An approach for modelling and forecasting research activity related to an emerging technology

Abstract

The understanding of emerging technologies and the analysis of their development pose a great challenge for decision makers, as being able to assess and forecast technological change enables them to make the most of it. There is a whole field of research focused on this area, called technology forecasting, in which bibliometrics plays an important role. Within that framework, this paper presents a forecasting approach focused on a specific field of technology forecasting: research activity related to an emerging technology. This approach is based on four research fields – bibliometrics, text mining, time series modelling and time series forecasting – and is structured in five interlinked steps that generate a continuous flow of information. The main milestone is the generation of time series that measure the level of research activity and can be used for forecasting. The usefulness of this approach is shown by applying it to an emerging technology: cloud computing. The results enable the technology to be structured into five main sub-technologies which are characterised through five time series. Time series analysis of the trends related to each sub-technology shows that privacy & security has been the most active sub-technology to date in this area and is expected to maintain its level of interest in the near future.

Keywords: Technology forecasting, research-activity forecasting, bibliometrics, text mining, trend analysis, structural time series models.

Introduction

Emerging technologies currently pose a great challenge for decision makers in both the private and public spheres. The analysis of their development and implications has an impact on critical decisions made at a wide range of organisations. Accordingly, today's private enterprises and policy makers need to be able to assess and forecast technological change in order to take advantage of it. Related to this, *technology forecasting* has sparked interest among researchers and practitioners, and as a result many methods have been proposed and applied. From the initial studies experts agree that these methods should be used in combination so as to offset the weaknesses of one forecasting method with the strengths of another (Martin and Daim 2012). In this context, the main objective of this article is to forecast trends in an emerging technology based solely on the research activity related to it.

Forecasting research activity is a complex task that involves different fields. This paper proposes an approach for dealing with this problem which combines four methodologies with the following specific objectives:

- *Bibliometrics:* to retrieve scientific publications related to an emerging technology and generate a cleaned data set based on them.
- *Text mining:* to generate time series based on the content of the data set. These series represent the research activity associated with different sub-technologies within the emerging technology and enable it to be depicted over time.
- *Time series modelling*: to specify appropriate models for the time series using the structural time series modelling methodology introduced by Harvey (1989).
- *Time series forecasting:* to obtain forecasts of the short-term development of the research activity related to the technology.

This paper makes four main contributions to technology forecasting: first, we propose an integrated forecasting approach for research activity; second, we characterise the research activity related to an emerging technology via monthly time series which represent predominant sub-technologies; third, we use Structural Time Series Models to forecast these time series, i.e. the research activity in the technology; and fourth, we show the usefulness of the approach proposed by applying it to an emerging technology, cloud computing (CC), which is a cutting edge technology with a clear impact on businesses today.

The article is organised as follows: the Background section defines the fields of research on which this approach is based and provides examples from the scientific literature. The Research Approach section describes the proposed approach step by step, detailing the tasks associated with each step. The Data section explains CC technology and the process for generating the time series data is described. The Results section sets out some insights into the future development of CC technology by analysing five sub-technologies. Finally, the Conclusions and Future Work section summarises the most important results obtained, stresses the relevance of this kind of analysis to decision makers and establishes some lines for future research to deal with the current limitations of the approach.

Background

Our theoretical and practical framework is based on the research fields of bibliometrics, text mining and technology forecasting, specifically research-activity forecasting using time series models. Bibliometrics structures the information contained in scientific publications, which is the basis for a consistent forecasting exercise. It can thus help researchers to map and profile their entire research domain (Börner et al. 2003). Text mining goes a step further and processes the contents of the publications (Kostoff and Geisler 1999). Text mining tools can be easily used to examine research trends and patterns in the fields of technology management due to the increasing appearance of software developed specifically for such applications (Porter et al. 2005¸ Mikut and Reischl 2011). Time series models forecasting uses statistical techniques to identify and estimate patterns in the trends in data and project them into the future.

Each research field offers a number of methods which can be combined to create an integrated approach to forecast the changes over time in a technology. In this context, numerous papers can be found where bibliometrics and text mining analysis methods are used to construct the basis for some kind of forecasting exercise (Bengisu and Nekhili 2006; Daim et al. 2006; Kajikawa et al. 2008; Chen et al.

2011). In terms of statistical methods, the most common approach for forecasting the future evolution of a technology based on bibliometrics data is growth curve analysis (see Martino 2003 for further discussion). However, a few authors obtain their forecasts using time series models such as moving average models (Jun and Uhm 2010) and autoregressive integrated moving average (ARIMA) models (Brunk 2003; Fernández-Cano et al. 2012).

Within this framework of time series forecasting, we propose to use the Structural Time Series Models (STSMs) presented by Harvey (1989). These models belong to the class of Unobserved Components models (Pagan 1975; Engle 1979; Kitagawa and Gersch 1984), which decompose the series into components such as trends, cycles and seasonalities. The advantage of STSMs is that each component is specified stochastically so they can change over time following changes in the data. For instance, a stochastic trend will change according to the average growth patterns in the data. Since our objective is to identify and forecast trends, STSMs are flexible enough to enable us to capture the changes that occur in the life cycle of the research activity in an emerging technology. Furthermore, these models can easily be extended to handle specific features of series that are difficult to deal with in other time series frameworks, such as heteroscedasticity, nonlinearity and non-Gaussianity (Proeitti 2002). Even though these are relatively new models, the said features have enabled them to be applied satisfactorily for forecasting different type of series (Gonzalez and Moral 1995 for tourism demand; Schlink and Herbarth 1997 for SO_2 data; Ghosh et al. 2009 for short-term traffic flow; Dilaver and Hunt 2011 for energy demand). We have found no applications to date in the field of bibliometrics.

Research Approach

Fig. 1 shows the layout of the integrated approach proposed for forecasting the research activity related to an emerging technology, with a step-by-step description of the methodology required and the outcome obtained. Roughly speaking, the purpose of the first two steps is to construct the data set, the third step focuses on time series generation and the fourth and fifth steps deal with the analysis of those time series, modelling and forecasting, respectively. The sub-sections below explain all the tasks required in each step.

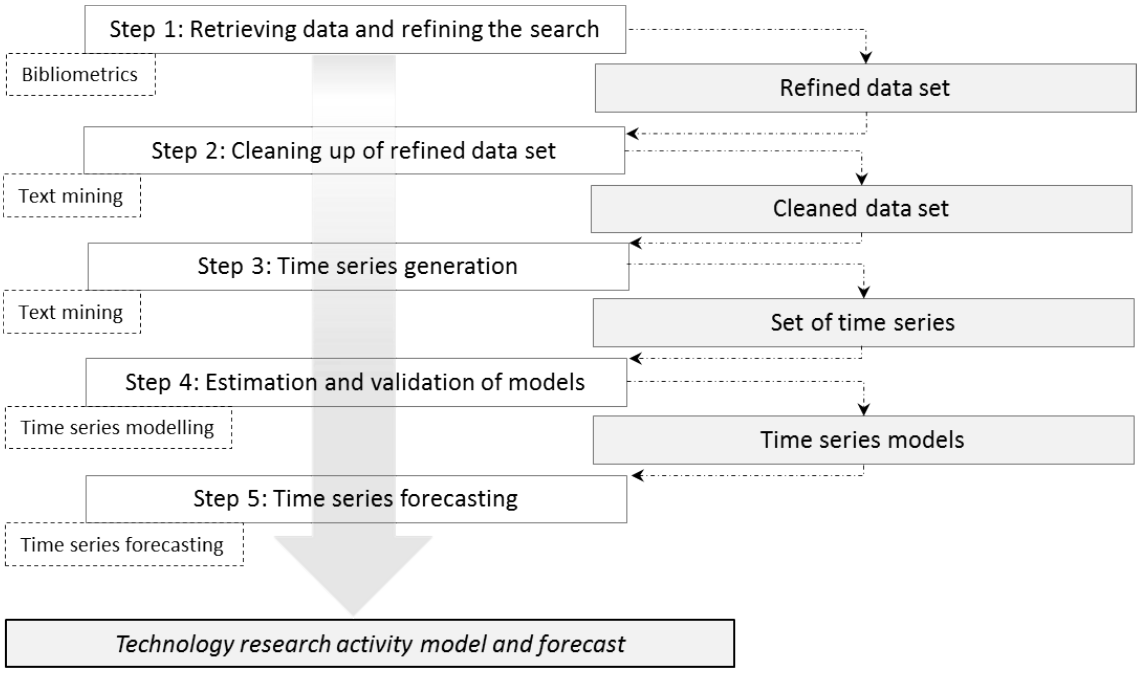
Step 1: Retrieving data and refining the search
Bibliometrics
Refined data set
Step 2: Cleaning up of refined data set
Text mining
Cleaned data set
Step 3: Time series generation
Text mining
Set of time series
Step 4: Estimation and validation of models
Time series modelling
Time series models
Step 5: Time series forecasting
Time series forecasting
Technology research activity model and forecast

Fig. 1 Research approach flow diagram with methodologies and outcomes step by step

Step 1. Retrieving data and refining the search

Bilbiometrics is used to generate a set of structured registers of research publications, whose subject is directly linked to research in the technology to be analysed. The registers retrieved must contain the following fields: title, abstract, publication date (at monthly level) and keywords. To carry out this task Boolean conditions are applied to multidisciplinary online databases to retrieve only those documents which fulfil the following requirements: they must belong to the same research field as the technology considered; their keywords must contain the name of the technology analysed; they must not be reviews of the field; and they must not describe technology applications. The data set created in this step is then cleaned up and exploited in subsequent steps.

Step 2. Cleaning up the refined data set

Once a data set of publications has been obtained, it must be processed so that the information that it contains can be exploited consistently. Thus, the goal of this step is to obtain a cleaned-up, integrated data set. The text mining tool selected to carry out this task is *VantagePoint* software (www.thevantagepoint.com).

The process begins with the importing of all the documents into the text mining tool in the form of structured registers. Then those registers that do not contain full information in the *Title*, *Abstract* and *Publication Date* fields are removed in order to avoid corrupt registers. Subsequently a natural language processing (NLP) method is applied to the *Abstract* and *Title* fields to obtain a list of terms representing topics related to the technology. These terms are combined with those contained in the *Keywords* field to generate a final list of terms sorted by frequency of appearance which represent the topics analysed in the document set. Finally, this list is treated with the fuzzy logic functionality to group into a single term all those terms which have equivalent meanings but are not written in exactly the same way.

This list of terms is used to identify the main sub-technologies that have structured the evolution of the technology and generate the corresponding time series in the following steps.

Step 3. Generating time series

The goal of this step is to generate time series which describe the evolution of the technology. These time series are a sequence of values that reflect the frequency of occurrence of certain terms of interest. To that end it is first necessary to identify the sub-technologies and the specific terms that describe them, and then to generate a monthly time series for each sub-technology based on the count and sum of the frequencies of its specific terms.

In order to identify the sub-technologies and their specific terms, factor analysis (FA) techniques are applied to the list of terms generated in the previous step. This reduces the initial list of terms (variables) to a number of *factors*, which are groupings of terms that frequently appear together within the documents (for further analysis of the application of FA to text mining see Kongthon 2004). Thus, these factors can be treated as sub-technologies, as they are a set of terms that represent concepts that appear together repeatedly in research publications.

To generate monthly time series, the list of terms generated in the previous step must be classified by months. To that end the registers in the full data set are split into different sub-data sets where those documents whose month and year coincide exactly are grouped together. This form of organisation divides the initial list into sub-lists that correspond to each month of the sample.

Finally, a counting process is applied to generate the time series. For example, if *sub-technology_1* is composed of three terms (*term_1, term_2 and term_3*) and these terms occur 3, 2 and 5 times respectively in the list of terms for a specific month of a given year, the value of the time series for that point in time is the sum of those frequencies: 10. A time series representing the evolution of each sub-technology is generated using this process.

Step 4. Estimation and validation of models

Once the time series have been generated, the objective is to identify their most important features and components and to construct an appropriate Structural Times Series Model for forecasting. An STSM is formulated directly in terms of unobserved components:

$$Y_t = \mu_t + \psi_t + \gamma_t + \varepsilon_t \qquad t = 1, 2, \ldots., T \tag{1}$$

where μ_t is the trend, ψ_t the cycle, γ_t the seasonal component and ε_t the irregular component. All these components are formulated stochastically so they can evolve over time according to changes in the data.

This general model (1) encompasses both the additive model when Y_t represents the original data and the multiplicative model when Y_t represents the logarithms. In this paper we do not consider the cycle for these series associated with emerging technologies because the sample is too small to identify such behaviour. The trend represents the long-term component of the series and can be divided into two elements: the level , μ_t, and the slope, β_t. The stochastic formulation of the trend has the following form:

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \qquad \eta_t \sim NID\left(0, \sigma_\eta^2\right) \tag{2a}$$

$$\beta_t = \beta_{t-1} + \zeta_t, \qquad \zeta_t \sim NID\left(0, \sigma_\zeta^2\right) \tag{2b}$$

where η_t and ζ_t are error terms independent of each other that enable the level and the slope to change over time stochastically. If $\sigma_\eta^2 = \sigma_\zeta^2 = 0$ the stochastic trend becomes a deterministic trend with the form $\mu_t = \mu_0 + \beta t$. The seasonal component pattern is formulated stochastically so the sum of seasonal components over a year is a zero mean random variable:

$$\gamma_t = \sum_{j=1}^{s-1} \gamma_{t-j} + w_t, \qquad w_t \sim NID(0, \sigma_w^2) \tag{3}$$

where *s* represents the total number of seasons. If $\sigma_w^2 = 0$ the seasonal component is deterministic, i.e. equivalent to including seasonal dummies in model (1). The irregular component is a Gaussian white noise process, $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$, and the disturbances ($\eta_t, \zeta_t, w_t$ and ε_t) are mutually independent.

Additionally, these models could include exogenous variables such as explanatory variables or interventions to model outliers and structural breaks. The intervention at period $\mathrm{t} = \mathrm{t}_0$ is defined as an impulse variable:

$$\mathrm{I}_{\mathrm{t}_0} = \begin{cases} 0 & \mathrm{t} \neq \mathrm{t}_0 \\ 1 & \mathrm{t} = \mathrm{t}_0 \end{cases} \tag{4}$$

and it can be used to deal with an outlier by introducing it directly in model (1) or to deal with a structural break by adding it in the level equation (2.a):

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \mathrm{I}_{t_0} \lambda_{t_0}^{\mu} + \eta_t \tag{5}$$

where $\lambda_{t_0}^{\mu}$ is the coefficient that measures the jump in the level trend.

The modelling strategy starts with an analysis of the characteristics of the time series to select the components to be included. The unknown coefficients of the model are estimated by maximum likelihood based on the prediction error decomposition computed using the Kalman filter. The model selected is validated by means of the usual set of diagnostic tests (non-normality, heteroscedasticity and serial correlation), which are based on standardized residuals. STSMs can be estimated and validated using the specific STAMP 8.2 software (Koopman et al. 2009).

Step 5. Forecasting time series

Once the model is validated it can be used to estimate and project the trend component, which is the element that will be used as a proxy for the evolution of research activity over the period considered.

The estimation of the components using all the information in the sample is called signal extraction. It can be performed using smoothing algorithms based on the Kalman filter (see Anderson and Moore 1979). In this way, estimates can be obtained of the level of the trend, m_t, and its slope b_t, t = 1, 2, ….,T. These estimates are computed as a weighted average of the observations with more weight given to more recent observations. The smoothing algorithms provide estimates of both the components and their Root Mean Square Errors (RMSE).

The forecast function of the local linear trend (2) is given by

$$m_{T+h} = m_T + b_T\, h, \quad h = 1, 2, 3\ \ldots \tag{6}$$

where m_{T+h} is the forecast of the trend *h* periods ahead and m_T and b_T are the estimates of the level and slope of the trend at the end of the sample. These forecasts along with their RMSE are computed using STAMP 8.2.

All this information is used to identify which sub-technologies are expected to dominate the field in the future and which ones will decrease their participation in the technology's future evolution. It also provides an accurate picture of that evolution.

Data

As mentioned, the technology whose research activity is modelled and forecasted here is cloud computing technology (CC). One of the most widely used definitions of CC is that of the NIST (Mell and Grance 2011), which states that "CC is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction". CC services are grouped into three models (Miller and Veiga 2009): Infrastructure as a Service (IaaS), which comprises a flexible infrastructure of distributed data centre services connected via Internet style networks; Platform as a Service (PaaS), middleware which provides application services and/or runtime environment for cloud applications; and Software as a Service (SaaS), which is made up of top-layer applications delivered on demand.

This technology is already impacting current businesses and is expected to increase its presence in all aspects related to information and communication technologies (see the study conducted by Cisco Systems (2016) for information about this impact). Moreover, CC is a perfect example of an emerging technology as it has the potential to play a major role in addressing inefficiencies and make a fundamental contribution to the growth and competitiveness of organisations, especially SMEs (Sahandi et al. 2013). Thus, any attempt to increase awareness of this technology will result in benefits for many companies and facilitate its implementation.

The raw data for the first step of our procedure was generated using the Web of Science database (WOS). Other multidisciplinary databases were considered, but WOS was the only one that provided monthly publication date information. The time period considered was from 2010 to 2015. A search conducted using the set of Boolean conditions resulted in a data set of 2649 entries.

In the second step, once the NLP analysis had been applied to the *Abstract* and *Title* fields and fuzzy logic functionality, a complete list of descriptive terms was generated for CC technology. Factor analysis was applied to this list of terms (step 3) in order to identify the most important factors and the terms included in them, which were taken as sub-technologies within this framework. Table 1 shows the resulting sub-technologies and the terms included in them. It should be noted that this result is directly derived from the factor analysis. Although more detailed sub-technologies may be considered, the aim of this approach is to rely as far as possible on the automatic results given by the tools used in each step. Thus, these sub-technologies provide a static representation of CC technology, as they represent its most important fields of research. Together with this, the complete description of the technology is based on the detailed analysis of their trends over time via time series generation. Applying the counting process described in step 3, a monthly time series for each technology is generated from January 2010 to December 2015.

Table 1 Factor analysis results, CC sub-technologies and constituent terms

Sub-technology	Terms
Data processing	Big data, data*, Hadoop, MapReduce, massive data, Monte-Carlo, parallel processing
Energy management	Energy*, green cc, power consumption, power saving, smart power, waste power
Mobile CC	Android, backend*, mobile*, offloading, smartphone
Privacy & security	Access control, authentication, confidentiality, encryption, privacy*, security*, trusted cc
Scheduling	Ant colony algorithm, load balancing, scheduling*, neural networks, particle swarm optimisation

Note: * The sub-technology includes a number of terms based on this word.

Fig. 2 shows the time series plots for each sub-technology. It can be seen that the variance of the series increases with their level, suggesting that a multiplicative STSM model would be appropriate. Therefore, the series are transformed into logarithms to work with a log-linear model. All series present a trend that grows rapidly in the early years and then stabilises at the end of the sample. Therefore, the specification of the model should include a local linear trend (2). In some series highly irregular behaviour is observed in the early years: see for example the jump in the level of the Mobile CC series in 2013. However, since we were mainly interested in extracting the trend, we opted to leave the decision on the inclusion of interventions in the model until the model validation step.

Visual analysis of the graph in Fig. 1 and the very nature of the series do not point to the existence of seasonality. However an autocorrelation analysis was performed on the detrended series to detect any significant values in the autocorrelation of order 12. The results show that the Scheduling series is the only one in which a seasonal component was detected. The model for that series therefore includes a seasonal component as well as the trend. This unexpected seasonality may be due to the presence of annual conferences or regular annual issues of journals, which may impact the number of publications related to the sub-technology.

The trend pattern can also be seen to be quite similar for all series, so multivariate analysis could be relevant in this case. STSM provide powerful methods for analysing multivariate time series. Seemingly unrelated time series equations are designed to handle a set of time series which are not directly related to each other but are subject to similar influences (Fernandez and Harvey 1990) and common factor models enable series to be handled with possible co-integration (Fernández 1990). Due to the small size of the samples worked with here, the multivariate approach (which entails estimating a huge number of parameters) was discarded. This does not affect the objectives of this paper, which focuses on forecasting the trend for each sub-technology rather than on the relationships between them.

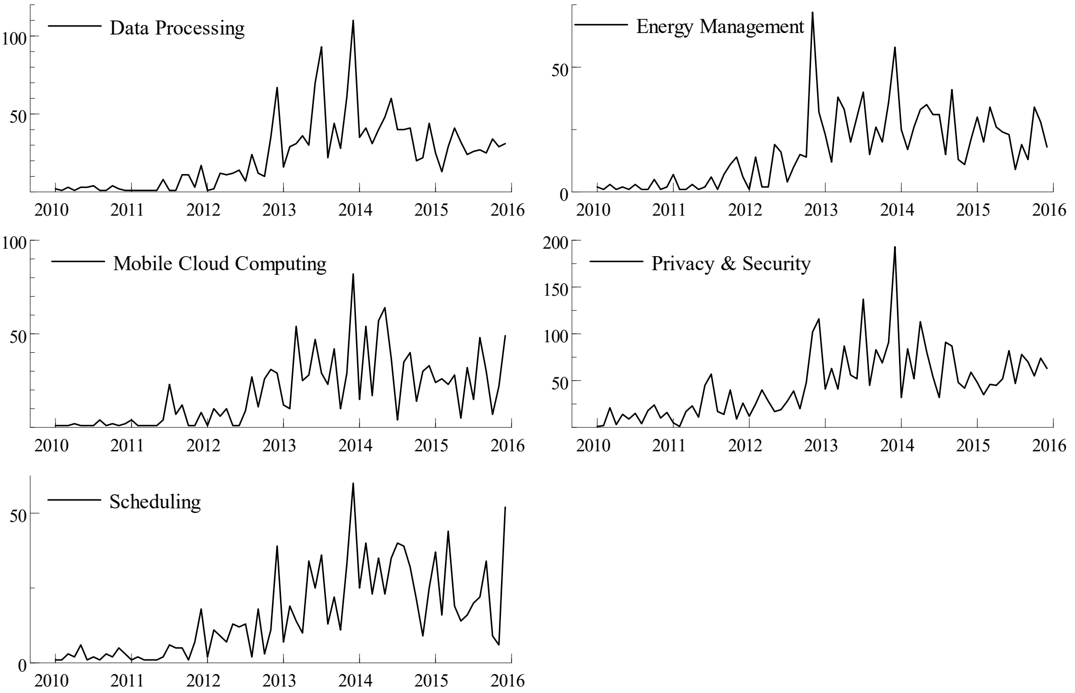

Fig. 2 Time series plots associated with the sub-technologies

Results

Steps four and five of our approach are concerned with identifying the trend component and forecasting it using STSMs. Given that we are mainly interested in the trend component of the series, the specification criterion is to select the simplest model able to represent the main features of the series. As indicated above, the series were modelled in logarithms.

The so-called local linear trend model is estimated for *Data Processing*, *Energy Management* and *Privacy & Security:*

$$Y_t = \mu_t + \varepsilon_t \qquad t = 1, 2, \ldots., T \tag{7}$$

where μ_t follows equation (2).

In the case of the *Mobile CC* series the graphical analysis suggests the same model but with a possible level break around 2012 (Fig. 2), which is confirmed by the residuals in the estimation of model (7). We use the automatic procedure given by STAMP to identify the date and the nature of this break. Based on this, a level break in July 2012 is included. As a result, the appropriate model for the *Mobile CC* series is the local linear trend model (7), where the trend follows equations (2.b) and (5).

Finally, the model for the *Scheduling* series is:

$$Y_t = \mu_t + \gamma_t + \varepsilon_t \qquad t = 1, 2, \ldots., T \tag{8}$$

where μ_t follows equation (2) and γ_t is the seasonal component (3) which is considered deterministic due to the small sample size that we have, i.e. $\sigma_w^2 = 0$.

Table 2 summarises the models formulated for each series and Table 3 contains the estimation results obtained, i.e. the estimated variances of the disturbances for all the components and the estimated coefficient of the level break included in the *Mobile CC* series.

Table 2 Model specification

	Logs	Trend	Seasonal	Interventions
Data Processing	Yes	Local linear	No	No
Energy Management	Yes	Local linear	No	No
Mobile CC	Yes	Local linear	No	Yes, jump (5)
Privacy & Security	Yes	Local linear	No	No
Scheduling	Yes	Local linear	Equation (3) with $\sigma_w^2 = 0$	No

According to the results, for the cases of the *Data Processing*, *Energy Management* and *Mobile CC* series, the estimate of σ_η^2 is zero and the slope is stochastic. This implies that the specific disturbance in the level component (η_t) is not included in equation (2.a), which results in a structural model known as a *smooth trend* model. For the cases of the *Privacy & Security* and *Scheduling* series the best fitting models for the trend include stochastic disturbances for both level and slope components.

Table 3 Parameter estimates

	Data Processing	Energy Management	Mobile CC	Privacy & Security	Scheduling
σ_η^2	0.000	0.000	0.000	0.0017	0.0108
σ_ζ^2	8.04x10^{-4}	3.69x10^{-4}	4.90x10^{-5}	1.1x10^{-4}	2.06x10^{-4}
σ_ε^2	0.386	0.437	0.622	0.418	0.318
$\lambda_{jul2012}^{\mu}$ $(RMSE)$	-	-	1.688 (0.422)	-	-

Once the models are estimated they must be validated in order to support the conclusions to be drawn. Table 4 provides the information related to this validation, which is based on diagnostic test statistics and appropriate measures of goodness of fit. Additionally to the information provided in Table 4, it is worth remarking that the χ^2 statistic of a joint test of significance for the seasonal component in the *Scheduling* model was found to be significant at the 1% significance level ($\chi^2 = 32.23, p\_value = 0.000$).

Table 4 Model validation. Diagnostics and goodness of fit

Time series	Diagnostics on residuals			Goodness of fit	
	Normality (B-S)	Heteroscedasticity H(h)	Serial Correlation Q(12)	Standard Error ($\tilde{\sigma}$)	Coefficient of Determination
Data Processing	7.35 (0.025)	0.140 (0.999)	15.59 (0.112)	0.714	$R_D^2 = 0.359$
Energy Management	0.85 (0.655)	0.277 (0.998)	9.67 (0.470)	0.735	$R_D^2 = 0.433$
Mobile CC	1.51 (0.470)	0.709 (0.792)	13.68 (0.188)	0.825	$R_D^2 = 0.477$
Privacy & Security	18.34 (0.001)	0.154 (0.999)	8.38 (0.592)	0.702	$R_D^2 = 0.380$
Scheduling	0.998 (0.607)	0.813 (0.672)	8.20(0.515)	0.591	$R_S^2 = 0.454$

Note: B-S = Bowman and Shenton (1975) normality test for residuals. H(h)= test for heteroscedasticity, distributed approximately as F(h,h) under the null hypothesis of homoscedasticity (h =19 for Scheduling and h =23 for the rest of the series). Q(12)=Box-Ljung Q-statistic based on the first 12 residual autocorrelations and distributed approximately as χ_q^2 (q = 9 for Scheduling and q = 10 for the rest of the series). The standard error is the square root of the prediction error variance. R_D^2 is the coefficient of determination based on first differences and R_S^2 is based on first differences around seasonal means. *p-values* are given in parentheses.

The results of the diagnostic tests are satisfactory in general. None of the models shows evidence against the homoscedasticity and serial uncorrelation hypotheses. Only the residuals of the *Privacy & Security* series show statistical evidence against the normality assumption at the 1% level. This non-normality can be attributed to two large residuals that modify the tails of the distribution, moving them away from normality. Finally, the coefficients of determination show that these models seem to fit well for the time series.

Once the models are validated they can be used to estimate the trend component at any moment in time (m_t, b_t) and to forecast. It should be noted that in the log-linear models the slope can be interpreted as the monthly rate of growth of the trend. The graphs in Fig. 3 show the estimated elements of the trend: the right column the percentage of growth, $100b_t$; and the left column, the level of trend obtained using anti-log analysis.

All this information enables us to draw some interesting conclusions about the evolution of research activity per sub-technology. When it comes to analysing the intensity of research activity, the trend component provides an image which is not distorted by variations caused by various factors, such as irregular or seasonal movements of the series. In this sense, it can be said that *Privacy & Security* is the sub-technology with the most intense research activity throughout this period within CC technology. It peaks at the beginning of 2014 and holds that same level until the end of the sample. Some major differences must be noted in the development of the rest of the sub-technologies considered. The *Energy Management* and *Data Processing* series follow similar paths: their activity starts to take off early in 2012, with yearly rates of growth above 100%, peaking in the final months of 2013 and then starting to decrease slightly. The only difference between them is observed at the end of the sample, when the rates of growth of *Data Processing* seem to be recovering faster than those of *Energy Management*. Nevertheless, it must be borne in mind that neither of these growth rates is statistically significant at the end of the sample.

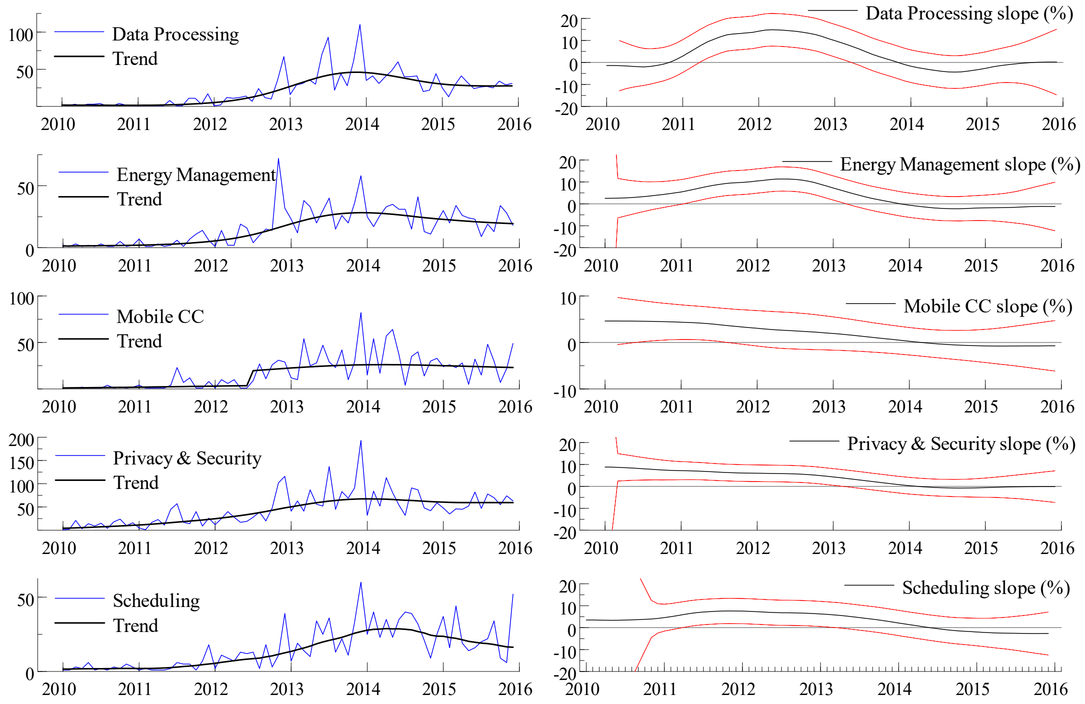

Fig. 3 Estimated trend and slope (with 2 standard errors) for each sub-technology

Mobile CC shows quite high growth rates at the beginning of the sample and a sharp increase in the level modelled by a level break in July 2012. It is as if interest on this sub-technology suddenly increases among the scientific community. However, after peaking at the beginning of 2014 activity related to this sub-technology also starts to decrease and the slope's final value is negative, though not statistically significant. Finally, the *Scheduling* series shows the longest period of growth: activity starts to grow at the beginning of 2012 (like the rest of the series) but does not peak until the middle of 2014. As a result its peak intensity is reached later than those of the other sub-technologies. Nonetheless, once the level of this series starts to decrease it does so faster, as if the increased interest is not strong enough to generate consistent research into it. This can be easily detected in the slope's final value, which shows a yearly decrease of 32.12%.

All these facts are linked to the forecasting results. Fig. 4 shows the forecasts of the trend component for 2016 (see equation (6)). Table 5 provides the following information: the first row displays the estimates for yearly percentage growth computed from the monthly growth rates estimated at the end of the sample and the following rows show trend forecasts for each month of 2016.

This information gives us an accurate picture of the future evolution of the research activity related to each sub-technology which supplements the previous analysis of the evolution of the trend. Based on the estimated yearly rates of growth, it can be concluded that research activity will decrease in the coming months for all sub-technologies except *Data Processing*. With respect to the level of activity of each sub-technology, it can be observed that *Privacy & Security* will continue to be a highly active research field, as it is forecast to maintain its activity levels in the near future. Without generating such a high level of activity but still showing an increase in the trend in the last period of the sample, *Data Processing* is the other sub-technology which will not lose interest on the part of the research community in the near future. By contrast, *Mobile CC*, *Energy Management* and *Scheduling* are expected to undergo a noticeable decline in scientific activity. This is especially noteworthy in the case of *Scheduling*, which is expected to become a sub-technology with little scientific activity in the last period of 2016.

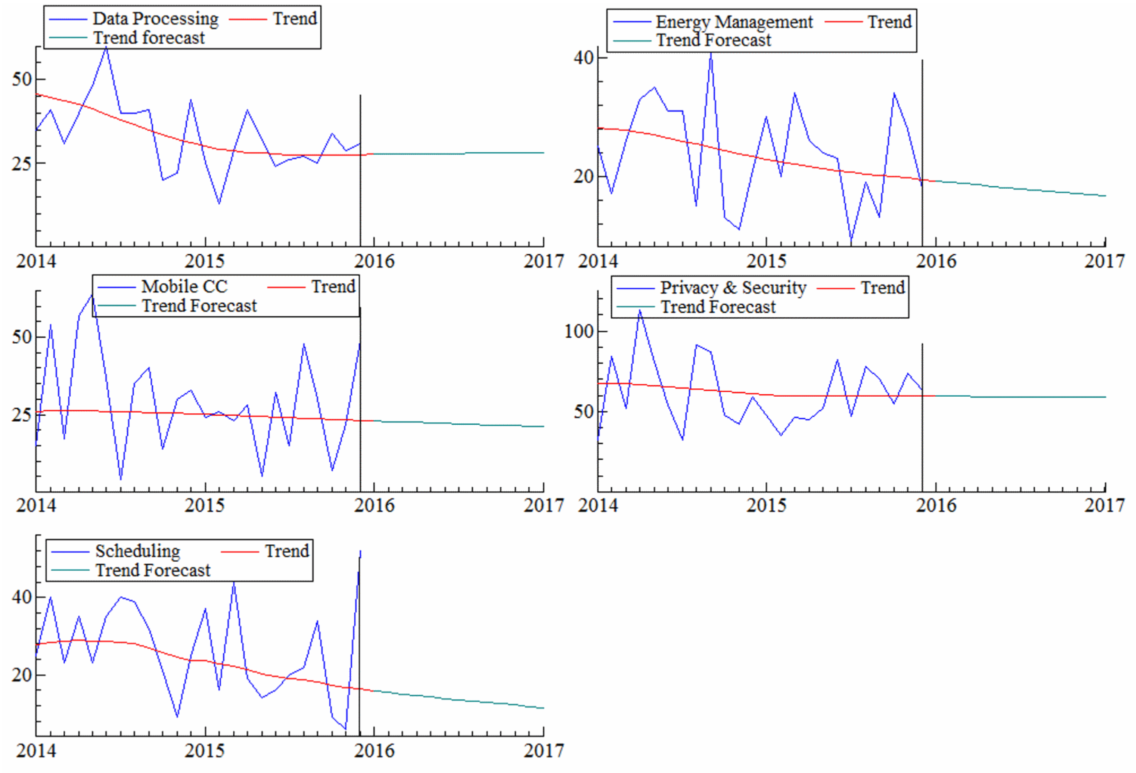

Fig. 4 Trend forecasts for each model 12 steps ahead of the final period (2016/01-2016/12)

Table 5 Trend forecast values for 2016

	Data Processing	Energy management	Mobile CC	Privacy & Security	Scheduling
Yearly % growth at Dec 2015 (p value)	1.967 (0.983)	-14.136 (0.832)	-8.571 (0.793)	-1.196 (0.978)	-32.12 (0.587)
			Monthly forecasts		
2016/01	27.70	19.19	22.94	59.45	15.89
2016/02	27.74	18.96	22.78	59.39	15.47
2016/03	27.79	18.74	22.61	59.33	15.06
2016/04	27.83	18.52	22.45	59.27	14.66
2016/05	27.88	18.30	22.29	59.21	14.28
2016/06	27.92	18.09	22.13	59.15	13.90
2016/07	27.97	17.87	21.98	59.09	13.53
2016/08	28.01	17.66	21.82	59.03	13.17
2016/09	28.06	17.46	21.67	58.97	12.83
2016/10	28.11	17.25	21.51	58.91	12.49
2016/11	28.15	17.05	21.36	58.86	12.16
2016/12	28.20	16.85	21.21	58.80	11.84

Conclusions and future work

This paper presents an approach which contemplates various specific goals, as stated in the introduction. Firstly, it aimed to apply bibliometrics to retrieve specific scientific publications related to an emerging technology and to generate a data set which could be used as raw data to characterise and forecast the scientific activity related to that technology. That data set was then to be comprehensively analysed via text mining for two purposes: to identify the main sub-technologies within the technology (this is achieved by applying factor analysis to the content of specific fields in the documents) and to generate time series associated with those sub-technologies based on monthly frequency in the appearance of terms that characterise them.

Once these time series were generated, the goal was to model them by means of STSMs, which would enable the series to be broken down into their unobserved components, with special attention to the trend component. The STSM for each of the series provides a full description of its characteristics, and the trend of each series is used as a proxy for the intensity of research activity, enabling us to compare the changes over time in the different sub-technologies. Finally, trend forecasts are used to analyse the short-term future of research activity in each sub-technology.

All these tasks yielded valuable information that was used to depict the research activity associated with CC technology. Thus, it may be concluded that *Privacy & Security* had been the most active sub-technology in the scientific community, and that it was expected to maintain this predominance. Other sub-technologies such as *Data Processing* were found to have had less impact among researchers but to be likely to increase their importance in the coming months. Additionally, sub-technologies were found that had sparked some interest within the development of CC technology, such as *Energy Management*, *Mobile CC* and *Scheduling* but had proved unable to maintain it. Research activity in these sub-technologies is expected to decrease in the near future, with the decline being especially sharp for *Scheduling*. This characterisation enables us to identify the most important fields of research in the initial development of CC technology and to anticipate which of them will continue leading the field and which will not.

The results obtained for CC technology enable us to identify the potential of the approach used for describing research activity related to an emerging technology and forecasting its future evolution. It is worth remarking that only five sub-technologies were analysed here because this application is focused on validating the approach. However, merely by applying the steps of the approach to several more sub-technologies an even more thorough description of research activity can be obtained. In addition, it should be noted that the approach can be applied to any type of technology regardless of its nature, provided that it is possible to obtain consistent bibliometric information. Linked to this, the main limitation of the approach is the short length of the series related to emerging technologies, which conditions the whole analysis and limits the horizon of the prospecting.

This paper is limited to describing the research activity associated with an emerging technology. We are aware that the development of a technology is not captured only by bibliometric data on research activity, but one of the advantages of this approach is that it can be integrated into a broader one in which other information sources are considered. Indeed, this approach is framed within a previous paper by the authors (Bildosola et al. 2015) that proposed a broader framework of analysis with two main objectives: first to generate and forecast a research activity profile, and secondly to obtain a complete picture of the technology including other data sources such as web content mining (Cooley et al. 1997) to forecast the technology as a whole. The present paper has provided a solution to the first objective. Thus, future work should concentrate on the second, i.e. the use of other data sources, the construction of a broad database that combines all the information generated and the integration and representation of this information using a single graphic element by means of technology roadmapping methods.

References

Anderson, B.D.O., & Moore, J. B. (1979). *Optimal filtering*. Englewood, NJ: Prentice-Hall.

Bengisu, M., & Nekhili, R. (2006). Forecasting emerging technologies with the aid of science and technology databases. *Technological Forecasting and Social Change*, *73*(7), 835-844.

Bildosola, I., Rio-Bélver, R., & Cilleruelo, E. (2015). Forecasting the Big Services Era: Novel Approach Combining Statistical Methods, Expertise and Technology Roadmapping. In *Enhancing Synergies in a Collaborative Environment* (pp. 371-379). Springer International Publishing.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual review of information science and technology*, *37*(1), 179-255.

Bowman, K. O., & Shenton, L.R. (1975). Omnibus contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, 62, 243-250.

Brunk, G. (2003). Swarming of innovations, fractal patterns, and the historical time series of US patents. *Scientometrics*, *56*(1), 61-80.

Chen, Y. H., Chen, C. Y., & Lee, S. C. (2011). Technology forecasting and patent strategy of hydrogen energy and fuel cell technologies. *International Journal of Hydrogen Energy*, *36*(12), 6957-6969.

Cisco, C. V. N. I. (2015). Global Mobile Data Traffic Forecast Update. 2014–2019 (white paper).

Cooley, R., Mobasher, B., & Srivastava, J. (1997, November). Web mining: Information and pattern discovery on the world wide web. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on* (pp. 558-567). IEEE.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, *73*(8), 981-1012.

Dilaver, Z., & Hunt, L. C. (2011). Modelling and forecasting Turkish residential electricity demand. *Energy Policy*, *39*(6), 3117-3127.

Engle, R. F. (1979). Estimating structural models of seasonality. In A. Zellner (Ed.) *Seasonal analysis of economic time series* (pp. 281-308). Washington D.C.: U.S. Bureau of the Census.

Fernández, F. J (1990). Estimation and Testing of a Multivariate Exponential Smoothing Model in the Frequency Domain", *Journal of Time Series Analysis*, 11, 89-105.

Fernández, F. J., & Harvey, A. C. (1990). Seemingly unrelated time series equations and a test for homogeneity. *Journal of Business & Economic Statistics*, *8*(1), 71-81.

Fernández-Cano, A., Torralbo, M., & Vallejo, M. (2012). Time series of scientific growth in Spanish doctoral theses (1848–2009). *Scientometrics*, *91*(1), 15-36.

Ghosh, B., Basu, B., & O'Mahony, M. (2009). Multivariate short-term traffic flow forecasting using time-series analysis. *IEEE Transactions on Intelligent Transportation Systems*, *10*(2), 246-254.

Gonzalez, P., & Moral, P. (1995). An analysis of the international tourism demand in Spain. *International Journal of Forecasting*, *11*(2), 233-251.

Harvey, A. C. (1989). *Forecasting, Structural Time Series and the Kalman Filter*, Cambridge: Cambridge University Press.

Jun, S., & Uhm, D. (2010). Technology forecasting using frequency time series model: Bio-technology patent analysis. *Journal of Modern Mathematics and Statistics*, *4*(3), 101-104.

Kajikawa, Y., Yoshikawa, J., Takeda, Y., & Matsushima, K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change*, *75*(6), 771-782.

Kitagawa, G., & Gersch, W. (1984). A smoothness priors–state space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, *79*(386), 378-389.

Kongthon, A. (2004). A text mining framework for discovering technological intelligence to support science and technology management (Doctoral dissertation, Georgia Institute of Technology).

Koopman S.J., Harvey A.C., Doornik J.A., & Shephard, N. (2009). *STAMP 8.2: Structural Time Series Analyser, Modeler, and Predictor*. London: Timberlake Consultants.

Kostoff, R. N., & Geisler, E. (1999). Strategic management and implementation of textual data mining in government organizations. *Technology Analysis & Strategic Management*, 11(4), 493-525.

Martin, H., & Daim, T. U. (2012). Technology roadmap development process (TRDP) for the service sector: A conceptual framework. *Technology in Society*, 34(1), 94-105.

Martino, J. P. (2003). A review of selected recent advances in technological forecasting. *Technological Forecasting and Social Change*, *70*(8), 719-733.

Mell, P. & Grance, T. (2011). The NIST definition of cloud computing. National Institute of Standards and Technology Special Publication 800-145, U.S. Department of Commerce.

Miller, H. G., & Veiga, J. (2009). Cloud computing: Will commodity services benefit users long term?. *IT Professional Magazine*, *11*(6), 57.

Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5), 431-443.

Pagan, A. (1975). A note on the extraction of components from time series. *Econometrica*, 43, 163-168.

Porter, A. L., Watts, R. J. & Anderson, T .R. (2005). Mining PICMET: 1997-2005 Papers Help You Track Management of Technology Developments. *Portland International Conference on Management of Engineering and Technology (PICMET)*, Portland, OR.

Proietti, T. (2002). Forecasting with structural time series models. In P. Clements & D.F. Henry (Eds.) *A Companion to Economic Forecasting*, 105-132. Malden, MA: Blackwell.

Sahandi, R., Alkhalil, A., & Opara-Martins, J. (2013). Cloud computing from SMEs perspective: a survey based investigation. *Journal of Information Technology Management*, *24*(1), 1-12.

Schlink, U., Herbarth, O., & Tetzlaff, G. (1997). A component time-series model for SO 2 data: Forecasting, interpretation and modification. *Atmospheric Environment*, *31*(9), 1285-1295.