Assigning Patents to Industries: Tests of the Yale Technology Concordance

SAMUEL KORTUM & JONATHAN PUTNAM

ABSTRACT *We describe a method to predict patent counts disaggregated by industry, using available data on patenting by technology field. This method—the Yale Technology Concordance (YTC)—exploits a data set of patents that have been individually assigned by the Canadian Patent Office to both an industry and a technology field. The procedure for predicting patents by industry is developed as a statistical model so that the standard errors of the predictions can be estimated. The YTC is tested on several subsets of Canadian patents by comparing out-of-sample predictions with industry assignments made by the Canadian Patent Office. We find that the predictions of patents by industry are quite accurate for the subset of patents from US inventors. The prediction errors are much greater for the subset of patents granted or published after 1989. This suggests that the relationship between the technology fields and industries has shifted in a way that the procedure does not capture. Nonetheless, predictions from the YTC do appear to give a reasonably accurate picture of the pattern of patenting by industry.*

KEYWORDS: *Patents, Yale Technology Concordance, industry classification, Canada*

1. Introduction

Patents have proved to be useful indicators of research activity and technological change in a number of economic studies (Griliches, 1990). However, patent data have not been available at the industry level—a unit of analysis at which data on research expenditure and productivity growth are collected. As a consequence, certain empirical issues, such as the link between patenting and productivity growth, have received little attention.[1] In this paper, we present a method of

Samuel Kortum, Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, USA. Jonathan Putnam, Charles River Associates, 200 Clarendon Street, T-33, Boston, MA 02116, USA. E-mail: kortum@bu.edu and jputnam@ziplink.net. This paper has been prepared for the 11th International Conference on Input–Output Techniques, New Delhi, India. We thank the Canadian Intellectual Property Office for providing the data on which this paper is based, and for answering our numerous questions. An earlier version of this paper was presented at the NBER Summer Institute. We thank participants for their comments. We have also benefited from the comments of Robert Evenson. Poorti Marino provided excellent research assistance, but we remain responsible for any errors. Kortum gratefully acknowledges the support of the National Science Foundation (NSF) under Grant 9309935–001. Funding for the Yale Technology Concordance was originally provided by the NSF under Grant SRS-8607692.

0953-5314/97/020161-15

predicting patents by industry, using widely available information on the distribution of patenting across technology fields.

Our method of predicting patents by industry relies on a unique data set from the Canadian Intellectual Property Office (CIPO). Similarly to most patent offices, the CIPO assigns a technology field from the International Patent Classification (IPC) system to each patent it issues.[2] However, unlike other patent offices, the CIPO also assigns an industry of manufacture (IOM) and a sector of use (SOU) to most patents.[3] For a product patent, the IOM is the industry that manufactures the product, and the SOU is the industry (or sector) that uses it; however, for a process patent, an IOM is only assigned if the process includes some apparatus (Ellis, 1981). Both concepts appear to be useful, so we provide a means of predicting patents by either IOM or SOU.[4] We examine the industry and technology assignments of over 250 000 patents issued in Canada from 1983 through 1993.[5]

The key assumption that underlies our strategy is that the probability of a patent being assigned to a given industry is a function of the technology field of the patent and nothing else. In particular, we assume that the probability of industry assignment conditional on technology assignment does not depend on the country where the patent originates, or the date when it is issued. We estimate these conditional probabilities from the Canadian data and then apply them to patents in other countries or time periods where we know only the technology field of the patents.

Using the Canadian data to assign patents to industries in other countries was originally pursued by Evenson at Yale; hence, the name Yale Technology Concordance (YTC) (Evenson *et al.*, 1991). The term 'concordance', however, is somewhat misleading, because it implies a deterministic assignment of each technology field to a specific industry. We develop the YTC in a probabilistic framework. The technology field of a patent determines the probability distribution over possible industries to which the patent might be associated. An advantage of this approach is that approximate standard errors of the predictions can be calculated.

A simple example will illustrate our approach. Suppose that there are only three technology fields (*a*, *b* and *c*) and two industries (1 and 2). The following matrix gives the probabilities of a patent being associated with a given industry conditional on the technology field of the patent:

technology\industry	1	2
a	0.5	0.5
b	0.3	0.7
c	1.0	0.0

A patent from technology *a* is equally likely to be associated with industry 1 or 2, while a patent from technology *b* is more than twice as likely to be associated with industry 2. A patent from technology *c* is always associated with industry 1. Assume that this set of conditional probabilities is known (in practice, it will be estimated from the Canadian data). Suppose that, in Japan, we observe 60 patents in technology *a*, 100 patents in technology *b* and 10 patents in technology *c*. The prediction of patents by industry in Japan would be $0.5 \times 60 + 0.3 \times 100 + 1.0 \times 10 = 70$ in industry 1, and $0.5 \times 60 + 0.7 \times 100 = 100$ in industry 2. In the following, we show that the standard error of the prediction

is 6. Taking account of uncertainty in the estimates of the conditional probabilities themselves would lead to a larger standard error.

We are not the first researchers to use patents by class to infer patents by industry. Schmookler (1966) assigned patents by US patent class (USPC) to selected industries of use.[6] His classification rule assigned all the patents in a USPC class to an industry if he determined that at least two-thirds of the patents in the class were used by that industry. The Office of Technology Assessment and Forecast (OTAF), which is part of the US Patent and Trademark Office, also developed a concordance similar to that of Schmookler.[7] In cases where a USPC was related to several industries, the patents in that class were assigned in equal fraction to each of those industries.

In principle, the YTC has several advantages over the OTAF concordance. First, the relation between technology fields and industries is inferred from the assignment to a technology field and an industry of over 250 000 patents. The individual industry assignments are made by trained personnel of the CIPO. There is no need to use judgement in determining the overall connection between technology fields and industries. Second, the YTC uses the IPC (an international standard for classifying patents by technology) rather than the USPC (which is only used in the US). Hence, the YTC can be applied to patent data from a wide set of countries. Third, we explicitly model the error that is inherent in predicting patents by industry. This allows us to quantify the degree of uncertainty attached to our prediction of patents by industry.

Of course, these advantages are in principle, and not necessarily in practice. In this paper, we test the YTC to obtain some idea of how accurate it will be in practice. First, we pretend that, for patents granted in Canada to US inventors, we have only information on the technology field. We then use the YTC to predict patents by industry for this subset of Canadian patents, and we compare these predictions with actual industry assignments made by the CIPO. We perform a similar exercise for patents granted or published after 1989. The results indicate that the YTC gives a reasonably accurate picture of the pattern of patenting by industry. However, as a result of shifts over time in the relationship between technology fields and industries, our estimated standard errors appear to be much too small.

In the next section, we develop the statistical model that underlies our procedure. The third section then tests the method on the Canadian data. The fourth section concludes.

2. The Statistical Model

We assume that every patent is associated with exactly one of $j = 1, \ldots, \mathcal{J}$ industries and is assigned to exactly one of $i = 1, \ldots, I$ technology fields.[8] Let a_{ij} be the probability that a patent will be associated with industry j, given that it is in technology field i. We have

$$a_{ij} = \Pr[\text{industry} = j \mid \text{technology field} = i]$$

The $\mathcal{J}$-dimensional column vector of these conditional probabilities for a given technology field i is

$$\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{i\mathcal{J}})'$$

These vectors, $i = 1, \ldots, I$, form the rows of an I by J matrix, i.e.

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_I' \end{pmatrix}$$

The key assumption that underlies the YTC is that the matrix **A** does not vary with time or with the country that issues the patent. In particular, we want to estimate **A** using the Canadian data set, and then use it to predict patents by industry in other countries and over other time periods.

2.1. *Inference Conditional on* **A**

In this section, we treat the matrix **A** as if it were known. We assume that we are given data on patents by technology field, $\mathbf{x} = (x_1, \ldots, x_I)'$. These data will be used to predict the unknown random vector, giving the number of patents associated with each industry, i.e. $\mathbf{Y} = (Y_1, \ldots, Y_J)'$. By construction, we have $\sum_{j=1}^{J} Y_j = \sum_{i=1}^{I} x_i$. In the numerical example above, $\mathbf{x} = (60, 100, 10)'$.

Our goal is to derive expressions for the expectation of **Y** conditional on **x** ($E[\mathbf{Y}|\mathbf{x}]$) and for the variance of expectation errors ($E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$, where $\boldsymbol{\varepsilon} = \mathbf{Y} - E[\mathbf{Y}|\mathbf{x}]$).

It is convenient to let Y_{ij} be the random number of patents in technology i and industry j. Then, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ is the random vector of patents in technology i by industry and $\mathbf{Y} = \sum_{i=1}^{I} \mathbf{Y}_i$.

If the matrix **A** of conditional probabilities is actually constant, then the random vector of patents by industry in technology i has a multinomial distribution,[9] i.e.

$$\mathbf{Y}_i \sim M(x_i, \mathbf{a}_i)$$

Thus, $E[\mathbf{Y}_i|\mathbf{x}_i] = x_i\mathbf{a}_i$. Letting $\boldsymbol{\varepsilon}_i = \mathbf{Y}_i - E[\mathbf{Y}_i|x_i]$, $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'] = x_i[\text{diag}(\mathbf{a}_i) - \mathbf{a}_i\mathbf{a}_i']$, where diag(**a**) refers to the diagonal matrix with elements of **a** on the diagonal.

We can now calculate the expectation of patents by industry over all technology fields as

$$E[\mathbf{Y}|\mathbf{x}] = \sum_{i=1}^{I} E[\mathbf{Y}_i|x_i] = \sum_{i=1}^{I} x_i\mathbf{a}_i = \mathbf{A}'\mathbf{x}$$

Furthermore, by the independence of $\mathbf{Y}_i$, we have

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sum_{i=1}^{I} x_i[\text{diag}(\mathbf{a}_i) - \mathbf{a}_i\mathbf{a}_i'] = \text{diag}(\mathbf{A}'\mathbf{x}) - \mathbf{A}'\,\text{diag}(\mathbf{x})\mathbf{A}$$

Returning to the numerical example from the introduction, we have

$$\mathbf{A} = \begin{pmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \\ 1.0 & 0.0 \end{pmatrix}$$

$\mathbf{A}'\mathbf{x} = (70, 100)'$ and

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \begin{pmatrix} 36 & -36 \\ -36 & 36 \end{pmatrix}$$

The square root of the diagonal elements yields the standard error of 6 mentioned in the introduction. Note that the two elements of **Y** are perfectly negatively correlated. An unexpectedly large number of patents associated with industry 1 must come at the expense of industry 2. With more than two industries, there is still no error in the prediction of the sum of patents over all industries, because the sum is known to equal $\sum_{i=1}^{I} x_i$.

2.2. Estimating **A**

To apply the concordance, we will estimate the matrix **A** from the Canadian data set. A superscript 0 will identify variables as being associated with the Canadian data. Thus, x_i^0 is the number of patents in the Canadian data set in technology field i and the random vector of how these patents distribute over industries is $\mathbf{Y}_i^0$. The maximum likelihood estimator (MLE) of the vector $\mathbf{a}_i$ is $\hat{\mathbf{a}}_i = \mathbf{Y}_i^0/x_i^0$. The matrix $\hat{\mathbf{A}}$ has its ith row given by $\hat{\mathbf{a}}_i'$.

In what follows, it is convenient to denote $E[\mathbf{Y}|\mathbf{x}]$ by the symbol $\boldsymbol{\Theta}$. Because $\hat{\mathbf{A}}$ is the MLE of $\mathbf{A}$, and $\boldsymbol{\Theta} = \mathbf{A}'\mathbf{x}$, it follows that the MLE of $\boldsymbol{\Theta}$ is

$$\hat{\boldsymbol{\Theta}} = \hat{\mathbf{A}}'\mathbf{x}$$

Note that $\hat{\boldsymbol{\Theta}}$ is simply a weighted sum of independent multinomial random variables. Its mean is $\boldsymbol{\Theta}$. Thus, letting $\boldsymbol{\eta} = \hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$, the variance matrix of $\boldsymbol{\eta}$ is

$$E[\boldsymbol{\eta}\boldsymbol{\eta}'] = \sum_{i=1}^{I} x_i \frac{x_i}{x_i^0} [\text{diag}(\mathbf{a}_i) - \mathbf{a}_i\mathbf{a}_i']$$

We can now combine the two sources of uncertainty associated with predicting patents by industry. The first source arises because the technology field of a patent generally does not exactly determine the industry with which it will be associated. The second source arises because the probability of an industry assignment, conditional on the technology field of the patent, is not known with certainty, but must be inferred from the finite sample of Canadian assignments. Thus, we have

$$\mathbf{Y} = \hat{\boldsymbol{\Theta}} + \boldsymbol{\varepsilon} - \boldsymbol{\eta} = \hat{\mathbf{A}}'\mathbf{x} + \mathbf{u}$$

where $\mathbf{u} = \boldsymbol{\varepsilon} - \boldsymbol{\eta}$. We make the conservative assumption that $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ are independent.[10] Thus, letting $\boldsymbol{\Omega} = E[\mathbf{u}\mathbf{u}']$, we have

$$\boldsymbol{\Omega} = \sum_{i=1}^{I} x_i\left(1 + \frac{x_i}{x_i^0}\right)[\text{diag}(\mathbf{a}_i) - \mathbf{a}_i\mathbf{a}_i']$$

An estimate $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ is obtained by replacing $\mathbf{a}_i$ in the above formula with an estimate of it, i.e. $\hat{\mathbf{a}}_i$.

Continuing with our numerical example, suppose the matrix $\mathbf{A}$ is an estimate, i.e. $\hat{\mathbf{A}}$, based on data from Canada. If patents by technology field in Canada are $\mathbf{x}^0 = (30, 100, 20)'$, then

$$\hat{\boldsymbol{\Omega}} = \begin{pmatrix} 87 & -87 \\ -87 & 87 \end{pmatrix}$$

The standard error in predicting patents by industry rises from 6 to over 9.

3. Testing the Concordance on the Canadian Data

In applying the YTC, we will use the Canadian data to estimate the conditional probabilities in the matrix **A**. However, the Canadian data also provide a means to test the underlying assumptions of our statistical model. To understand how this can be achieved requires a more complete description of the data set.

3.1. The Canadian Data Set

The Canadian data set (referred to as PATDAT) includes all inventions patented in Canada during 1978–93. We omit patents issued prior to 1983, because industry assignments in the earlier years are less reliable.[11] Essentially all the remaining patents (about 250 000) are assigned an IPC code, at the level known as an 'IPC group'.[12] Almost 7000 distinct groups appear in the data set.

Over 92% of the patents are assigned a four-digit CSIC for the primary SOU, and over 87% are assigned a primary IOM.[13] Less than 8% of patents are assigned more than one IOM, and about 20% are assigned to more than one SOU. In about half these instances of multiple-industry assignment, all four-digit industries are in the same two-digit industry. Because the multiple assignments are relatively infrequent, and are often made to the same broadly defined industry, we ignore all but the primary IOM and the primary SOU.

We maintain two versions of the Canadian data set and construct two concordances: one for SOU and one for IOM. Within each data set, we simply delete patents that are not assigned to an industry.

The inventions patented in Canada come from residents of many countries. Only 7% come from Canadian inventors. About half are from US inventors.

To test the YTC, we split the Canadian data along two dimensions. We split it over time, considering patents issued during 1983–89 against patents issued after 1989. We also split it by country of inventor, considering patents issued to US inventors vs all other patents. For each split, we estimate the matrix $\mathbf{A}$ from the first subsample and then use it to predict patents by industry in the second subsample, based on patents by IPC in the second subsample.[14]

To apply this test, we need to determine the number of technology fields and the number of industries. For the analysis that follows, we consider technology fields at the level of IPC groups.[15]

We aggregate all four-digit CSIC into 27 industries, as listed in Table 1. The 20 manufacturing industries correspond closely to the industries for which the National Science Foundation (NSF) collects R&D expenditures.[16] The seven non-manufacturing industries are defined quite broadly, because they generally use few patents and almost never manufacture them. Although the Canadian data have more industry detail, using this detail increases the potential errors in applying the concordance. For example, a patent in a certain technology field may be quite likely to be used in the primary metals industry, but there may be little information in the technology field about whether it is used in the ferrous or non-ferrous metals subindustry. By considering fairly aggregate industries, we reduce the errors generated by the concordance.

We calculate two summary statistics that give some feel for the relationship between technology fields and industries. These calculations are based on the SOU concept. The patents in a given technology field may all be assigned to one industry, or they may be assigned to several different industries. We find that it is most likely for patents to be in a technology field that maps into nine distinct industries. Thus, it appears that patents in a given technology field will often be dispersed over many industries. However, we find that 63% of patents are assigned to the industry that is the most likely industry, given the technology field of the patent. This statistic suggests that the technology field of the patent contains a lot of information about its SOU.

Table 1. Industry definitions

Sector		CSIC (1980)	SIC (1987)
Manufacturing			
1	Food and related products	10, 11	20
2	Textiles and apparel	18–24	22, 23
3	Lumber and furniture	25, 26	24, 25
4	Paper products	27	26
5	Industrial chemicals	371, 373	281, 282, 286
6	Drugs	374	283
7	Other chemicals	372, 375–379	284, 285, 287–289
8	Petroleum ref. and extraction	36, 7, 9	13, 29
9	Rubber products	15, 16	30
10	Stone, clay and glass products	35	32
11	Primary metals	29	33
12	Fabricated metal products	30	34
13	Office and computing machines	336	357
14	Other non-electrical machinery	31	351–356, 358, 359
15	Communication and electronic	335	366, 367
16	Other electrical equipment	331–334, 337–339	361–365, 369
17	Transportation equipment	323–329	371, 373–375, 379
18	Aircraft and missiles	321	372, 376
19	Prof. and scientific instruments	391	38
20	Other manufacturing	12, 17, 28, 392–399	21, 27, 31, 39
Non-manufacturing			
21	Agriculture and forestry	1–5	1–9
22	Mining (except oil)	6, 8	10, 12, 14
23	Construction	40–44	15–17
24	Transportation and utilities	45–49	40–49
25	Wholesale and retail trade	50–69, 92	50–59
26	Finance, insurance and real estate	70–76	60–67
27	Services and government	77, 81–91, 93–99	70–99

Based on US Bureau of the Census (1991). The following adjustments are made at the three-digit and four-digit levels to the above groupings: (#, Canadian SICs); (2, 3257); (3, 2714); (4, 1691); (11, 3381, 3922); (14, 3081); (19, 1994); (20, 2581); (22, 0921, 0929); (24, 7794, 8631, 996); (26, 4491); (27, 0239, 4411, 4522, 6213, 635, 639, 6562).

3.2. Tests of the YTC

We create two subsets of the data (1 and 2) by breaking the sample along the time dimension or along the residence-of-inventor dimension.[17] The first subset of the data is used to estimate the matrix **A**. This, together with patents by IPC in the second subset, is used to predict patents by industry in the second subset of the data. Let $\hat{\mathbf{A}}(1)$ be the matrix of conditional probabilities estimated from subset 1. Our prediction of patents by industry in subset 2 is

$$E[\mathbf{Y}(2)|\mathbf{x}(2)] = \hat{\mathbf{A}}(1)'\mathbf{x}(2)$$

We compare these predictions with $\mathbf{Y}(2)$ and with a naïve prediction. Our naïve prediction is obtained by scaling patents by industry in subset 1 to account for differences in the number of patents in the two subsets, i.e. n_2/n_1 $[\mathbf{Y}(1)]$, where n_k is the number of patents in subset k, for $k = 1, 2$.

Table 2. Patents from US inventors, by IOM

Sector		Actual	Actual less predicted		Estimated SE
			Naïve	Concordance	
Manufacturing					
1	Food and kindred products	1 294	339.60	141.54	40.940
2	Textiles and apparel	1 576	135.59	106.69	48.943
3	Lumber and furniture	1 214	−300.88	−64.57	42.788
4	Paper products	1 639	410.24	206.23	48.268
5	Industrial chemicals	11 785	−4 624.94	−359.50	110.913
6	Drugs	6 207	180.78	92.67	80.846
7	Other chemicals	8 113	467.05	397.29	102.459
8	Petroleum ref. and extraction	481	103.75	−16.36	28.452
9	Rubber products	5 198	219.26	112.68	90.419
10	Stone, clay and glass products	1 544	−88.47	43.50	48.081
11	Primary metals	1 508	−148.97	−52.45	49.777
12	Fabricated metal products	7 010	−581.08	−161.19	105.772
13	Office and computing machines	5 813	2 021.87	384.56	95.661
14	Other non-electrical machinery	23 042	−3 066.73	−1 096.26	180.286
15	Communication and electronic	13 149	−116.53	−190.75	134.941
16	Other electrical equipment	9 883	1 225.82	277.50	124.320
17	Transportation equipment	5 323	743.07	28.33	88.710
18	Aircraft and missiles	379	51.72	12.74	26.695
19	Prof. and scientific instruments	13 951	3 933.75	286.23	148.080
20	Other manufacturing	4 693	−699.24	−44.08	87.368
Non-manufacturing					
21	Agriculture and forestry	37	−47.27	−16.81	9.686
22	Mining (except oil)	18	−15.32	−14.11	8.016
23	Construction	188	−107.92	−8.14	17.921
24	Transportation and utilities	54	−4.79	−17.07	14.158
25	Wholesale and retail trade	30	−4.30	−1.90	8.055
26	Finance, insurance and real estate	1	−1.00	−1.00	0.000
27	Services and government	116	−27.06	−47.77	18.994
28	Total	124 246	0.00	0.00	N/A

'Actual' refers to patents as classified by the CIPO. The 'naïve' prediction is that the fraction of patents in each industry is the same between patents granted to US inventors and all others. The 'concordance' prediction is as described in the paper, with conditional probabilities estimated from patents granted to non-US inventors. The estimated standard errors (SE) are for the predictions from the concordance.

3.2.1. Patents granted by US inventors. In this test, subset 2 represents patents granted to US inventors and subset 1 is patents granted to non-US inventors. Each subset contains about 130 000 patents.

Table 2 shows the results for the IOM. The first column shows the actual assignment to the IOM of patents granted to US inventors. Note that essentially all the patents are assigned to one of the 20 manufacturing industries. The second column shows the difference between the actual industry assignment and the naïve prediction of patents by industry (assuming that US-inventor patents and those belonging to inventors from elsewhere are assigned to the different industries in the same proportions). The third column is the difference between the actual industry assignment and the predictions by industry based on the YTC. For example, the first row shows that there were 1294 US-inventor patents assigned to the food industry; the naïve procedure underpredicted this number by 370 patents, while the YTC underpredicted this number by 142 patents (hence, the actual prediction

Table 3. Patents from US inventors, SOU

Sector		Actual	Actual less predicted		Estimated SE
			Naïve	Concordance	
Manufacturing					
1	Food and related products	2 748	− 399.81	− 145.879	62.142
2	Textiles and apparel	1 722	− 103.67	20.099	50.767
3	Lumber and furniture	865	− 328.60	− 112.877	38.660
4	Paper products	1 989	− 571.19	− 14.747	53.728
5	Industrial chemicals	8 572	− 1 229.99	− 134.342	111.257
6	Drugs	6 608	− 3 639.53	− 369.479	90.081
7	Other chemicals	4 900	− 1 693.29	− 632.862	94.968
8	Petroleum ref. and extraction	5 258	2 260.96	930.601	91.380
9	Rubber products	5 457	− 26.78	203.208	87.312
10	Stone, clay and glass products	1 740	− 144.63	− 21.490	51.915
11	Primary metals	2 271	− 1 777.56	− 329.253	60.445
12	Fabricated metal products	4 413	76.43	229.587	83.980
13	Office and computing machines	5 943	2 144.75	369.170	99.190
14	Other non-electrical machinery	11 432	951.55	− 88.520	140.615
15	Communication and electronic	12 096	− 40.99	− 43.396	128.601
16	Other electrical equipment	6 710	459.81	− 257.588	107.113
17	Transportation equipment	7 951	1 350.94	269.340	108.214
18	Aircraft and missiles	791	255.57	135.399	37.278
19	Prof. and scientific instruments	4 548	765.22	16.598	91.992
20	Other manufacturing	3 020	− 25.36	− 39.303	70.704
Non-manufacturing					
21	Agriculture and forestry	2 137	− 792.38	− 289.374	57.228
22	Mining (except oil)	653	− 377.26	− 171.889	37.310
23	Construction	5 015	− 1 098.92	− 163.723	88.839
24	Transportation and utilities	4 588	− 269.50	− 125.111	91.335
25	Wholesale and retail trade	2 080	126.75	51.982	61.978
26	Finance, insurance and real estate	119	− 9.54	− 12.843	17.196
27	Services and government	15 928	4 137.01	726.693	153.178
28	Total	129 554	0.00	0.000	N/A

'Actual' refers to patents as classified by the CIPO. The 'naïve' prediction is that the fraction of patents in each industry is the same between patents granted to US inventors and all others. The 'concordance' prediction is as described in the paper, with conditional probabilities estimated from patents granted to non-US inventors. The estimated standard errors (SE) are for the predictions from the concordance.

from the concordance was 1152). In all the manufacturing industries, with the exception of communication equipment, the YTC prediction error is less than the naïve prediction error. The most dramatic improvements are in 'industrial chemicals' (5), 'computing' (13), 'other electrical equipment' (16) and 'instruments' (19).

For each manufacturing industry, except 'food' (1) and 'paper' (4), the prediction error from using the YTC is less than 10% of the actual number of patents assigned to the industry. The last column of Table 2 shows the estimated standard errors of the predictions. In most industries, the absolute value of the YTC prediction error is less than twice its estimated standard error. None the less, there is some evidence that the estimated standard errors understate the true uncertainty attached to the predictions from the YTC.

Table 4. Patents granted or published after 1989, by IOM

Sector		Actual	Actual less predicted		Estimated SE
			Naïve	Concordance	
Manufacturing					
1	Food and related products	1 282	− 12.26	− 62.06	42.722
2	Textiles and apparel	1 613	− 268.01	− 599.91	68.967
3	Lumber and furniture	1 608	95.84	6.36	56.984
4	Paper products	1 754	258.90	− 21.39	61.597
5	Industrial chemicals	17 421	2 836.26	77.29	152.315
6	Drugs	8 924	4 417.71	96.50	158.492
7	Other chemicals	9 111	178.49	− 662.07	128.262
8	Petroleum ref. and extraction	477	− 33.62	− 82.40	33.929
9	Rubber products	6 416	1 345.27	628.92	110.787
10	Stone, clay and glass products	1 652	− 393.09	− 428.45	65.080
11	Primary metals	1 596	− 508.16	− 391.61	62.043
12	Fabricated metal products	7 615	− 1 759.87	− 1 104.23	125.133
13	Office and computing machines	6 662	2 697.83	2 904.12	83.827
14	Other non-electrical machinery	26 544	− 3 834.43	729.17	193.210
15	Communication and electronic	13 000	− 4 966.10	− 3 496.15	154.727
16	Other electrical equipment	8 041	− 5 974.06	− 3 001.29	130.245
17	Transportation equipment	6 092	966.14	1 500.18	84.269
18	Aircraft and missiles	496	211.16	150.73	24.974
19	Prof. and scientific instruments	15 598	4 340.80	4 059.67	143.702
20	Other manufacturing	5 665	− 282.57	− 1 016.63	112.598
Non-manufacturing					
21	Agriculture and forestry	123	− 123.00	− 123.00	0.000
22	Mining (except oil)	12	− 40.51	− 28.79	8.787
23	Construction	367	205.55	230.65	17.272
24	Transportation and utilities	114	− 114.00	− 114.00	0.000
25	Wholesale and retail trade	64	62.69	62.24	2.188
26	Finance, insurance and real estate	1	− 1.00	− 1.00	0.000
27	Services and government	243	218.06	211.14	5.710
28	Total	142 491	0.00	0.00	N/A

'Actual' refers to patents as classified by the CIPO. The 'naïve' prediction is that the fraction of patents in each industry is the same for patents granted before or after the end of 1989. The 'concordance' prediction is as described in the paper, with conditional probabilities estimated from patents granted prior to 1989. The estimated standard errors (SE) are for the predictions from the concordance.

Table 3 presents broadly similar results for SOU. There are now 21 of the 27 industries in which the concordance predictions improve on the naïve predictions. In this application, for both the IOM and the SOU, the YTC performs quite well. It does appear, however, that the estimated standard errors must be viewed sceptically.

3.2.2. Patents granted or published after 1989. In this test, subset 2 is patent applications published after 1989 and subset 1 is patents granted during 1983–89. Subset 2 contains more patents, because it includes applications that were published but that may never be granted.

Table 4 shows the results for the IOM. Once again, the YTC generally outperforms the naïve predictions. However, in this case, there are a number of manufacturing industries—'textiles' (2), 'petroleum' (8), 'stone clay glass' (10), 'primary metals' (11), 'fabricated metals' (12), 'computing' (13), 'communication

Table 5. Patents granted or published after 1989, by SOU

Sector		Actual	Actual less predicted		Estimated SE
			Naïve	Concordance	
Manufacturing					
1	Food and related products	3 956	1 255.48	523.35	85.282
2	Textiles	2 026	− 62.98	− 250.11	65.100
3	Lumber and furniture	1 225	71.78	− 153.73	55.543
4	Paper products	3 011	866.66	402.04	75.133
5	Industrial chemicals	10 295	− 801.00	− 899.47	143.969
6	Drugs	11 565	4 123.73	1 512.80	157.318
7	Other chemicals	6 875	486.79	351.70	114.619
8	Petroleum ref. and extraction	3 317	− 3 328.21	− 1 295.74	84.071
9	Rubber products	6 836	1 175.31	395.63	121.174
10	Stone, clay and glass products	2 061	− 85.98	76.69	61.697
11	Primary metals	3 222	− 1 045.59	365.54	66.715
12	Fabricated metal products	4 567	− 1 143.77	− 677.82	105.717
13	Office and computing machines	6 581	2 242.24	2 110.91	88.303
14	Other non-electrical machinery	12 470	− 454.03	708.23	143.295
15	Communication and electronic	12 546	− 3 411.99	− 2 032.45	148.210
16	Other electrical equipment	6 477	− 2 353.41	− 505.34	110.896
17	Transportation equipment	8 562	366.86	800.73	113.128
18	Aircraft and missiles	883	274.10	150.37	38.639
19	Prof. and scientific instruments	5 893	2 507.13	2 746.86	81.780
20	Other manufacturing	3 509	0.56	− 78.46	89.421
Non-manufacturing					
21	Agriculture and forestry	2 625	− 726.60	− 485.02	72.388
22	Mining (except oil)	635	− 793.68	− 383.87	42.273
23	Construction	6 336	− 260.45	− 549.01	114.640
24	Transportation and utilities	4 743	− 1 676.84	− 934.35	105.671
25	Wholesale and retail trade	2 342	23.69	− 327.19	84.768
26	Finance, insurance and real estate	135	− 19.20	− 69.48	23.754
27	Services and government	17 188	2 769.40	− 1 502.81	197.778
28	Total	149 881	0.00	0.00	N/A

'Actual' refers to patents as classified by the CIPO. The 'naïve' prediction is that the fraction of patents in each industry is the same for patents granted before or after the end of 1989. The 'concordance' prediction is as described in the paper, with conditional probabilities estimated from patents granted prior to 1989. The estimated standard errors (SE) are for the predictions from the concordance.

equipment' (15), 'other electrical equipment' (16), 'transportation equipment' (17), 'aircraft' (18), 'instruments' (19), and (20) 'other'—in which the concordance errors exceed 10% of the actual patents assigned. The prediction errors exceed 25% of the actual patents assigned for six of these industries. In most industries, the estimated standard errors appear to be far too small.

The results for the SOU (Table 5) are much the same. The YTC performs particularly poorly in 'petroleum' (8), 'computing' (13), 'instruments' (19), 'mining' (22) and 'finance' (26). In these five industries, the prediction errors exceed 25% of the actual patents assigned.

In trying to predict patents by industry after 1989, the concordance makes substantial errors.[18] It is likely that errors of this magnitude or greater are a fact of life in using the YTC to predict patents by industry in other countries. Unfortunately, our estimated standard errors may not pin-point the problem industries. This is because the shifts in the underlying conditional probabilities that

generate large prediction errors also invalidate our estimated standard errors. However, it is still the case that the IPC assignments of patents are useful in predicting their industry assignments. This can be seen in Tables 4 and 5, by the tendency of the concordance prediction error to be less than the naïve prediction error.

3.2.3. Annual estimates. We now investigate more carefully the time-series properties of the concordance prediction errors. If the errors in predicting patents by industry after 1989 arose from persistent shifts in the underlying YTC probability matrix **A**, then we should observe the prediction errors within a given industry as being highly correlated over time. However, if errors are highly persistent, then it may be that the variation over time of the predicted patents for a given industry is quite informative about the variation over time in patents actually assigned to the industry.

To pursue these hypotheses, we compare the actual and predicted patents by industry for each year during 1989–93. The matrix **A** is estimated using patents granted during 1983–88. To avoid having the results affected by the discontinuity in patenting between 1989 and 1990, we look at shares of patenting by industry. In other words, both the actual and the predicted patents by industry for a given year are divided by the total patents in that year.

A series of plots presenting the results are available on request from the authors. The bottom line is that the errors in the concordance do display persistence over time. Nevertheless, the concordance predictions often succeed in picking up the time-series variation in actual patent assignments. This is important, because it is often the time-series variation—rather than the overall level—which is of most interest for economic analysis.

4. Conclusions

In this paper, we have presented a formal statistical method for predicting patent counts by industry. It would be a great help to economic researchers if each national patent office made these assignments; however, in fact, only the Canadian patent office does so. We have used the industry and technology assignments of Canadian patents to develop a method that can be used to predict patent counts by industry in all countries that group their patents according to the IPC system (most countries do so).

The method's out-of-sample prediction errors are evaluated on subsets of the Canadian data. The results ae encouraging. The prediction error is generally smaller than would be the case if the predictions were based on constant shares of patents being assigned to each industry. Furthermore, the prediction error is generally moderate relative to the number of patents assigned to an industry (usually less than 10% for patents granted to US inventors and usually less than 25% for patents published after 1989). Even when errors are large, the predictions from the YTC often reflect the time-series variation in the patents assigned to an industry. When compared with alternatives, such as manual assignment of each patent to an industry, the YTC looks quite attractive.

By developing the YTC within a statistical framework, we can estimate standard errors for our predictions. In theory, these standard errors could be

incorporated into subsequent studies that use the patents by industry as data. In practice, it appears that these standard errors have a considerable downward bias. Our hypothesis is that the instability in the concordance relationships over time leads us to underestimate the true standard errors.[19]

If the YTC proves to be a useful tool for researchers, then it will be applied over different time periods to patent data from many different countries. This paper provides some evidence on the magnitude of the industry assignment errors that are likely to result. We have shown that shifts in industry assignments conditional on technology assignments can lead to large prediction errors over time in some industries. We also provide indirect evidence on how well the concordance will work when applied in countries other than Canada. As shown in Tables 2 and 3, we predict quite accurately the industry assignments of US-inventor patents, based on the industry and technology assignments of non-US-inventor patents. Thus, variation in the source country of patented inventions may pose few problems for the YTC. Of course, all our evidence is based on the selected sample of inventions that were worth patenting in Canada. It is possible that the country in which patent protection is sought matters more for the reliability of the YTC than does the source country. We await comparisons with industry-level data in other countries to resolve this issue.

Notes

1. A number of papers have used early versions of the Yale Technology Concordance. For example, Evenson (1991) and Kortum (1993) look at the behavior of the patent: R&D ratio at the industry level, while Kortum and Lach (1995) estimate the relationship between patenting and productivity growth in US industries. Earlier efforts to count patents at the industry level include those of Schmookler (1966) and Scherer (1984).
2. We use the IPC at a level of detail that includes about 6000 distinct technology fields.
3. The industries are defined at the four-digit level of the Canadian Standard Industrial Classification (CSIC) system. The task of assigning industries is performed by a staff of patent classifiers. These individuals are trained in the industry classification system, although their main job is to assign technology codes. The CSIC codes are similar, but not identical, to US SIC codes. For example, there is no classification for 'rockets' under the CSIC. However, it is possible (with the help of the US Bureau of the Census (1991)) to aggregate the Canadian codes to the level at which internationally comparable R&D data are collected by the Organization for Economic Cooperation and Development (OECD). In this paper, we construct aggregates that match industries for which R&D data are collected in the US.
4. See Johnson (this issue) (1995) for a comparison of patents by industry as assigned by the CIPO and industry-level data from innovation surveys.
5. Since October 1989, Canada has published patent applications 18 months after the date of application, regardless of whether or not the patents would eventually be granted. The data set includes all these published applications.
6. The USPC is very detailed, including 100 000 classes (about 50 000 in Schmookler's time). It is even more 'technologically oriented' (hence, less 'industry oriented') than the IPC. This means that patents are classified by how they accomplish a certain task, rather than by the service they provide. For example, a heart pump is a 'pump' under the USPC, and not a 'medical device'.
7. The OTAF concordance is described in Marmor *et al.* (1979). Scherer (1982) points to several deficiencies in the original concordance. An extensive review and assessment are provided in Patent and Trademark Office (1985).
8. We assume that a patent is assigned to only one IOM and only one SOU. In reality, the CIPO assigned up to three IOMs and three SOUs; also, patent offices often assign multiple IPCs to a single invention. Incorporating this additional information, by identifying 'principle' and 'secondary' assignments, complicates the methodological issues and requires additional assumptions on the appropriate weights to attach to each assignment. Therefore, we assume that the first classification is the only one. In the following section, we discuss why this is a reasonable assumption.

9. See Bishop *et al.* (1977, p. 63) for a discussion of the sampling scheme that generates the multinomial distribution.
10. Because the same invention may be patented in several countries, it is likely that ϵ and η are positively correlated. By assuming independence, we obtain an upper bound on the variance of the difference between them ($\mathbf{u}$).
11. Tabulations of the data indicate a discontinuity in patenting by the SIC, pre- and post-1982. Individuals at the CIPO recommended using the post-1982 data, because industry assignment procedures improved and industry definitions stabilized.
12. For example, group D05C 009 contains patents on automatic embroidering machines.
13. Some patents, which are of very general use, are not assigned an SOU, and process patents, with no tangible apparatus, are not assigned an IOM. Canadian patents are assigned to four-digit industries as defined in the 1980 version of the CSIC. A concordance between the 1980 version of the CSIC and the 1987 version of the SIC has been developed by the US Census and Statistics Canada (US Bureau of the Census, 1991). About 4% of patents are assigned to a four-digit SIC ending in zero, such as 2710. Such industries do not appear in the CSIC system. In these cases, the industry assignment is only meaningful at the three-digit level.
14. In Kortum and Putnam (1989a), χ^2 tests of the stability of the $\mathbf{A}$ matrix were performed. Those tests always reject the null hypothesis of a stable $\mathbf{A}$ matrix, but they are not informative about how well the YTC will make out-of-sample predictions in practice.
15. In principle, because we condition on patents by technology field, the more detailed the technological classification is, the better will be the results. In practice, moving to greater detail entails considerable computational and data storage burdens. Furthermore, the definitions in the IPC are more likely to change over time at the subgroup level of the IPC. Keeping track of such changes as they are adopted in different countries would be extremely difficult.
16. The relationship between the CSIC and the SIC shown in Table 1 was checked against the concordance produced by the US Census and Statistics Canada (US Bureau of the Census, 1991).
17. The objective was to break the sample along a dimension in which shifts in the concordance relationships may have occurred. Another approach would be to create two data sets by randomly sampling from the original data set. This second approach would essentially mimic the theory used to derive the YTC. In this sense, it would not provide a powerful test of the YTC.
18. A factor that may be contributing to errors over time is changes in the IPC. The IPC system is supposed to be updated by the World Intellectual Property Organization every 5 years. We have chosen to ignore these updates, but it may be possible to keep track of the version of the IPC in applying the concordance.
19. A possible source of instability in the concordance is shifts over time in the industry structure in Canada. One method of modelling these shifts, but one that requires additional restrictions along other dimensions, is pursued by Kortum and Putnam (1989b). Unfortunately, in practice, that method did not turn out to be an improvement over the approach followed here.

References

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1977) *Discrete Multivariate Analysis: Theory and Practice* (Cambridge, MA, MIT Press).

Ellis, E. D. (1981) The philosophy, construction and uses of the Canadian patent data base PATDAT, *World Patent Information.*

Evenson, R. E. (1991) Patent data by industry: evidence for invention potential exhaustion?, in: *Technology and Productivity: The Challenge for Economic Policy* (Paris, Organization for Economic Cooperation and Development), pp. 233–248.

Evenson, R. E., Putnam, J. & Kortum, S. (1991) Estimating patent counts by industry using the Yale–Canada concordance, *Final Report to the National Science Foundation.*

Griliches, Z. (1990) Patent statistics as economic indicators: a survey, *Journal of Economic Literature*, December, pp. 1661–1707.

Kortum, S. (1993) Equilibrium R&D and the patent–R&D ratio: US evidence, *American Economic Review: Papers and Proceedings*, 83, pp. 450–457.

Kortum, S. & Putnam, J. (1989a) Estimating patents by industry: Part I, unpublished.

Kortum, S. & Putnam, J. (1989b) Estimating patents by industry: Part II, unpublished.

Kortum, S. & Lach, S. (1995) Patents and productivity growth in US manufacturing industries, unpublished.

Marmor, A. C., Lawson, W. S. & Terapane, J. F. (1979) The technology assessment and forecast program of the United States Patent Office, *World Patent Information*, 1, pp. 15–23.

Patent and Trademark Office (1985) *Review and Assessment of the OTAF Concordance between the US Patent Classification and the Standard Industrial Classification Systems: Final Report* (Office of Technology Assessment and Forecast, January).

Schmookler, J. (1966) *Invention and Economic Growth* (Cambridge, MA, Harvard University Press).

Scherer, F. M. (1982) The Office of Technology Assessment and Forecast industry concordance as a means of identifying industry technology origins, *World Patent Information*, 4, pp. 12–17.

Scherer, F. M. (1984) Using linked patent and R&D data to measure interindustry technology flows, in: Z. Griliches (ed.) *R&D, Patents, and Productivity* (Chicago, IL, University of Chicago Press), pp. 417–464.

US Bureau of the Census (1991) *Concordance between the Standard Industrial Classifications of the United States and Canada: 1987 United States SIC—1980 Canadian SIC* (US Bureau of the Census and Statistics Canada, February).