predicting recessions: one is to use leading indicators about financial conditions such as bank stress [16,17] and another is to estimate probit models of whether a recession will occur or not [18,19]. These approaches are limited in providing only short term forecasts and in predicting only the occurrence of cycles and not their intensity. They also ignore the fact that cycles have boom periods as well as recessions. Our approach has demonstrated how to construct longer term, out of sample, predictions of the effects of business cycles of varying sizes and to consider deviations above and below trend GDP.

Another possible application could be to help predict the development of patent systems in developing countries or regions, in connection with studies that connect their business cycles with those in developed countries (EG Ref. [20]).

The approach to analyse the lognormal data that was taken here may also be of interest for some other more general areas of application. It is important to avoid bias in forecasts that can otherwise creep in due to a failure to consider that the lognormal mean contains a term described from the variance as well as the normal mean on the log scale (see Annex 1).

Acknowledgements

We thank the European Patent Office for its sponsorship. The suggestions that were made are the opinions of the authors, who are responsible for any errors. The filings counts reported may not be the same as those in EPO publications. Specific forecasts that were shown are not per-se the official forecasts as used by EPO for its internal planning purposes.

Annex 1. The calculation of the forecasts and their variabilities.

Here we give a description of the calculation of the forecasts for the countries and their variances. Then we describe the amalgamation of the forecasts and variance estimates to produce the estimate and confidence limits for Total filings.

At the level of filings (P), the assumed distribution of the error, and hence also that of P itself, is assumed to be lognormal [21]. The technique for estimation that will be described takes this into account. Let $v = P/L$, with $\log(v)$ distributed as $N(\mu,\sigma^2)$, a normal distribution with mean μ and variance σ^2.

A suitable formulation of the model has a separate intercept for each of 28 countries (α_{01}, α_{02}, …,α_{028}), and common slope parameters (α_1, α_2, α_3, α_4 and α_5). The linear model is fitted to the transformed data for the various countries simultaneously to determine the parameter estimates for each country.

Let Z be the $n \times p$ design matrix of independent variables, including the country specific intercepts, and B denote the $p \times 1$ parameter vector. Here $n = 672$, since there are 28 countries and 24 years.

The parameter estimates $\widehat{B}$ are calculated by least squares and the associated error variance of ε is calculated as $\widehat{\sigma}^2$.

$$\widehat{B} = Z \cdot \left(Z^T Z\right)^{-1} \cdot Z^T \cdot \log(v)$$

$$\widehat{\sigma}^2 = \left(\log(v) - Z\widehat{B}\right)^T \cdot \left(\log(v) - Z\widehat{B}\right)/(n-p)$$

where $\log(v)$ is the $(n \times 1)$ vector of the dependent variable.

For the model in differences, $\log(v)$ is substituted by year to year differences in the logs and the transformed versions of the independent variables are also taken as differences, except for u. Taking differences removes one time point, so then $n = 28 \times 23 = 644$.

On the logarithmic scale, the fitted values of the observations are given by the matrix inner product Z. $\widehat{B}$ within the training set. The forecasts for each country for each future time point are given by projecting further $(1 \times p)$ rows z, that are equivalent to rows of Z but with independent variables beyond the data set. The forecast is calculated as $z.\widehat{B}$.

The fitted values on the scale of v are taken from the linear model as $\widehat{\gamma} = \exp((z.\widehat{B}) + \widehat{\sigma}^2/2)$. The estimated number of filings from a country at a given time point is then $w = L \cdot \widehat{\gamma}$, with an estimated variance $\mathrm{Var}[w] = L.^2\, \widehat{\gamma}^2\, (\exp(\widehat{\sigma}^2) - 1)$ [21].

It should be noted that, during the forecast period beyond the training data set, it is also necessary to forecast L and the independent variables Y^T, u and R. Forecasts for L are made by straight line regression projection from the 10 most recent available years in the training set. Using these, forecasts for log(R/L) are then made by regression over 10 earlier years. For Y^T and u, forecasts for Y are first obtained by a second order autoregressive model without an intercept, after which the forecasts for $\log(Y^T/L)$ and u are obtained from the Hodrick and Prescott filter [6].

The goal is to forecast Total filings as the sum of the forecasted filings per country of origin. This is $\widehat{w}_{TOTAL} = \Sigma\, w_i = \Sigma\, L_i.\, \widehat{\gamma}_i$, where Σ indicates summation over countries with L_i and $\widehat{\gamma}_i$ for country i.

In the following, Ξ is the (28×28) covariance matrix between countries on the log scale, that is estimated from the linear model by $\widehat{\Xi} = (Z^T Z)^{-1}.\, \widehat{\sigma}^2$. The variance of $\widehat{w}_{TOTAL}$ is estimated as the sum of the estimated covariances for all pairs of countries. This is based on an extension of the formula for the variance of the mean of a lognormal distribution [22].

$$Var\left[\widehat{w}_{TOTAL}\right] = Var\left[\sum_i L_i \cdot \widehat{\gamma}_i\right] = \sum_i \sum_j L_i \cdot L_j \cdot \widehat{\gamma}_i \cdot \widehat{\gamma}_j \left(\exp\left(\widehat{\Xi}_{ij}\right) - 1\right)$$

Summation is over all country pairs i and j; 1, …, i, …, 28; 1, …, j, …, 28.

Since the model includes autoregressive terms that relate to lags of filings at one and two years, the forecasts for more than two years out themselves use inputs of filings forecasts from one and

Please cite this article in press as: P. Hingley, W. Park, A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office, World Patent Information (2015), http://dx.doi.org/10.1016/j.wpi.2015.07.002

two years previously. These inputted forecasts are subject to variability by the same error process. It is likely that Var[$\widehat{w}_{TOTAL}$] is not great enough to cover all the variability that is inherent in this approach. In order to cope with this to some extent, a pragmatic compound variance method is used. The variance of the filings forecast for a future year s is given by the sum of the variances taken over all the forecasted years, up to and including s.

$$\mathrm{Var}\left[(\widehat{w}_{\mathrm{TOTAL}})_s\right] = \sum_{i=1}^{s}\left[\mathrm{Var}(\widehat{w}_{\mathrm{TOTAL}})_i\right]$$

From this variance, 95% confidence limits for the forecast of Total filings in a future year are calculated by the usual normal assumption, $(\widehat{w}_{TOTAL})_s \pm 1.965 \times \mathrm{SE}[(\widehat{w}_{TOTAL})_s]$, where SE indicates standard error and is the square root of Var[$(\widehat{w}_{TOTAL})_s$]. These confidence limits are appropriate for the predicted values of the mean. It is considered that the mean is forecasted because the process uses essentially unchanging historical training data for all years up to the last year in the data set, with only one added data point per country in each successive annual forecasting exercise.[8]

References

[1] P. Hingley, M. Nicolas, Improving forecasting methods at the European Patent Office, in: P. Hingley, M. Nicolas (Eds.), Forecasting Innovations, Springer, Heidelberg, 2006.
[2] P. Hingley, M. Nicolas, Methods for forecasting numbers of patent applications at the European Patent Office, World Pat. Inf. 26 (3) (2004) 191–204.
[3] A. Hidalgo, S. Gabaly, Use of prediction methods for patent and trademark applications in Spain, World Pat. Inf. 34 (1) (2012) 19–35.
[4] European Patent Office, Patent Filings Survey 2013: Intentions of Applicants Regarding Patent Applications at the European Patent Office and Other Offices, 2014. See archive at, http://www.epo.org/service-support/contact-us/surveys/future-filings.html.
[5] W. Park, International patenting at the European Patent Office: aggregate, sectoral and family filings, in: P. Hingley, M. Nicolas (Eds.), Forecasting Innovations, Springer, Heidelberg, 2006.
[6] P. Hingley, W. Park, Do business cycles affect patenting? Evidence from European Patent Office Filings. Working paper.
[7] World Bank, World development indicators http://data.worldbank.org/data-catalog/world-development-indicators.
[8] Organisation for Economic Cooperation and Development, Main Science and Technology Indicators, 2014. http://www.oecd.org/sti/msti.htm.
[9] University of Pennsylvania, Penn World Tables. https://pwt.sas.upenn.edu.
[10] G. Dikta, Time series methods to forecast patent filings, in: P. Hingley, M. Nicolas (Eds.), Forecasting Innovations, Springer, Heidelberg, 2006.
[11] R. Hodrick, E. Prescott, Postwar Business Cycles: An Empirical Investigation, J. Money, Credit, Bank. 29 (2007) 1–16.
[12] J. Hamilton, Time Series Analysis, Princeton University Press, 1994.
[13] World Intellectual Property Review, China Seeks to Treble Patent Filings by 2020, 2015. http://www.worldipreview.com/news/china-seeks-to-treble-patent-filings-by-2020-7603.
[14] Federal Reserve Bank of Philidelphia, Livingston Survey of Economists' Expectations, 2015. http://www.philadelphiafed.org/research-and-data/real-time-center/livingston-survey/.
[15] N. Meade, An assessment of the comparative accuracy of time series forecasts of patent filings: the benefits of disaggregation in space and time, in: P. Hingley, M. Nicolas (Eds.), Forecasting Innovations, Springer, Heidelberg, 2006.
[16] C. Hakkio, W. Keeton, Financial stress: what is it, how can it be measured, and why does it matter? Federal Reserve Bank of Kansas City, Econ. Rev. 2Q (2009) 1–50.
[17] J. Hatzius, P. Hooper, F. Mishkin, K. Schoenholtz, M. Watson, Financial Conditions Indexes: a Fresh Look after the Financial Crisis, National Bureau of Economic Research Working Paper, 2010. No. 16150.
[18] H. Kauppi, P. Saikkonen, Predicting U.S. Recessions with dynamic binary response models, Rev. Econ. Stat. 90 (4) (2008) 777–791.
[19] H. Nyberg, Dynamic probit models and financial variables in recession forecasting, J. Forecast. 29 (2010) 215–230.
[20] A. Pesce, Economic Cycles in Emerging and Advanced Countries – Synchronization, Internal Spillovers and the Decoupling Hypothesis, Springer, Heidelberg, 2015.
[21] N. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, vol. 1, Wiley, 1994.
[22] P. Soderlind, Lecture Notes – Econometrics: Some Statistics, 2013. http://home.datacomm.ch/paulsoderlind/Courses/OldCourses/EcmXSta.pdf.
[23] N. Draper, H. Smith, Applied Regression Analysis, second ed., Wiley, New York, 1981.

[8] Alternative confidence interval/prediction interval formulae are given in the statistical literature (EG Ref. [23]), depending whether the intervals are sought for the mean or for an individual new observation.

Please cite this article in press as: P. Hingley, W. Park, A dynamic log-linear regression model to forecast numbers of future filings at the European Patent Office, World Patent Information (2015), http://dx.doi.org/10.1016/j.wpi.2015.07.002