Chapter 3

From theory to time series*

Peter Hingley[1] and Walter G Park[2],

[1] Financial Controlling and Statistics, European Patent Office, Munich, Germany
[2] Department of Economics, American University, Washington, DC, US

1 Introduction

In order to develop forecasting models for numbers of patent filings, it is important to consider that patents protect more than just intellectual property; since they are an intrinsic component of the larger economic picture. This occurs through the process of innovation, technological and scientific change, economic productivity and growth. In this chapter, we will show one particular theoretical formulation of the underlying process that leads to patenting. Then we will briefly review the various types of time series based regression analysis that are available and show how they could be used to fit a model within a framework of econometric methodology that uses techniques of cointegration and error correction. Various approaches to time series could be used and this book necessarily concentrates on only some of them.

2 A theoretical model

This section develops a conceptual model of international patenting flows as a basis for the empirical analysis. The following model is adapted from Eaton and Kortum (1996), Kortum and Lerner (1998), and Park (2001). In this approach, a decision-theoretic model of patenting is formulated where inventors weigh the costs and benefits of filing for a patent. Other, more macro level, approaches are possible; for example, Abdih and Joutz (2005) use time-series methods (including cointegration analyses, see Sect. 3.1) to

* Frederick Joutz and Robert Trost contributed to Sect. 3.1 during the preliminary phase of the research programme.

derive a knowledge production function which relates patent filings to the past stock of filings and to the number of R&D scientists and engineers.

First some definitions and notation are in order. Let the source country be the country of origin of patent applications (or the country to which patents are granted). Let the destination country be the country in which a patent application is filed (or the country which grants a patent right). For now, we are primarily interested in the special case where the destination is not a specific country but a regional office (such as the EPO).

The variations in international patenting depend on three kinds of heterogeneity: (1) market heterogeneities (some destinations are more attractive than others); (2) invention heterogeneities (some inventions are more valuable than others); (3) heterogeneity between source countries (some source countries are more inventive than others).

Let the source country be indexed by i and the destination country by j, such that i = j refers to domestic patent applications, and $i \neq j$ to foreign patent applications. Let P_{ij} denote the number of patent applications from country i to country j. Each source country produces each period a flow of inventions. Let α_i denote this flow of ideas, some of which may be patentable. Of these, some fraction, f_{ij}, is applied for a patent in country j[1]. Hence the number of patent applications from country i in country j is:

$$P_{ij} = \alpha_i f_{ij} \tag{1}$$

We will model each of the three components on the right hand of this equation. First, we assume inventive output to be produced according to a linearly homogeneous production function.

$$\alpha_i = \alpha\ R_i^{\beta_1} S_i^{\beta_2} L_i^{1-\beta_1-\beta_2}, \tag{2}$$

where $\alpha > 0$ and R denotes the stock of research and development (R&D) capital, S the supply of scientists and engineers, and L labor. The exponents (β_1, β_2, . . .) represent elasticities: the percentage change in inventive output for a 1% change in input.

It will be shown that those inventions which cross a particular quality threshold will be the ones for which patents are sought in the destination market. To develop this idea, assume that each of the various inventions of the source country can be indexed by its “quality” level, associated say with the inventive step of an invention. The quality level is assumed to be

[1] Of course, this is not to say that they are "patentable" – only that patent applications are filed for them. Whether inventions qualify or meet the standards of patentability (as set out in country j's laws) is determined at the patent granting phase.

a random variable, Q, drawn from a negative exponential distribution. Let the cumulative distribution function be as follows.

$$F(q) = \Pr (Q < q) = 1 - \exp(-\psi q) , \tag{3}$$

This essentially captures the stylized fact that the distribution of invention quality is skewed: a small percentage of the top inventions account for a large majority of the total value of patent rights.[2] It can be seen from the fact that $F'(q) > 0$ and $F''(q) < 0$. Thus the median quality is less than the mean quality (hence the distribution of Q is positively skewed – or skewed to the right).

The mean inventive step from such a distribution is $1/\psi$. An invention of size q is assumed to augment a firm's productivity by a factor of exp(q). For example, if A is an index of productivity, the new level of productivity would be $A' = \exp(q) A$. Thus, under this formulation, q is the growth rate of productivity. Firm productivity could be enhanced either because the invention improves production potential or is a cost-saving innovation. We assume that the productivity increase is reflected in firm profits. A tractable formulation is the following.

$$\pi = \pi(q) = \exp(q)\, \pi_0 , \qquad \text{where} \quad \pi'(q) > 0 ,$$

and π denotes the instantaneous flow of profits and π_0 some base flow.

The derivative of π with respect to q under this formulation is of course assumed to be positive. However, it is also necessary to adjust the level of a firm's profits due to imitation activities. Because of imitation or infringement, the returns to inventions are not fully appropriable. In each market or destination j, firms face hazards of imitation. Assume that imitation acts like a tax on profits, and denote by h the rate at which profits can actually be appropriated. Thus net instantaneous profits are

$$\pi = h \exp(q)\pi_0 \tag{4}$$

Where $0 \leq h < 1$. Since $h < 1$, this signifies that the returns are imperfectly appropriated and a value of $h = 0$ denotes that the returns are completely dissipated. However, with patent protection the ability to appropriate increases, depending on the strength of the intellectual property regime. Let θ denote the increase in the rate of appropriation due to patent protection. θ is assumed to be a positive function of the strength of patent rights[3]. Thus with patent protection, the rate at which returns are appropriated equals

2 See Putnam (1996) for some empirical evidence and review of the literature.

3 If the enforcement of patent rights is completely ineffective or if patent protection is, for some reason, not necessary for the appropriation of investment returns, then θ would be zero.

$h + \theta$, where $0 \leq h + \theta < 1$[4]. The value of a firm equals the presented discounted value of the future stream of profits, and depends on whether or not a firm has a patent. With a patent, the value is

$$V^{PAT} = \int_0^\infty (h+\theta)\pi_0 \exp(q)\exp(-rt)dt = \frac{(h+\theta)\pi_0 \exp(q)}{r} \tag{5}$$

where r is the real interest rate.

Without a patent, the value of a firm is the above expression with θ set to zero:

$$V^{NO\ PAT} = \int_0^\infty h\,\pi_0 \exp(q)\exp(-rt)\,dt = \frac{h\,\pi_0 \exp(q)}{r} \tag{6}$$

Hence the value of patent protection is:

$$\Delta V = V^{PAT} - V^{NO\ PAT} = \frac{\theta\pi_0 \exp(q)}{r} \tag{7}$$

That is, a patent in market j enables the patentee to *purchase* a reduction in the incidence of imitation, the benefit from which is reflected in an increase in firm value. Thus, a firm will seek patent protection if the net benefit of patenting exceeds the cost of filing for protection.

$$\Delta V = V^{PAT} - V^{NO\ PAT} \geq c \tag{8}$$

where c denotes the cost of obtaining a patent (e.g. filing fees, agent fees, and possibly translation fees).

The underlying logic is that inventors have means other than patent protection to appropriate the rewards from their innovation (such as lead times, reputation, secrecy). Thus the value of a patent is the incremental return an inventor can get above and beyond what can be realized by alternative (non-patenting) means.

Equation (8) helps determine which of the source country inventions will be applied for patent protection. A critical threshold quality of inventions can be identified using (7) and (8), namely:

$$q^* = \ln\left(\frac{r\,c}{\theta\pi_0}\right) \tag{9}$$

---

[4] Note that the increase in appropriability could have been modeled multiplicatively (as $h\,\theta$). However, the results are qualitatively similar but analytically less tractable.

The more expensive it is to file a patent (i.e. the higher c is), the higher the quality threshold (indicating that only inventions of higher quality are worth patenting). Furthermore, the stronger the patent regime (i.e. the higher θ is), the lower the quality threshold. Thus, not surprisingly, patent rights are more valuable, holding other factors constant, if patent protection is stronger. A higher base flow of profits (π_0) also contributes to a lower critical threshold quality. Firms that in general face larger markets (or produce goods and services that the destination market more highly values) are likely to have a higher base flow of profits.

The number of patents filed can now be determined. Recall the cumulative distribution function F(q). Given a critical threshold quality q*, F(q*) is the fraction of source country inventions that are not patented and $1 - F(q^*) = \exp(-\psi q^*)$ is the fraction that are. Thus, using our notation above, the third term in Eq. (1) is:

$$f_{ij} = \exp(-\psi q^*) = \left(\frac{r\, c_j}{\theta_j \pi_j}\right)^{-\psi} \tag{10}$$

The subscript j has been brought back to clarify that it is the filing cost and the strength of patent rights of the destination that matter. The base flow of profits π_0 has been renamed π_j to indicate that the profits would be derived from exploiting inventions in the destination market. Note that while these profits are derived from the destination market, they accrue to the firm in source country i. Now, putting it all together by substituting (2) and (10) into (1), yields the prediction for patenting flows from i to j:

$$P_{ij} = \alpha\, R_i^{\beta_1}\, S_i^{\beta_2}\, L_i^{1-\beta_1-\beta_2} \left(\frac{rc_j}{\theta_j \pi_j}\right)^{-\psi} \tag{11}$$

Taking natural logs of both sides of (11) yields:

$$\ln\left(\frac{P_{ij}}{L_i}\right) = \gamma_0 + \gamma_1 \ln\left(\frac{R_i}{L_i}\right) + \gamma_2 \ln\left(\frac{S_i}{L_i}\right) + \gamma_\pi \ln\left(\pi_j\right) + \gamma_\theta \ln\left(\theta_j\right) - \gamma_c \ln\left(c_j\right) + \mu \tag{12}$$

where μ denotes the error term[5].

Empirical measures of these variables are available; for example, data on research and development expenditures and science and engineering

[5] Furthermore, the parameters in (11) are functions of previous parameters; i.e. $\gamma_0 = \ln \alpha - \psi \ln r$, $\gamma_1 = \beta_1$, $\gamma_2 = \beta_2$, $\gamma_\pi = \gamma_\theta = \psi$, and $\gamma_c = -\psi$.

personnel can be used for R and S respectively. Patent filing, translation (if any), and search costs can be used for c. An index of patent strength can also be used for θ.[6]

In Eq. (12), the base flow of profits, π_j, depend on the characteristics of the destination market. For example, the size of the destination market (e.g. the markets of the member countries that comprise the EPO) should influence the profitability of commercializing innovations. As a measure of the market size, the real gross domestic product (GDP) of the destination could be employed. For example, the GDPs of the European Patent Convention (EPC) contracting states could be summed (each year) to provide a measure of the market size of the EPC area. Another modification to (12) is to add time subscripts since data on international patenting are available over time. Finally, given the panel data nature, the error term may consist of important individual (country) effects, which may be fixed or random. The individual effect may capture any specific interaction effect that individual source countries have with the EPO.

Thus the equation to be estimated is:

$$\ln\left(\frac{P_{ijt}}{L_{it}}\right) = \gamma_0 + \gamma_1 \ln\left(\frac{R_i}{L_i}\right) + \gamma_2 \ln\left(\frac{S_i}{L_i}\right) + \gamma_\pi \ln\left(GDPC_{jt}\right) + \gamma_\theta \ln\left(IPR_{jt}\right) - \gamma_c \ln\left(c_{jt}\right) + \mu_{it} \tag{13}$$

The error term is motivated by the fact that some profitable inventions fail to be patented while some unprofitable ones are patented. Note that, in (13), j refers to the EPO. Thus, the dependent variable is the natural log of EPO patent applications from country i divided by the labor force in country i.

Equation (13) constitutes a basic model specification for explaining patenting behavior in the EPO. The underlying premise is that inventive and patenting behaviors respond to economic incentives (due to market size, patenting costs, and property rights) and to technological capabilities (such as R&D resources and productivity).

3 Time series regression based approaches

Variants of the model that is developed in Eq. (13) are discussed in Chap. 7, where a time series approach is used to fit the models to available data sets on EPO filings. However, in this book there are several other con-

[6] See, for example, Ginarte and Park (1997).

tributions involving time series analyses, and we would now like to introduce methods that are in principle available.

The problem involves setting up a regression analysis based system that is consistent with an economic formulation of the problem. In an empirical regression setting, it is typical that the series to be analysed contains a rather short set of data and so estimation can proceed for only a severely limited number of parameters. It is unlikely that the full economic formulation can be realised in the analysis, since parameters may be essentially collinear and indistinguishable. This is the question of identifiability of a regression model and the basic rule of thumb is to take the simplest alternative that is adequate for the data.

In the self determining approach, the historical development of the patent series itself is used to project the future trend. Usually this is done within the framework of Box-Jenkins based ARIMA methods, state space models or vector autoregression (VAR) models. Chapter 4 introduces these techniques and they appear again in later chapters.

In the predictive approaches, concomitant series of other variables are used to explain the movements in the patent series. Extensions of the above mentioned techniques are generally available to cope with this situation. Chapters 5 to 7 deal with these kinds of models.

Typically, series of pooled national research and development expenditures (R&D) or gross domestic product (GDP) data are used as predictors. Where a model can be developed in which these predictor series act on patents with a lag, forecasts can then be made for the future that are based on predictor values available today. Otherwise, the forecasting problem may just be transferred from one of forecasting the patent series to one of forecasting the values of the concomitant variables. This can in fact make sense if the concomitant variable is something important, such as GDP, where considerable forecasting efforts are available from outside agencies for different purposes.

As with more straightforward linear regression modelling in a conventional statistical setting, simple approaches such as the self determining models can be formulated with ordinary algebra while more complex multivariate models are manageably represented using matrix and vector expressions.

These approaches can be used pragmatically, that is without worrying too much about the underlying economic mechanism. Indeed, for the limited problem of forecasting future numbers of total patent filings, a mechanism may not be needed if the forecasting performance is sufficiently high.

There are of course many developments in time series that have been ignored in the above description, e.g. Nonlinear time series regression methods (Tong, 1990). However we suggest that the methods should be kept simple unless evidence emerges that a more complicated formulation is required.

3.1 Econometric methodology – cointegration, vector autoregression, error correction models and count data analysis

Now, we would like to look a little closer at ways that the economic process can be taken into account. The following suggestions involve the VAR methodology but also involve cointegration analysis, error correction models and count data models that otherwise received little attention within the projects that were carried out in the research programme.

The general-to-specific econometric modelling approach advocated by Hendry (1986, 1993, and 1995) can be used in the modelling and forecasting of patent application filings. However, to be consistent with our stated aim in the last section to keep modelling simple, we suggest that even the initial general specification should not be too elaborate. The general-to-specific modelling approach is a relatively recent strategy used in econometrics. It attempts to characterize the properties of the sample data in simple parametric relationships that remain reasonably constant over time, account for the findings of previous models, and are interpretable in an economic and financial sense. Rather than using econometrics to illustrate theory, the goal is to “discover” which alternative theoretical views are tenable and test them scientifically. The general-to-specific approach starts with over-parameterized models (in terms of lags and variables) and through specification tests reduces to a parsimonious representation.

Consider for example a plan to develop models to describe patenting at the EPO by applicants from various countries using this approach. The initial models can rely on a database that would at a minimum include the filings from various countries at the EPO, domestic filings, R&D measures, and real GDP. There may be a long-run relationship in the level or accumulated R&D effort and patenting activity. Basic scientific developments and technological progress can lead to important or break-though innovations, which produce large flows of patent activity and filings. Also, the size of the domestic market for innovation can influence the propensity to file a patent at the EPO. The mechanism of the patent system, which involves a priority forming first filing at one office that is followed by subsequent filings at other offices, suggests that there is a lag between the domestic filings and subsequent filings at the EPO. Real GDP captures the demand for new products, processes, and services. In the short run there can be another relationship between year-to-year changes in R&D effort, which provides incremental technological improvements, and modifications that affect patent application filings in the short run. In the next or final stage the national level models can be combined in either a panel data framework or in a system of equations.

Once the variables have been selected, the process begins with examining the time series properties of the data. The (statistical) implication of a series with a unit root is that it is non-stationary or integrated of order one, I(1). There is no tendency towards a mean value and the standard error is not defined. We can look at the change of these series or their growth rates and these will be stationary with a defined mean and standard error. Simple univariate forecasting models can be developed based on the time series properties. These models can be used as a comparison for the econometric models. In some cases, macroeconomic series with unit roots tend to “trend” together or are cointegrated. They might not synchronously rise and fall every year, but in the long run they move together. Examples of this include exchange rates and interest rate differentials, consumption expenditures and (disposable) income, levels of sales and inventories or capital stock, short run and long-run interest rates, and patenting activity or filings and R&D effort. These examples represent equilibrium relationships that are consistent with economic theory and intuition. The standard test procedure follows the multivariate cointegration approach of Johansen (1988).

The procedures begin with a multivariate specification. Patent filings can be represented as the flow of new knowledge which is a function of the stock of knowledge derived from previous patents. The labor effort devoted to knowledge production can be represented by R&D effort and real GDP. A preliminary national level VAR model might look like:

$$\begin{bmatrix} EPOFil_t \\ RD_t \\ RGDP_t \\ DOMFil_t \end{bmatrix} = A(0) + A(L) \begin{bmatrix} EPOFil_{t-1} \\ RD_{t-1} \\ RGDP_{t-1} \\ DOMFil_{t-1} \end{bmatrix} + B(L) X_t + e_t$$

where EPOFil are the filings at the EPO, RD is a measure of R&D effort, RGDP is real gross domestic product, and DOMFil are first filings nationally or the size of the domestic innovation market. Patent application filings are an indicator for knowledge production. Country specific subscripts are omitted for simplicity here, but specific models may be developed for each country. A(0) is a set of deterministic variables like constant, trend, or dummies for changes in patenting rules. The specification and interpretation of these variables will be discussed shortly. A(L) is a lag polynomial operator (seee also Chap. 4 and Chap. 6), X are other exogenous variables, contemporaneous and lagged depending on the conformable operator B(L). The term e represents a vector of disturbances with means of zero, constant variance covariance matrix, and serially uncorrelated disturbances.

The long-run properties of trending variables can be used to explain growth or speeds of adjustment to equilibrium over time through an error correction (ECM) model. A special case of the ECM is the partial adjustment model. The ECM captures both the short-run dynamics and the long-term trends in patent filings. The specification is in growth rates or first differences in natural logarithms. This approach avoids the resulting nonsense or spurious regression problem when the data series are I(1). Research by Diebold and Lutz, 2000 and Clements and Hendry, 2000 suggest that the cointegration testing and possibly an error correction model is the appropriate methodology to use for estimation and forecasting purposes.

Exogeneity is an important conceptual and empirical issue for inference analysis, forecasting, and policy analysis in the presence of cointegration. Ericsson and Irons (1994) discuss these issues and the appropriate empirical tests. Joutz and Maxwell (2002) have applied these techniques in modelling and forecasting high yield bonds. The factorization or conditioning of models following these tests suggests the appropriate specification of the forecasting models.

Cointegration tests are a multivariate form of integration analysis. Individual series may be I(1), but a linear combination of the series may be I(0). The error correction model is a generalization from the traditional partial adjustment model and permits the estimation of short-run and long-run elasticities.

An approach to this is based on the findings of Nelson and Plosser (1982), in which many macroeconomic and aggregate level series are shown to be well modeled as stochastic trends, i.e. integrated of order one, or I(1). Simple first differencing of the data will remove the non-stationarity problem, but with a loss of generality regarding the long run "equilibrium" relationships among the variables.

Engle and Granger (1987) solve this filtering problem with the cointegration technique. They suggest that if all, or a subset of, the variables are I(1), there may exist a linear combination of the variables which is stationary, I(0). The linear combination is then taken to express a long-run "equilibrium" relationship. Series which are cointegrated can always be represented in an error correction model. The error correction model is specified in first differences, which are stationary, and represent the short run movements in the variables. When an error correction term is included in the model, the long-run, or equilibrium, relations in patent activity are accounted for. Lags of the independent and dependent variables may be included to capture additional short- and medium-term dynamics of patenting activity. The advantage of the first difference model is that the specification is stationary so that estimation and statistical inference can be performed using standard statistical methods. The contemporaneous coefficients are interpreted as short run elasticities.

The vector(s) obtained in the cointegration analysis represent the long-run relationship among the variables. To model patent filing or activity more generally, however, a short-run error correction model is employed. The error correction framework models the variables in differences, and then the coefficients on the differenced variables correspond to short-run elasticities. The model furthermore contains an error correction term, ECM. This term is obtained from the long-run relationship and expresses deviations in patent filing (knowledge production) from its long-run mean. The coefficient in front of the ECM term measures the speed of adjustment in current consumption to the previous equilibrium demand value. The model in its most general form is as follows:

$$\Delta y_t = \alpha + \Sigma \delta_i \Delta y_{t-i} + \Sigma \beta_i \Delta x_{t-i} + \gamma ECM_{t-1} + e_t \quad i = 1, \ldots, t$$

where y are the dependent variables, x is a vector of independent variables and ECM is the error correction term.

Once these first round models are built, the effort can turn to simultaneous modelling of the patent filings at the EPO. There are several approaches that can be considered. These will depend on data quality, sample size, and results from the preliminary models.

One approach is to use dynamic panel data techniques. In a panel data setting, there are time-series observations on multiple countries. We can denote the cross-section sample size by N, and, in an ideal setting, have t=1,...,T time-series observations covering the same calendar period, a balanced panel. In practice, it often happens that some cross-sections start earlier, or finish later. When T is large, N small, and the panel balanced, it is then possible to use the simultaneous-equations modelling procedures. If T is small, and the model is dynamic (i.e. includes a lagged dependent variable), the estimation bias can be substantial (see Nickell, 1981). Methods to cope with such dynamic panel data models include the GMM-type estimators of Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (1998). Li, Maddala, Trost, and Joutz (1997) have used shrinkage estimation techniques for ECM modelling of energy demands.

The second approach uses the simultaneous-equation modelling framework. The approach may start with the unrestricted VAR and impose conditions on the "exogeneity" of the variables and the variance-covariance matrix.

It should be noted that patent filings are count data in nature (e.g. non-negative integers). In that regard, a complementary approach is to use count data analysis.[7] Under this approach, the researcher would work directly with the raw data. Patent counts would not be transformed (for example, through differencing, expressing filings as ratios of other variables like labor or GDP, and/or taking logarithms); otherwise by definition they would no longer be count data. In working with raw data, however, we would encounter issues concerning the stationarity of the variables; the patent counts could be integrated of order 1. Secondly, the number of filings is generally non-zero and generally "large" in value for most countries during most periods. Relatively few zeros or small numbers (e.g. under 20) are encountered. The advantages of using count models are therefore not especially great for international patent filing data. Nonetheless, future work could better incorporate count data modeling in patent forecasting analyses.

4 Conclusions

Some of the approaches described in Sect. 3.1 have been applied to modeling the development of knowledge using USPTO patent data (Abdih and Joutz, 2005). This seems to be a good path to follow in future approaches to EPO data. It would be possible to take even deeper account of econometric theory in formulating time series based regression models. For example, a commonly used approach in macro-economic model building is to posit a model including a whole set of contributory variables, to test and isolate for exogenous versus endogenous effectors, and then to isolate combinations of endogenous effectors in reduced form equations that can be used for the actual fitting (Favero 2001). This approach is of importance when looking at frequently measured sets of correlated variables (e.g. interest rates, GDP, etc.), but may be more suitable for considering policy implications and intervention analysis, rather than for forecasting numbers of patent filings from historical annual counts.

The knowledge processes that are considered are amenable to varying interpretations in terms of modeling approaches. It remains an open question as to whether a detailed examination of knowledge production is appropriate when constructing a forecasting model for budgetary purposes. This is because the data that are likely to be available for modeling may lack the richness required for uncovering the processes that are described.

[7] See the seminal work of Hausman et. al. (1984).

Nevertheless, any simplified model that is used for regression analysis should at least be consistent with an accepted underlying mechanism of knowledge production.

However, the studies in the following chapters do not generally take up on the suggestions made here. Differencing as a form of dealing with cointegration is approached via Box-Jenkins ARIMA methodology in Chap. 4 and Chap. 6, while in Chap. 5 Blind uses an alternative approach that uses a trend effect as an additional variable, and in Chap. 7 an extension of the approach from Sect. 2 of this chapter is used. The various approaches in the following chapters are practical and useful methods for the problem at hand.