

Reflections on patent data analysis¹

By

Elena Kosmopoulou

Preliminary Draft June 2005

(Please do not quote)

1 Introduction

This work reflects upon the research approach and methodological issues faced in the patent data evaluation of diversification. Aims to clarify the syllogism and to introduce a justified perspective upon which analysis can be based. It respects the view that different methods are appropriate 'according to the materials available, the state of investigation reached and the object in view' (Keynes 1897/[1890], p. 6) and that 'specific methods and criteria of analysis are appropriate to the illumination of *some* kinds of objects or materials *but not others*' (Lawson 2003, p. xvi). It aims to present a convincing argument for the method used and provide evidence for the soundness of its selection.

The data used are patent granted in the United States to the world's largest 792 industrial firms as of 1982, derived from the listings of Fortune 500, and compiled to form the Reading database. The empirical evidence concerns the corporate technology, of the very large manufacturing firms, in the period 1969-95.

Section 2, discusses the general research design, *Section 3* examines statistically alternative approaches for the conduct of the study and identifies the main parameters affecting international diversification. *Section 4*, gives a brief account of some of the methodological limitations and identifies some of the research challenges faced by this thesis and lastly, *Section 5* gives a brief description of the research design that follows.

2 Research design

The aim of this study is to improve the understanding of the process of innovation in the context of its (international) environment. Before embarking into further analysis the main factors affecting the innovation process and its internationalisation, need to be identified and the best line of enquiry established. Common sense indicates that the nationality of a firm, the type of industry it is in, the nature of its technology, and the period considered are factors contributing to the rate and direction of internationalisation. Each of these factors is incorporated in either implicit or explicit manner in the classification of the data and research methodology of established literature in the area (see for example Cantwell 1989; Andersen 1997; Fai 1998).

Patents are organised according to the location of origin of the first named inventor. For various different research objectives patent data can be organised in alternative ways according to the industry of the innovating firm to which a patent is assigned, the technological field of the patent (originally organised by the USPTO and then regrouping at a more aggregate level by the Reading group of researchers). Firms and industries tend to have a specific technological profile. The industry of the firm can often be identified from its technological structure². The choice of technological groupings it is to some extent arbitrary. The selection of technological fields (and further groupings into Broad Technological Sectors), industries (and Broad Industrial Groups) is a personal and ad hoc choice, that is directed by first, the issues at hand and second, the desire to be consistent with other work by researchers of the 'Reading Group'.

This work focuses on recent years and the time period of the twenty-seven years from 1969 until 1995. The overall period has further been divided into sub-periods as different levels of aggregation have been considered. In order to avoid or ease down the 'noise' from practices and institutional processes, specific to firms or industries, but not related to their actual level of innovative activity, periods (ranging from 4 to 9 years in length) and not yearly data are considered.

Before embarking upon a more detailed analysis, a preliminary examination of the data through statistical processes is conducted first to identify the most (and least) important parameters in the exploration of patterns of international technological diversification. If

one parameter (for example, 'time') is not very important, then it can be excluded from subsequent analysis. The second objective is to determine the balance between qualitative and quantitative research in what follows. Since the main subject of the thesis is centred on the internationalisation of technological activity, the location of activity is of paramount importance and cannot be disregarded, even if location were to turn out to be less important than other parameters. 'Time' (or 'period') is considered across six periods, 'organisation' is considered in terms of industry technological structure, and lastly the third dimension refers to 'technology'. In sum, the statistical procedures that follow aim to identify first, if the 'period', the 'industry' or the 'technology' are important sources of variation in patenting (and if so, to what extent), and are thus worthy of further more detailed analysis. If, on the other hand, one or more are not important, then the patent data can be 'pooled' and the parameter can be 'dropped' and need not be considered for separate analysis. However, if all (or some) of the parameters ('Period' (P), 'Industry' (I), 'Technology' (T)) are important, which of them most affect the evolution of technological activity?

3 Description of the statistical procedures and interpretation of the results

The analysis was undertaken in steps. First, an ANOVA test was used to determine how much of the variation in patent levels can be attributed to differences between categories and how much is due to variation within each category and then second a regression analysis is run in which patents in each PIT category are regressed on a series of dummy variables that represent specific classes within each category or some combination of classes across categories to allow for interaction effects, and establish the most important sources of variation³. The variables used in both the ANOVA procedure and the regression analysis relate to the PIT categories. First, the overall period 1969-95 is divided into six periods: 1969-72, 1973-77, 1978-82, 1983-86, 1987-90, and 1991-95. Second, the twelve (12) industries considered are: Aircraft, Chemical, Coal & Petroleum, Electrical Engineering, Food, Instruments, Machinery, Metals, Motor Vehicles, Office Equipment, Other Manufacturing, and Pharmaceuticals. Finally, the Technology variable is constructed across the following ten (10) technological fields: chemical,

communication, computing, engines, instruments, mechanical, other, other electrical, pharmaceutical, and transport.

Tables A.1 and A.2 and Figures A.1, A.2, and A.3 are derived from the ANOVA procedure. The ANOVA procedure and the regression analysis tests show essentially the same thing, based on the general linear model (Cohen 1968, p. 426). However, there are some fundamental differences in their theoretical basis and use. First, the use of analysis of variance does not presuppose an explanatory value for independent variables - but likewise here, a regression on a set of categorical dummies does not suppose any causal association either, but is intended merely to establish the patterns of variation found in the data. Second, according to some researchers, the analysis of variance is considered to be a more able and 'versatile' tool to handle 'analytical problems [that may involve] variables that are related to each other, but only over part of the range of studied values, and not necessarily in a simple linear or polynomial fashion. [Especially when the analyst is] interested in uncovering the fact that the two variables *are* related, albeit in a complex and perhaps indescribable [statistically] manner. [...] In [...] situations [in which] the predictor variable is composed of values which differ in *kind*, rather than in *quantity*.[...] It is more general in scope than regression analysis, in that it can be used for identifying relationships between criterion variables and predictor variables, whether those predictor variables are *quantitative* or *qualitative* in nature.' (Kachigan 1991, p. 194). For the ANOVA procedure a square transformation of the observations was used⁴ to reduce the negative skewness of the distribution (fewer firms, technological sectors, or industries have a higher number of patents than lower). The basic model used here is:

$$Y_{ijk} = f(P_i, I_j, T_k, (PI)_{ij}, (PT)_{ik}, (IT)_{jk}, \Sigma_{ijk}E) + \mu$$

Where Y is a function of the factors P, I, T, (PI), (PT), (IT) and ΣE plus the constant μ , the general mean of the model. The main effects are P the period (or 'time', with $i=1, \dots, 6$); I the industry ($j=1, \dots, 12$); and T the technology ($k=1, \dots, 10$). The interactions in the model are (PI), the interaction between period and industry, (PT), the interaction between period and technology, and (IT), the interaction between industry and technology, while ΣE is the cumulative error.

Table A.1 here

The null hypothesis H_0 is that all the parameters have the same effect. The H_1 hypothesis states the opposite, that not all parameters have the same effect. If true then the null hypothesis is rejected. As Table A.1 indicates the F value is large (>0) for $p < 0.0000$, so the null hypothesis can be rejected. Thus, the observed fluctuations in patenting are not due merely to random chance. In fact, the results of the ANOVA procedure show high levels of F-values at the 99% level of significance for all the interactions and main effects. Considering the main effects, the most significant source of variation is attributed to technology, followed by the industry effect, while the time period is much less significant. Of the three interactions due to factors in combination with one another, it is the technological structure of industries that is persistently responsible for the distinctiveness of the patterns observed. Although changes over time are also significant their importance is less than that due to technology. Moreover, the effect of time is more significant at the level of industries than at the level of technological fields⁵.

Figures A.1, A.2, A.3 and Table A.2 here

Figures A.1, A.2, A.3 are concerned with the normality of the distribution. In Figure A.1, residual values are plotted. In Figure A.2 the observed probability of the model is plotted against the expected probability for the Residuals⁶, while Figure A.3 shows the detrended plot of the Residuals. An additional test has been conducted to test the hypothesis that all the variables have the same mean. The results of the test are reported in Table A.2.

Table A.3 here

Having established the significance of the various interactions and main effects at an overall level, it is now possible to proceed to a more detailed analysis of the relative importance of patenting associated with specific industries, technological fields and periods. To this end, a regression analysis has been conducted that distinguishes specific

industry, technology, and period effects⁷. The model is similar to that used in the ANOVA process. The constant here is the intercept b and κ_i , λ_j , μ_k , ν_{ij} , ξ_{ik} , ζ_{jk} are the coefficients - that are of interest now - of the collection of dummy variables in the model, and ε is the error:

$$Y_{ijk} = b + \sum \kappa_i P_i + \sum \lambda_j I_j + \sum \mu_k T_k + \sum \nu_{ij} (PI)_{ij} + \sum \xi_{ik} (PT)_{ik} + \sum \zeta_{jk} (IT)_{jk} + \varepsilon_{ijk}$$

The dummy variables of the regression analysis include all feasible linear combinations of industry, technological field and period, in total 177 dummies having allowed for a control category in each case (so for example 11 industry dummies with 12 industries). For the direct effects the industry 'Other Manufacturing', the technological field 'other' and the period 1973-77 were used as the control categories; and for the purpose of the interaction effects the industry 'Coal and Petroleum', the 'other electricals' technological field, and the period 1987-90 were taken as the control categories. The results are shown in Tables A.3, A.4.1, A.4.2, A.5.1 and A.5.2 and in the residuals as plotted in Graphs A.1 until A.11. Table A.3 reports the results of the Analysis of Variance for the regression of patents on the industry, technology, and period dummies to assess the overall fit of the model. The value F is significant at the 99% level.

Tables A.4.1 and A.4.2 here

Tables A.4.1 and A.4.2 report the categories whose dummy variable representations have estimated coefficients with the most significant t values. The results are consistent with those of the ANOVA test discussed above. In particular, the model fit depends most on some specific interactions, especially those concerning a few given technological fields, to a lesser degree some particular industries, while the period effect is much less pronounced. The main industry effects present in the table relate to the patenting characteristics of the *Chemical* and *Mechanical* industries, and the only period effect shows a tendency towards higher patenting in the most recent period, 1991-95. However, the pattern of innovative activity of the largest multinationals in the period 1969-95 is best described by a large number of specific interactions, each with high level of

significance. Many of the significant effects are interactions of positive value, and all the significant effects with negative values are due to interactions.

Most positive interactions (Table A.4.1) are clustered around mainly, *Equipment and Instruments* industries. The *Electrical and Electronics* industry, but also the 'related' *Office* The *Electrical and Electronics* industry is present in its relationship with *communication, computing*, other *electrical* and *instrument* technologies and the period 1991-95. The *Office Equipment* industry features prominently through significant interactions with *computing*, other *electricals*, *pharmaceutical* technologies, and the period 1991-95 again, while the *Instrument* industry has high patenting in interaction with the other *electrical* technologies category. Lastly, *Chemical* and *Pharmaceutical* industries feature relatively less, but the interactions with *pharmaceutical* technologies are significant for both industries, and with *chemical* technologies for the *Chemical* industry (showing that the inter-industry variation in patenting in any given industry's primary technology is most pronounced in the case of the commitment shown by the chemical industry to its own primary technological field). The strong specificity of industry-primary technology effects can also be seen in the pharmaceutical industry/pharmaceutical technology, and the office equipment/computing technology combinations. In other words, it is in the science-based sectors in which industries are most responsible for their 'own' technologies. All variables that have negative coefficients (that relate to the regression model in an inverse fashion) are interactions related mostly to chemical but also mechanical technologies. Both these technologies have significant interactions with the *Food*, *Office Equipment*, and *Instrument* industries. *Mechanical* technologies have also a significant negative association with the *Pharmaceutical* industry, while chemical technologies do so with the *Motor Vehicle*, *Electrical Engineering*, *Machinery* and *Aircraft* industries. In fact, *chemical* technologies are inversely related to the model with a negative interaction in combination with all industries except the *Chemical* and *Pharmaceutical* ones with which they have significant positive interaction. This reinforces the observation above about the degree of focus on the primary technologies in the *Chemical* and *Pharmaceutical* industries, compared to the development of these technologies in other industries. In sum, attention needs to be directed especially towards the *Electronics Office Equipment* and *Instrument*

industries' interactions with technologies in the latest period examined (1991-95), and on chemical and mechanical technologies in relation to many industries as well as within the *Chemical* and *Mechanical* industries themselves.

Graphs A.1, and Tables A.5.1 and A.5.2 here

In Graphs, A.1, A.2, A.3, A.4, and A.5 all of the residuals resulting from the regression analysis procedure have been plotted and organised in a number of different ways. In the first graph, Graph A.1, the residuals have been organised by ascending value. The value of the residuals is identical to the simple numerical ordering of observations given along the x-axis, except for the very low values, where the residuals are negative, and the very high residual values. Only a few outliers can be distinguished, a distinctly positive subset and a negative group, but these do not influence the general trend (overall the residuals are approximately normally distributed, like in Figure A.1). The two furthest outlying observations have been identified as the combination of the *Electrical and Electronics* industry, and *computing* technology in the period 1991-95 (which has the highest positive value), and the combination of the *Chemical* industry, and *chemical* technology in the period 1983-86 (which has the highest negative value).

Graph A.2 here

In Graph A.2, the residuals have been plotted against the absolute number of patents in ascending order (which was the dependent variable in the regression analysis). The residuals tend to spread out moving rightwards along the x-axis with both higher and lower residuals as patenting rises, but many more outliers can be observed. Although a few outliers can be observed at a lower level of patenting, most of them are grouped at the higher end, since the residual variance rises with the level of patenting activity. This is to be expected. Furthermore, it is more likely that at the higher levels of patenting, positive residuals will occur. However, negative residuals may occur throughout the range of patenting activity. Finally, the outliers at the high end of the spectrum are the same as those noted above (in Graph A.1 and Tables A.5.1 and A.5.2).

Graph A.3, A.4, and A.5 here

Graphs A.3, A.4, and A.5 observations are grouped by period, industry, and technology categories respectively, and within each category by increasing residual values. Graph A.3 draws a comparative picture of residuals across time (the periods from right to left are: 1969-72, 1973-77, 1978-82, 1983-86, 1987-90, and 1991-95). In Graph A.4, the industries are ordered from left to right in alphabetical order: *Aircraft, Chemical, Coal & Petroleum, Electrical Engineering, Food, Instruments, Machinery, Metals, Motor Vehicles, Office Equipment, Other Manufacturing, and Pharmaceuticals*. Finally, in Graph A.5 the technological fields are ordered again from left to right as follows: *chemical, communication, computing, engines, instruments, mechanical, other, other electrical, pharmaceutical, and transport*.

The Graphs A.3, A.4, and A.5 show a noticeably higher variability in patenting in the periods 1973-77 and even more in the last period 1991-95, having controlled for all the systematic industry and technology effect. In Graph A.4, the industries that exhibit most cross-technological field (and cross-period) variability in patenting are the *Chemical and Electrical and Electronics* industries (the outliers being chemical and computing technologies as shown also in Graph A.1, and A.2). Lastly, in Graph A.5 the chemical and computing technological fields (the first and third segments in Graph A.5 respectively) can be distinguished through their distinctly different cross-industry (and cross-period) patterns.

Tables A.5.1, A.5.2 and Graphs A.6 to A.11 are concerned with those observations that have the highest residual values from the regression analysis. These are separated by the highest positive values (Table A.5.1 and Graphs A.6, 5.7, and A.8) and the most negative values (Table A.5.2 and Graphs A.9, A.10, and A.11) and will be discussed in this order. In particular, Graphs A.6-A.11, show the relative importance of the highest 30 (positive and negative) residual values of the regression. The greater the number of observations and the higher the aggregate absolute size of residuals attributable to observations that fall into that category, the greater the surface area it will occupy in the graph (for example in the Graph A.6 residuals of observations in the period 1991-95). The

representation takes into account both the value of the residual and the number of residuals included in the top/bottom 30 residuals as shown in Tables A.5.1 and A.5.2.⁸ Table A.5.1 presents the residuals from the regression analysis with the highest positive value to show the main individual peculiarities in patenting that exist after having allowed for the systematic sources of variation. The *Coal and Petroleum* industry stands out as having positive residuals in the *chemical* technological field in all six periods (since it was the 'control' industry of interaction effects, this combination had not been separately allowed for in the regression). The *Electrical and Electronics* industry has a strong presence, mostly in earlier periods, but also in the most recent times in *computing, communication* and *other technologies*. The *chemical* and *mechanical* technological fields feature prominently, especially in the early periods. Two other technological fields that deserve a mention are the *pharmaceutical* and *motor vehicle* fields (see also Kosmopoulou 2004). The pharmaceutical technological field has a high residual in chemical technologies in the earlier years (1973-77) and in *pharmaceutical* technologies in the latest period (1991-95). The *Motor Vehicles* industry has an earlier high value in *engine* technologies (1983-86) and a later one in *mechanical* technologies (1991-95).

Graphs A.6, A.7, and A.8 here

Graphs A.6, A.7, and A.8 present an aggregation by category of the residual values from Table A.5.1 that includes those residuals with the highest positive values. Graph A.6 presents the distribution of the residual values, organised by period. In the period 1969-95, the highest positive degree of individual variability across industries and technological fields is observed in the periods 1991-95, and to a lesser degree 1973-77, while much less variability occurred in the other four periods. Graph A.7 presents the distribution of the residual values, organised by industry. Four industries have observation with the highest residual values, and the one that most stands out is the *Electrical and Electronics* industry. Much cross-field (and cross-period) variability also occurs within the *Coal and Petroleum* and the *Chemical* industries, and to a lesser extent in the *Pharmaceutical* industry. Graph A.8 presents the distribution of the residual values, organised by technological field. Here the *chemical* technological field stands out

for the extent of its unexplained cross-industry (and cross-period) variance, followed at a distance by *computing* and *mechanical* technological fields.

Table A.5.2 presents the residuals from the regression analysis with the highest negative values to show which individual combinations have patenting lower than one might have expected from the systematic effects built into the regression model. Occupying prominent place in this table are residuals from observations within the *Electrical and Electronics* and *Other Manufacturing* industries. Much cross-industry variability is observed in the early periods in the computing and communication technological fields, while in the latest period feature residuals from the *chemical, pharmaceutical, engine, transport* and *other technologies* (also in the period 1983-86). The industries grouped under the label *Other Manufacturing* (an amalgamation of industries not included in any of the other categories) have high negative residual values in chemical technologies throughout the period 1969-95 (in all six periods), an interaction effect that was not valid for as the *Other Manufacturing* industry was used as a control category. The *Chemical* industry, also features often in the table, with high negative residual values in a variety of technological sectors and periods (*pharmaceutical* technologies in the period 1969-72, *chemical* technologies in the periods 1983-86 and 1987-90, *mechanical* technologies in the period 1983-86, and *computing* technologies in the period 1991-95).

Graph A.9, A.10, and A.11 here

Graphs A.9, A.10, and A.11 present the distribution of the residual values from Table A.5.2 that shows the observations that have residuals with the most negative values. Graph A.9 is organised by the period of the observations that have these residual values. As in Graph A.6 the period 1991-95 saw the highest variability but this time its prominence is less distinct and is followed by the periods 1983-86, and 1969-72. In Graph A.10 the residual values are organised by industry and again as in Graph A.7 the industry that stands out is that of *Electrical and Electronics*. The *Other Manufacturing, Chemical* and *Pharmaceutical* industries are the industries that had higher negative cross-field variability in the period 1969-95. Finally, Graph A.11, which is organised by

technology and shows that the chemical technological field is the one that suffered most from unexpected drops of activity, while computing technologies follow.

4 Methodological limitations

The limitations of the methodology used in this thesis will be summarised under two main headings. First, the limitations posed by the use of an inductive method; and second, the general problem of the 'categorisation' or 'systematisation' of the data. Both are common problems in empirical research.

Patent data are well suited for quantitative analysis and the method of induction⁹. Schmookler in his book *Invention and Economic Growth* (1966) was one of the first to employ this process in search of the determinants of the inventive activity and although his interpretation of some of his results have been later contested (see Rosenberg 1982, p. 18), his approach has been followed broadly since then. However, induction involves detailed empirical study out of which emerges the idea for a hypothesis and abstraction, still the identification of common characteristics - and the most important ones - poses many problems. Writing earlier Keynes¹⁰ had warned of the dangers of 'the slippery problem of passing from statistical description to inductive correlation. [...] The statistician, who is mainly interested in the technical method of his science, is less concerned with discovering the precise conditions in which a description can be legitimately extended by induction. He slips somewhat easily from one to the other, and having found a complete and satisfactory mode of description he may take less pains over the transitional argument, which is to permit him to use this description for the purpose of generalisation' (Keynes 1973, p. 315-16).

If the problem of induction is one that is often encountered in the statistical handling of data, it can also be seen as a part of the more general problem of extrapolation, the basis of abstraction and theoretical thinking. There is no conclusive argument about how far theoretical abstraction can go. At one end of the spectrum are scientists like Wilson (2000) who believes in what he terms his 'Ionian Enchantment', an expression 'coined [that] means a belief in the unity of sciences - a conviction, far deeper than a mere working proposition, that the world is orderly and can be explained by a small number of

natural laws' (p. 4). He traces back this idea to Thales of Miletus, in Ionia, in the sixth century B.C. who according to Aristotle was the founder of the physical sciences and in more recent days to Einstein who proclaimed: 'It is a wonderful feeling to recognize the unity of a complex phenomena that to direct observation appear to be quite separate' (Einstein quoted in Wilson 2000, p. 5). According to Wilson '[a]ll scientists¹¹, Einstein not excepted, are children of Tantalus.' (Wilson 2000, p. 5) This is certainly true, if not for any other reason, because the promise of the unified theory that seemed so close to Einstein still remains illusive. However, it is debatable whether a unifying theory exists, is an ideal to be sought, or is not even an appropriate objective for a social scientist. Lawson (2003) expresses an alternative view. 'In fact, once we accept the open, dynamic and holistic nature of features of reality [...], it comes as little surprise to find that different social scientists [...] regularly produce competing explanations of given concrete phenomenon' (p. 177).

The second issue related to that above is the 'systematisation' of empirical data for the purposes of theoretical analysis. Since,

'Abstraction entails the identification of what is essential to, and enduring in an entity, ignoring the accidental and superficial. More fundamentally, the identification of features, relations and structures depends upon acts of taxonomy and classification, involving the assignment of sameness and difference. Classification, by bringing together entities in discrete groups, must refer to enduring common qualities.' (Hodgson 1999, p. 143).

Systems and their boundaries are not easily defined. Boundaries are blurred and their delineations vague. The problem of limits has been discussed extensively¹². Any categorisation is problematic and it is not only empirical work in the social sciences, and indeed the field of the analysis of patent statistics that faces this challenge. Charles Darwin finishing his book *On the Origin of Species* in 1859 wrote in his final chapter about the difficulties researchers face in categorising plants, animals and primitive humans:

'Systematists will be able to pursue their labours as in present; but they will not be incessantly haunted by the shadowy doubt whether this or that form be a true species. This, I feel sure and I speak after experience, will be no slight relief. [...] Systematists will have only to decide (not that this will be easy) whether any form be sufficiently constant and distinct from other forms, to be capable of definition; and if

intermediate gradations, are looked at by most naturalists as sufficient to raise both forms to the rank of species.' (Darwin 1969/[1859], p. 13-14)

No definite answers exist and Hodgson refers to Alfred Whitehead (1926) who contended that though 'abstractions are essential [...] they always do some violence to the complex, changing reality' (Hodgson 1999, p. 143)¹³. This seems in line with Darwin's ideas who argued about categorisations in his field of 'species' by referring to 'naturalists' and their approach to the problem of 'genera' as 'merely artificial combinations made for convenience'. He ascertained that though this is not ideal it will 'free' researchers 'from the vain search for the undiscovered and undiscoverable essence of species' (Darwin 1964/[1859], p. 13-14).

The challenge then is not only to 'systematise' empirical data but also to try and ensure that the chosen classification is both small enough to be consistent internally, but also large enough to be the appropriate for the level of analysis necessary. Lawson draws a metaphor from Marx to delineate the problem. He reminds us that in 'the analysis of economic forms ...neither microscopes nor chemical reagents are of "use"' (Capital, vol. I, 19 cited in Lawson, 2003, p. xvi). In addition, statistical procedures may impose requirements as to the size or other dimensions of boundaries, as for example in the varying size of different categories of technological fields when measuring innovation across fields, 'the impact of a radical innovation and its diffusion is substantially ameliorated by the use of large number of patents' (Cantwell and Barrera 1999).

5 Concluding Remarks

The results set out here lay the foundations for the direction of the analysis. Both the ANOVA procedure and the Regression analysis show a high and significant systematic effect on patenting activity of the technological field, industry and period, as well as their interactions. In fact, the overall system is divided into a number 'of interconnected exchange relationships. Any attempt to explain the evolution of this system must be founded upon an understanding of the interaction over time between the evolution of the parts and the evolution of the whole system' (Lundgren 1991, p. 43-44).

Thus, the limitations of a solely quantitative analysis being apparent, this work will proceed by using an amalgamation of both quantitative, mainly descriptive statistics¹⁴, and qualitative analysis. After all, in a work following an evolutionary theoretical framework, one should take seriously the warning given by C.H. Waddington: 'The whole real guts of evolution - which is, how do you come to have horses and tigers, and things - is outside the mathematical theory' (Waddington, 1967 cited in Gould 2002, p. 584).

Further, the regression analysis has revealed many and significant effects in a large number of interactions of specific dual relationships among the main categories of variables, only a fraction of which have been separately identified (with a > 0.001). These results paint a very interesting although complicated picture of a large number of relationships concerning the evolution of industries' technological structure over time.

- **Endnotes:**

¹ 'Everything depends upon the security of our starting-point' Descartes

² As Professor Keith Pavitt used to say: 'show me its technological profile and I will tell you which industry the firm is in'.

³ Another difference is that '[t]he ANOVA framework has developed hand in hand with the experimental method starting with Fisher's work in Rothamstead Experimental Station of Agriculture (in England) [...] The regression and correlation techniques have developed mostly from non-experimental research. They are more general statistical techniques and now can be run relatively easily (and quickly) with modern computers.' (Wright 1997 p. 132)

⁴ A square transformation of observations is commonly suggested to improve the fit of the model when a distribution is negatively skewed (see for example Warrack 2000). The ANOVA test was also run using observations measured in absolute number of patents with similar results (not reported here). However, the use of a square transformation improved the symmetry of the distribution and the value of R-square adjusted from 0.942 (without transformation, when the ANOVA was run with absolute numbers of patents) to 0.075 (with the transformation when the values of the observations were substituted with their square equivalent).

⁵ The results reported are consistent with other empirical studies: 'An empirical analysis of some dimensions of SIS, based on patent data, has been provided for six countries. The result confirm, by and large, that both country specific and technology specific factors affect in essential ways the sectoral as well as the spatial organization of innovative activities within given industries. (Breschi and Malerba 1997, p. 153)

⁶ The standard SPSS procedure plots the Expected Normal quantiles calculated using Blom's proportional estimation formula and assigning the mean to ties.

⁷ There are difficulties in using the ANOVA when the aim is to analyse a large number of factors. Although the ANOVA computer procedures can handle with ease the increased complexity of four or five factors or

groupings of variables, interpretation is not straightforward. This 'apparent ease can be problematic. It is important to realise that by increasing the number of factors/variables, the models become more complex, and therefore more difficult to interpret. [...] The p values quoted by most packages assume that each individual effect is the only one being tested. In most cases this is not true.' (Wright 1997 p. 195).

⁸ The colours in the graphs of industrial and technological groups have been chosen to roughly correspond to the Broad Technological Sectors and Industrial groupings. Hence, shades of red, orange and yellow correspond to pharmaceutical, chemical, coal and petroleum industries and core technologies; shades of green to metal, mechanical and motor vehicle industries and technological fields; shades of blue to electrical and electronics; and grey to the category 'others'.

⁹ According to the definition of 'induction' as the research approach involving the development of a theory as a result of the observation of empirical data.

¹⁰ The problem of induction already highlighted by Hume attracted wide reference from Keynes as Lawson and Perasan (1985, p. 91) highlight: 'Thirty years ago I used to be occupied in examining the slippery problem of passing from statistical description to inductive correlation in the case of simple correlation; and today in the era of multiple correlation I do not find that in this respect practice is much improved' (Keynes 1973, p. 315-16).

¹¹ One could have a prolonged argument for the definition of such a contested word as this of 'science' and 'scientist'. In this text the word is defined according to Pingree (and not necessarily to Wilson's approval) as: 'a systematic explanation of perceived or imaginary phenomena, or else is based on such an explanation. [This definition further qualifies that m]athematics finds a place in science only as one of the symbolical languages in which scientific explanations may be expressed' (Pingree 2000, p. 35).

¹² See for example Penrose 1957, p. 109 and Hodgson, 1999, p. 143.

¹³ Hodgson (1999) also refers to Nicholas Georgescu-Roegen (1971) for the same argument from the perspective of a theoretical economist. Georgescu-Roegen recognises the same problem in model building and he comments: 'operational concepts have a contradictory of dialectical quality and cannot correspond precisely with real phenomena. [Hew then forewarns that a]ll acts of categorisation and abstraction must, therefore, be provisional. All theoretical foundations must forever be under scrutiny.' (Hodgson, 1999: 143)

¹⁴ 'This kind of mathematical statistics can be conveniently called "descriptive". The word "descriptive" is used here in its narrow sense, implying nothing but the formulation of results in mathematical terms, without attempting to include them in a more general logical system, based on the concept of probability.' (Von Mises 1957, p. 166)

References

- Andersen, H.B. (1997) *Technological Change and the Evolution of Corporate Innovation: The Structure of Patenting, 1890-1990*. Ph.D. thesis, University of Reading)
- Breschi, S. and Malerba, F. (1997) 'Sectoral Innovation Systems: Technological Regimes, Schumpeterian Dynamics, and Spatial Boundaries' in Edquist, C. (ed). *Systems of Innovation: Technologies, Institutions and Organisations*. London: Pinter, 6: 130-156.
- Cantwell, J.A. (1989) *Technological Innovation and Multinational Corporations*, Oxford, Basil Blackwell.
- Cantwell, J.A. and Barrera, P. (1999) *The History of Technological Development in Europe and the United States*. Oxford: Oxford University Press.
- Darwin, C. (1969/[1859]) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Cambridge, MA (USA): Cambridge University Press. (introduction by Ernst Mayr) (first edition in London: Murray)
- Darwin, C. (1985/[1859]) 'Recapitulation and Conclusion', in Martin Gardiner (ed.) *Great Essays in Science*, Oxford: OUP, 2: 7-17.
- Fai, F.M. (2003/[1998]), *Corporate Technological Competence and the Evolution of Technological Diversification*. Cheltenham: Edward Elgar. (1998 PhD thesis, University of Reading)
- Georgescu-Roegen, N. (1971) *The Entropy Law and the Economic Process*. Cambridge, MA (USA): Harvard University Press.
- Gould, S.J. (2002) *The Structure of Evolutionary Theory*, Cambridge, Mass. (USA): The Belknap Press of Harvard University Press.
- Hodgson, G. M. (1999) *Economics and Utopia*. London: Routledge.
- Kachigan, S.K. (1991) *Multivariate Statistical Analysis: A Conceptual Introduction*. New York: Radius Press. (2nd ed.)
- Keynes, J.N. (1897/[1891]) *The scope and Method of Political Economy*. (2nd edition) London: McMillan.London: McMillan & Co.
- Keynes, J.M. (1973) 'The General Theory and After: Part II, Defense and Development' in *The Collected Writings of John Maynard Keynes*, Royal Economic Society by Macmillan and Cambridge University.
- Lawson, Tony (2003) *Reorienting economics*. London: Routledge.

- Lundgren, A. (1991) *Technological Innovation and Industrial Evolution: The Emergence of Industrial Networks*. Stockholm: Stockholm School of Economics, The Economic Research Institute.
- Marx K. (1974/[1885]) 'A Critical Analysis of Capitalist Productions' in Engels, F. (ed) *Capital*. London: Lawrence and Wishart, **1**. (first published in 1867)
- Rosenberg, N. (1982) *Inside the Black Box: Technology and Economics*. Cambridge: Cambridge University Press.
- Schmookler, J.A. (1966) *Invention and Economic Growth*. Cambridge, Mass., (USA): Harvard University Press.
- Wilson, E.O. (2000) *Consilience: The Unity of Knowledge*. New York: Alfred A. Knopf.
- Wright, D.B. (1997) *Understanding Statistics: An Introduction for the Social Sciences*. London: Sage Publications.

Table A.1: Results of Anova Test (Between-Subjects Effects), using a square transformation

| Source | Sum of Squares | df | Mean Square | F | Sig. |
|-----------------------|-----------------------|-----------|--------------------|-----------|-------------|
| Corrected Model | 415626.489 | 224 | 1855.475 | 126.746 | 0.000 |
| Intercept | 525304.043 | 1 | 525304.043 | 35883.065 | 0.000 |
| INDUSTRY | 71997.899 | 11 | 6545.264 | 447.101 | 0.000 |
| technology | 164568.527 | 9 | 18285.392 | 1249.059 | 0.000 |
| PERIOD | 3258.229 | 5 | 651.646 | 44.513 | 0.000 |
| INDUSTRY * technology | 166743.329 | 99 | 1684.276 | 115.051 | 0.000 |
| INDUSTRY * PERIOD | 5476.674 | 55 | 99.576 | 6.802 | 0.000 |
| TECHNOLOGY * PERIOD | 3581.830 | 45 | 79.596 | 5.437 | 0.000 |
| Error | 7246.469 | 495 | 14.639 | | |
| Total | 948177 | 720 | | | |
| Corrected Total | 422872.957 | 719 | | | |

R Squared = .983 (Adjusted R Squared = .975)

| Class | Levels | Values |
|-------------------|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| period | 6 | 1969-72, 1973-77, 1978-82, 1983-86, 1987-90, 1991-95 |
| industry | 12 | Aircraft, Chemical, Coal & Petroleum, Electrical Engineering, Food, Instruments, Machinery, Metals, Motor Vehicles, Office Equipment, Other Manufacturing, Pharmaceutical |
| technology | 10 | chemical, communication, computing, engines, instruments, mechanical, other, other electrical, pharmaceutical, transport |

Figure A.1: Residual Plot (SQ_PATs)

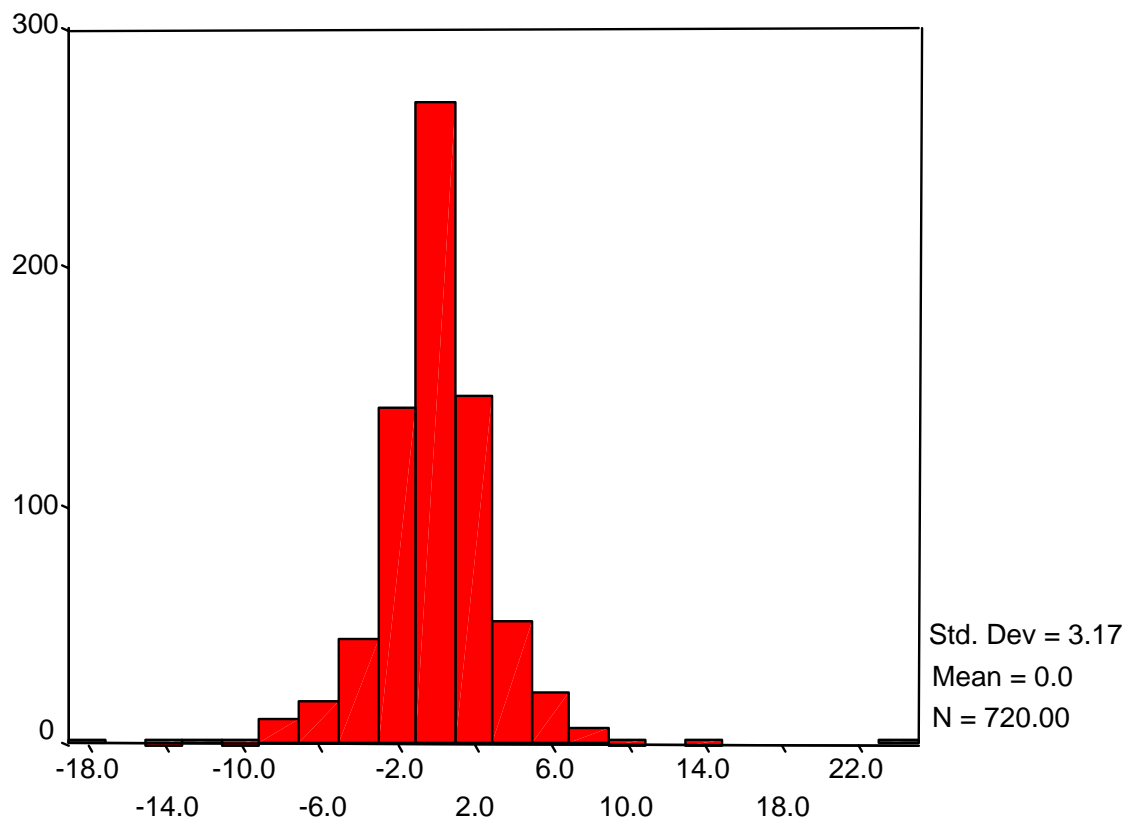


Figure A.2

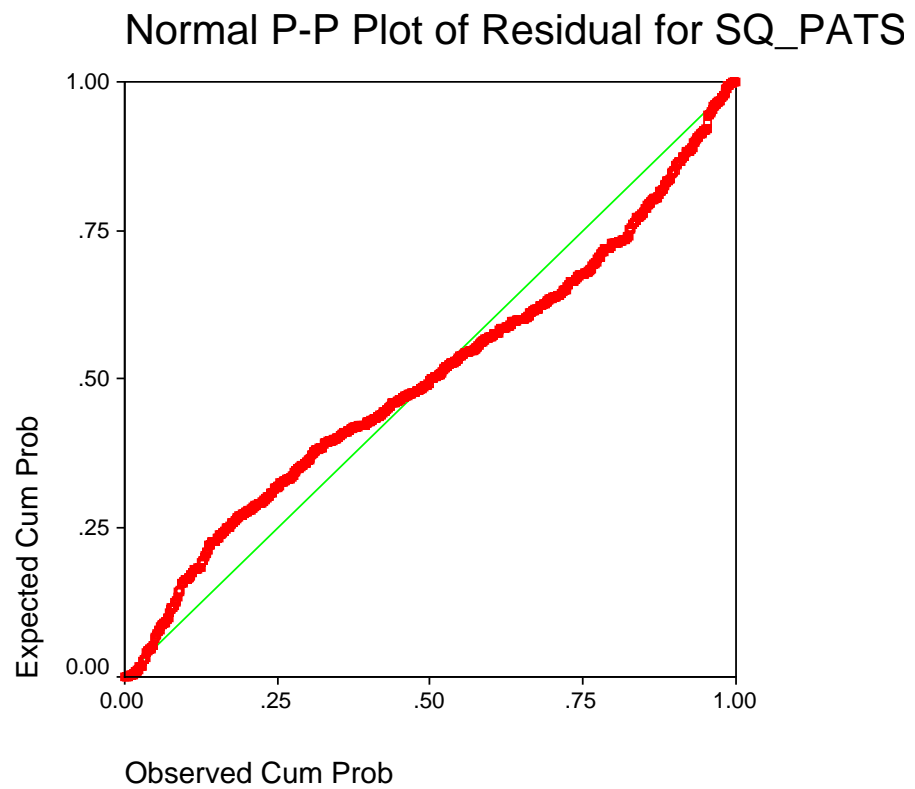
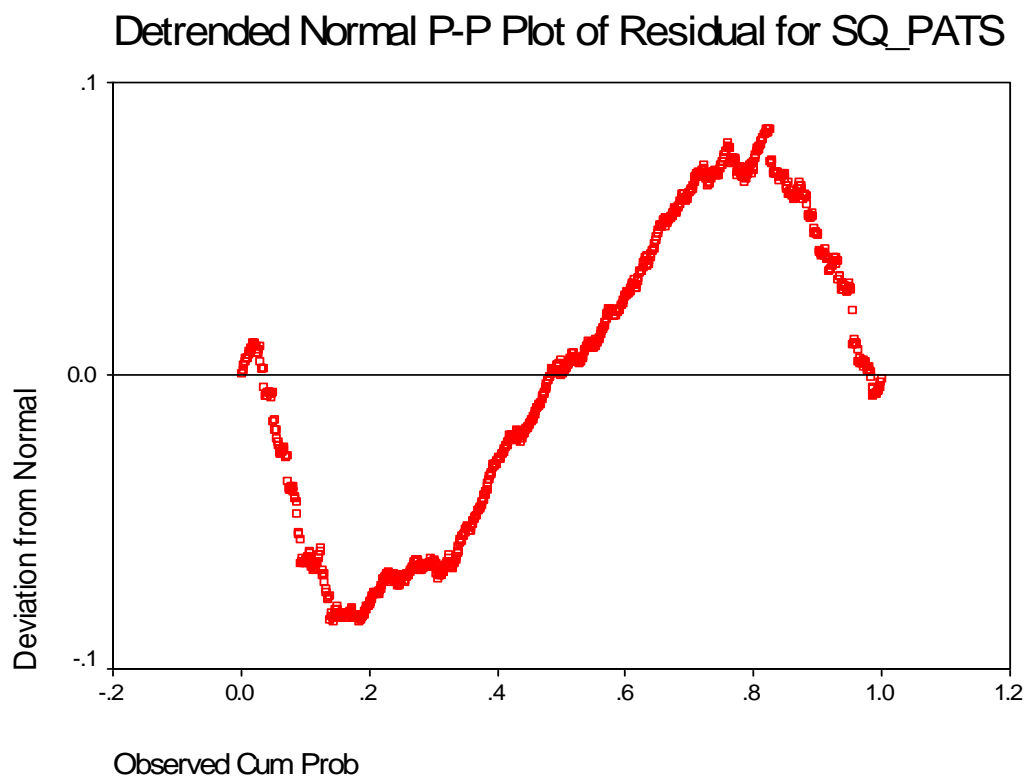


Figure A.3*



*Normal distribution parameters estimated: location=.00000000 scale=3.1746713

Table A.2: Test of Contrast of Means

| TIME (Polynomial) Contrast* | Dependent Variable: | |
|------------------------------------|----------------------------------------------------|-------|
| Linear | Contrast Estimate | 3.785 |
| | Hypothesized Value | 0 |
| | Difference (Estimate - Hypothesized) | 3.785 |
| | Std. Error | 0.349 |
| | Sig. | 0.000 |
| | 95% Confidence Interval for Difference Lower Bound | 3.098 |
| | Upper Bound | 4.471 |

* Metric = 1.000, 2.000, 3.000, 4.000, 5.000, 6.000

Table A.3: Analysis of Variance for the regression analysis of industry, technology, period at the patent level on dummies for these three effects

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|--------------------------|-----------|-----------------------|--------------------|----------------|------------------|
| Model | 177 | 4156812856 | 23484818 | 52.18 | <.0001 |
| Error | 542 | 243960804 | 450112 | | |
| Corrected Total | 719 | 4400773659 | | | |
| Root MSE | 670.90 | | R-Square | 0.945 | |
| Dependent Mean | 1316.91 | | Adj R-Sq | 0.927 | |
| Coefficient of Variation | 50.95 | | | | |

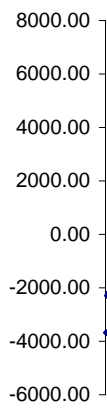
Table A.4.1: Dummy variables with the highest positive t-values, at a (higher than) 99% level of significance.

| | Dummy Variable | Parameter Estimates | Standard Error | t-Value |
|----|-------------------------------------------------------------------|----------------------------|-----------------------|----------------|
| 1 | Chemical Industry * chemical technology | 13334 | 410.843 | 32.45 |
| 2 | Electrical and Electronics Industry * other electrical technology | 10702 | 410.843 | 26.05 |
| 3 | Electrical and Electronics Industry * computing technology | 6676.38 | 410.843 | 16.25 |
| 4 | Chemical Industry | 4071.38 | 279.543 | 14.56 |
| 5 | Electrical and Electronics Industry * communication technology | 5866.96 | 410.843 | 14.28 |
| 6 | Machinery Industry | 3635.47 | 279.543 | 13.01 |
| 7 | Electrical and Electronics Industry * 1991-95 | 3043.60 | 318.238 | 9.56 |
| 8 | Electrical and Electronics Industry * instrument technology | 3513.46 | 410.843 | 8.55 |
| 9 | Pharmaceutical Industry * pharmaceutical technology | 3434.96 | 410.843 | 8.36 |
| 10 | Office Equipment Industry * computing technology | 3280.21 | 410.843 | 7.98 |
| 11 | Electrical and Electronics Industry * mechanical technology | 2697.04 | 410.843 | 6.56 |
| 12 | Office Equipment Industry * other electrical technology | 2242.88 | 410.843 | 5.46 |
| 13 | Chemical Industry * pharmaceutical technology | 2210.63 | 410.843 | 5.38 |
| 14 | computing technology * 1991-95 | 1441.21 | 290.510 | 4.96 |
| 15 | Instrument Industry * other electrical technology | 1969.54 | 410.843 | 4.79 |

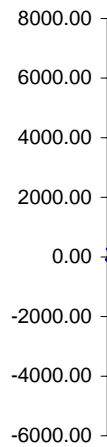
Table A.4.2: Dummy variables with the most negative t-values, at a (higher than) 99% level of significance.

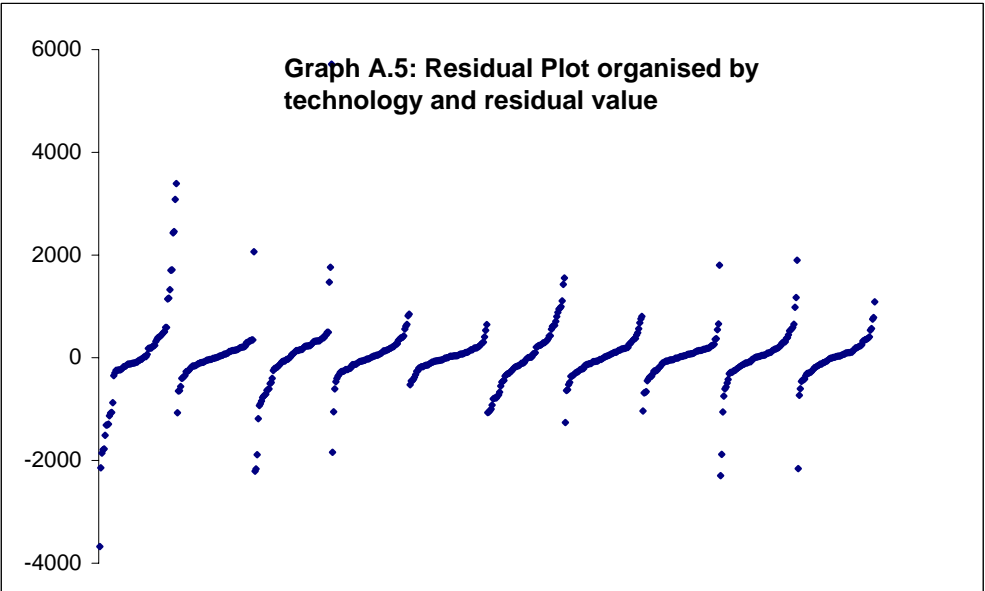
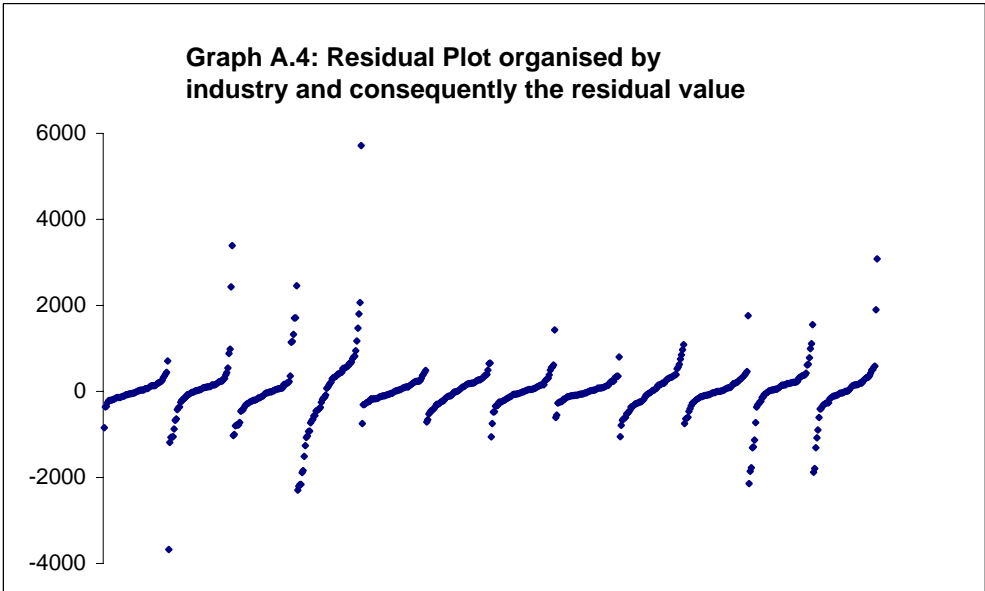
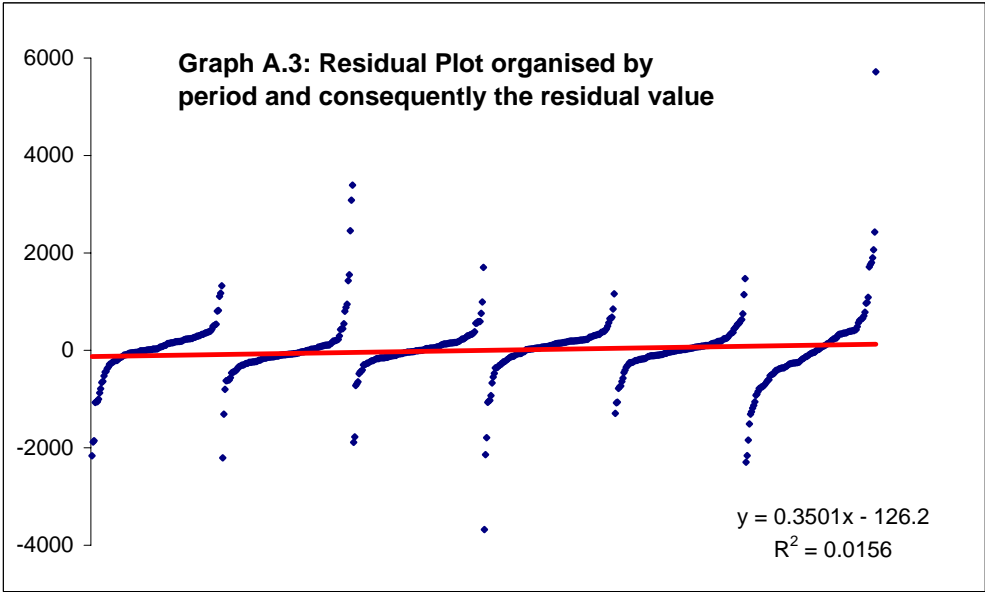
| | Dummy Variable | Parameter Estimates | Standard Error | t-Value |
|----|-----------------------------------------------------------|----------------------------|-----------------------|----------------|
| 1 | Motor Vehicle Industry * chemical technology | -4243.29 | 410.843 | -10.3 |
| 2 | Machinery Industry * chemical technology | -3615.71 | 410.843 | -8.8 |
| 3 | Food Industry * chemical technology | -3309.29 | 410.843 | -8.05 |
| 4 | Office Equipment Industry* * chemical technology | -2960.88 | 410.843 | -7.21 |
| 5 | Aircraft Industry * chemical technology | -2916.71 | 410.843 | -7.1 |
| 6 | Metal Industry * chemical technology | -2711.46 | 410.843 | -6.6 |
| 7 | Instrument Industry * mechanical technology | -2486.96 | 410.843 | -6.05 |
| 8 | Instrument Industry * chemical technology | -2286.38 | 410.843 | -5.57 |
| 9 | Pharmaceutical Industry * mechanical technology | -2279.79 | 410.843 | -5.55 |
| 10 | Office Equipment Industry * mechanical technology | -1874.63 | 410.843 | -4.56 |
| 11 | Food Industry * mechanical technology | -1786.54 | 410.843 | -4.35 |
| 12 | Electrical and Electronics Industry * chemical technology | -1647.04 | 410.843 | -4.01 |

Graph A.1: Residual Plot organised by residual value



Graph A.2: Residual Plot organised by ascending number of patents





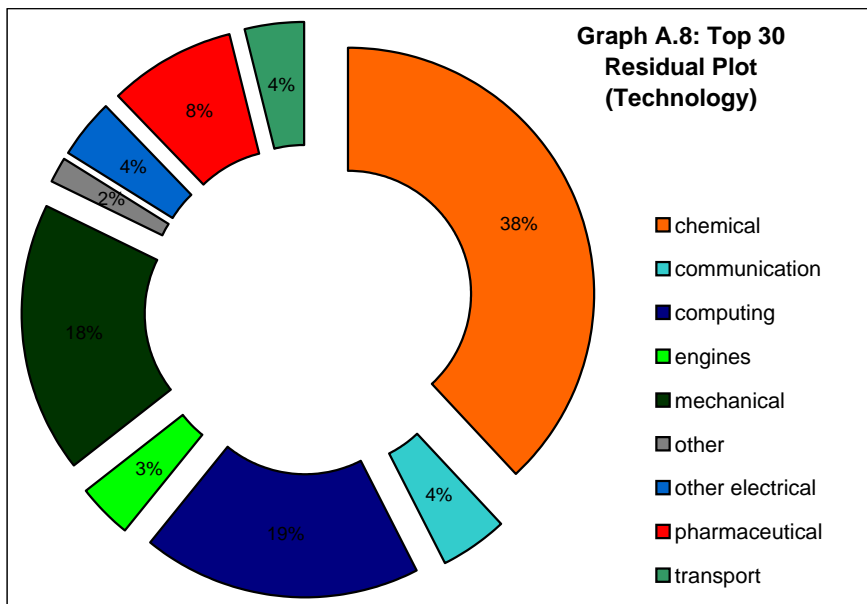
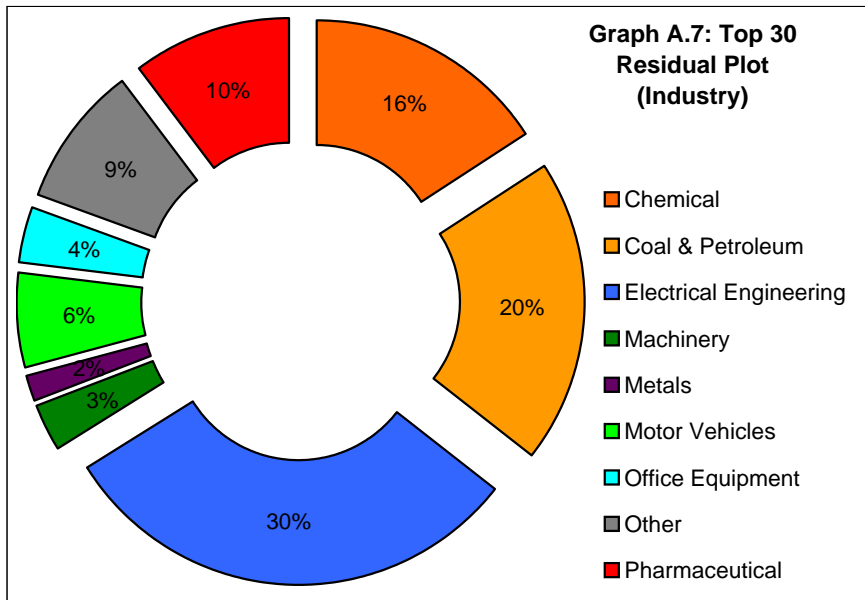
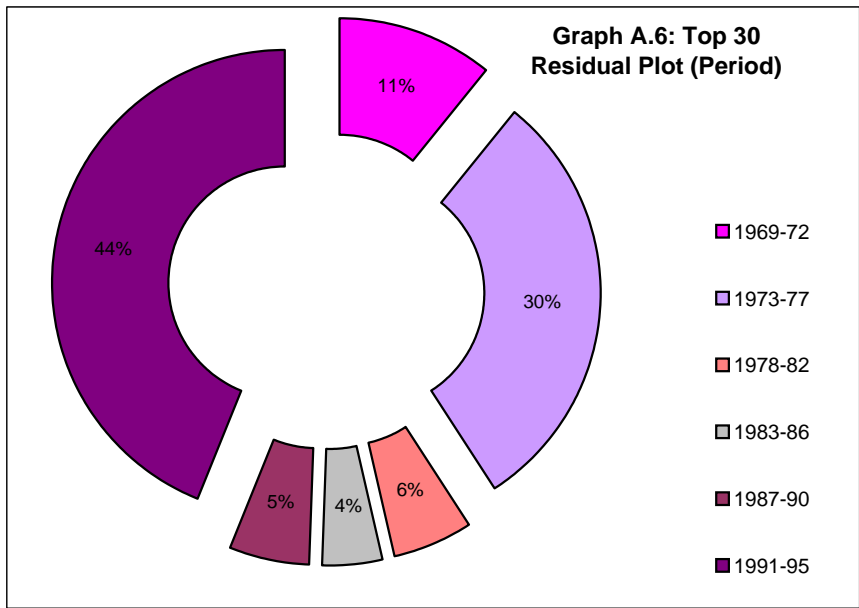
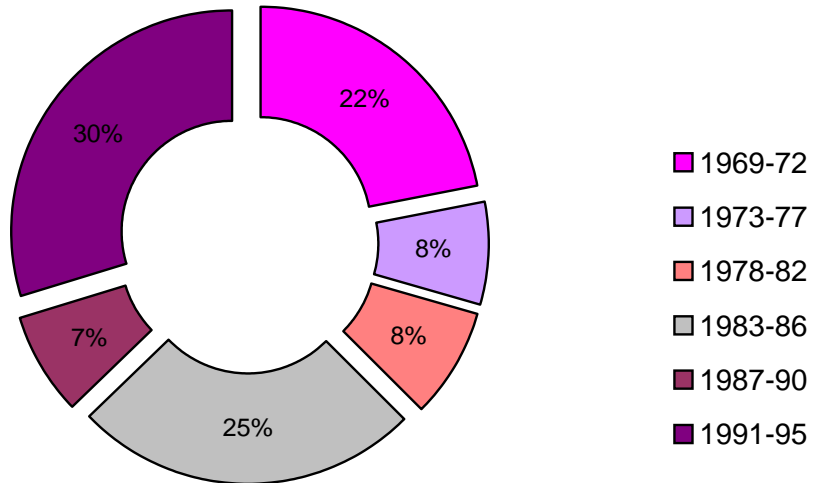


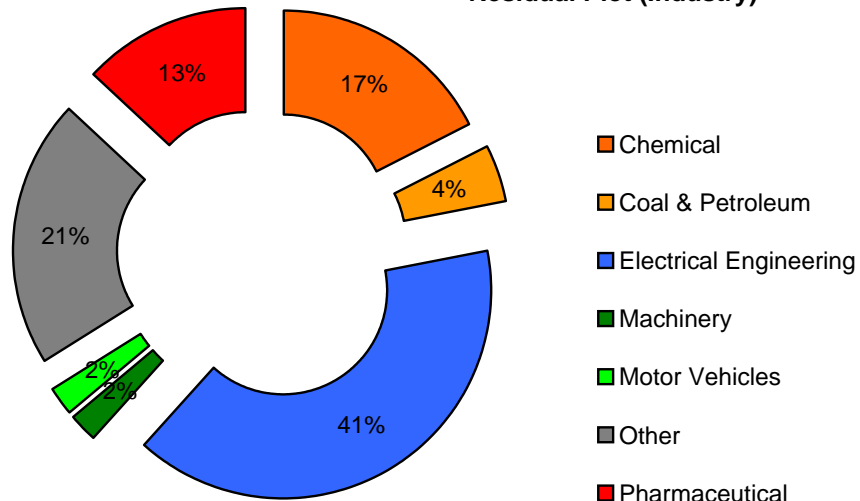
Table A.5.1: Observations whose residuals have the highest positive values.

| | Period | Industry | Technology | Residuals | Nos of Pats |
|----|---------------|------------------------|-------------------|------------------|--------------------|
| 1 | 1991-95 | Electrical Engineering | computing | 5716.03 | 15921 |
| 2 | 1973-77 | Chemical | chemical | 3391.66 | 21275 |
| 3 | 1973-77 | Pharmaceutical | chemical | 3081.91 | 8311 |
| 4 | 1973-77 | Coal & Petroleum | chemical | 2456.26 | 7070 |
| 5 | 1991-95 | Chemical | chemical | 2430.60 | 20659 |
| 6 | 1991-95 | Electrical Engineering | communication | 2065.45 | 11097 |
| 7 | 1991-95 | Pharmaceutical | pharmaceutical | 1899.81 | 5252 |
| 8 | 1991-95 | Electrical Engineering | other electrical | 1802.24 | 15815 |
| 9 | 1991-95 | Office Equipment | computing | 1761.88 | 6509 |
| 10 | 1991-95 | Coal & Petroleum | chemical | 1713.10 | 5895 |
| 11 | 1978-82 | Coal & Petroleum | chemical | 1703.84 | 6198 |
| 12 | 1973-77 | Other Manufacturing | mechanical | 1551.58 | 5668 |
| 13 | 1987-90 | Electrical Engineering | computing | 1472.14 | 9229 |
| 14 | 1973-77 | Machinery | mechanical | 1429.95 | 6915 |
| 15 | 1969-72 | Coal & Petroleum | chemical | 1326.00 | 5698 |
| 16 | 1969-72 | Electrical Engineering | pharmaceutical | 1173.91 | 46 |
| 17 | 1983-86 | Coal & Petroleum | chemical | 1161.68 | 5498 |
| 18 | 1987-90 | Coal & Petroleum | chemical | 1143.36 | 4706 |
| 19 | 1969-72 | Other Manufacturing | mechanical | 1108.62 | 4901 |
| 20 | 1991-95 | Motor Vehicles | transport | 1089.26 | 1977 |
| 21 | 1978-82 | Other Manufacturing | mechanical | 994.53 | 4771 |
| 22 | 1991-95 | Chemical | pharmaceutical | 982.58 | 4019 |
| 23 | 1991-95 | Motor Vehicles | mechanical | 967.90 | 6469 |
| 24 | 1973-77 | Electrical Engineering | mechanical | 947.25 | 7386 |
| 25 | 1973-77 | Chemical | mechanical | 882.82 | 6572 |
| 26 | 1983-86 | Motor Vehicles | engines | 850.35 | 2596 |
| 27 | 1969-72 | Electrical Engineering | engines | 819.26 | 49 |
| 28 | 1969-72 | Electrical Engineering | other | 806.35 | 342 |
| 29 | 1973-77 | Metals | mechanical | 801.72 | 5474 |
| 30 | 1991-95 | Other Manufacturing | transport | 784.12 | 827 |

Graph A.9: Bottom 30 Residual Plot (Time)



Graph A.10: Bottom 30 Residual Plot (Industry)



Graph A.11: Bottom 30 Residual Plot (Technology)

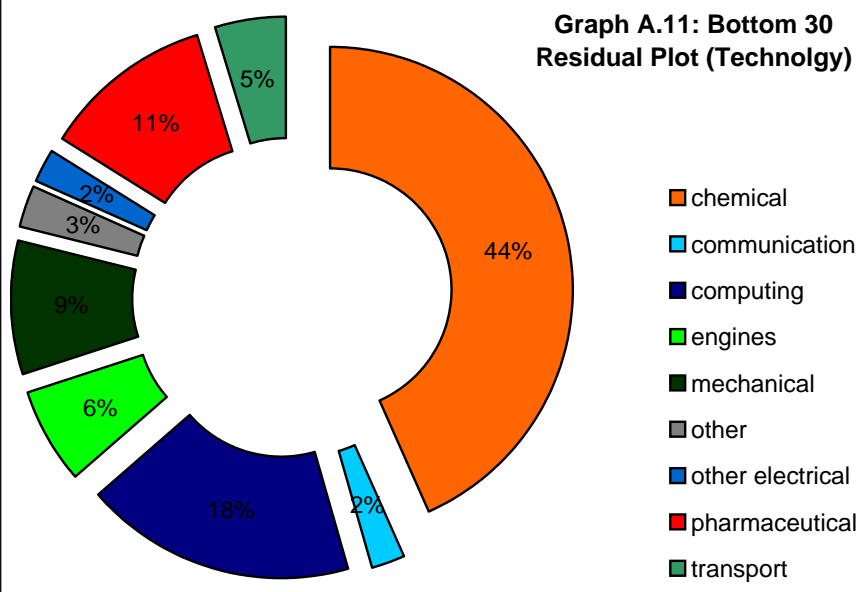


Table A.5.2: Observation whose residuals have the most negative values.

| Period | Industry | Technology | Residuals | Nos of Pats |
|---------------|------------------------|-------------------|------------------|--------------------|
| 1 1983-86 | Chemical | chemical | -3676.92 | 13929 |
| 2 1991-95 | Electrical Engineering | pharmaceutica | -2296.02 | 105 |
| 3 1973-77 | Electrical Engineering | computing | -2206.75 | 4150 |
| 4 1969-72 | Electrical Engineering | computing | -2164.95 | 3524 |
| 5 1991-95 | Electrical Engineering | transport | -2159.77 | 552 |
| 6 1983-86 | Other | chemical | -2141.90 | 2133 |
| 7 1978-82 | Electrical Engineering | computing | -1887.13 | 4076 |
| 8 1969-72 | Pharmaceutical | pharmaceutica | -1880.06 | 1079 |
| 9 1969-72 | Other | chemical | -1855.58 | 2455 |
| 10 1991-95 | Electrical Engineering | engines | -1841.59 | 430 |
| 11 1983-86 | Pharmaceutical | chemical | -1791.67 | 3160 |
| 12 1978-82 | Other | chemical | -1774.75 | 2658 |
| 13 1991-95 | Electrical Engineering | chemical | -1511.33 | 3631 |
| 14 1973-77 | Other | chemical | -1310.32 | 3242 |
| 15 1991-95 | Pharmaceutical | chemical | -1308.00 | 3639 |
| 16 1987-90 | Other | chemical | -1292.22 | 2209 |
| 17 1991-95 | Electrical Engineering | other | -1260.25 | 1253 |
| 18 1991-95 | Chemical | computing | -1185.70 | 459 |
| 19 1991-95 | Other | chemical | -1129.48 | 2991 |
| 20 1987-90 | Pharmaceutical | chemical | -1077.49 | 3102 |
| 21 1969-72 | Electrical Engineering | communicatio | -1069.78 | 4103 |
| 22 1983-86 | Chemical | mechanical | -1066.77 | 4345 |
| 23 1987-90 | Chemical | chemical | -1063.29 | 15915 |
| 24 1991-95 | Machinery | mechanical | -1055.50 | 3503 |
| 25 1969-72 | Chemical | pharmaceutica | -1053.69 | 1029 |
| 26 1969-72 | Motor Vehicles | engines | -1053.16 | 470 |
| 27 1983-86 | Electrical Engineering | other electrical | -1037.08 | 9511 |
| 28 1983-86 | Coal & Petroleum | mechanical | -1026.42 | 2874 |
| 29 1969-72 | Coal & Petroleum | mechanical | -999.80 | 2854 |
| 30 1983-86 | Electrical Engineering | computing | -929.33 | 5150 |