QUCEH WORKING PAPER SERIES
http://www.quceh.org.uk/working-papers

THAT’S CLASSIFIED!
INVENTING A NEW PATENT TAXONOMY

Stephen D. Billington (Queen’s University Belfast)
Alan J. Hanna (Queen’s University Belfast)

Working Paper 2018-06

QUEEN’S UNIVERSITY CENTRE FOR ECONOMIC HISTORY
Queen’s University Belfast
185 Stranmillis Road
Belfast BT9 5EE

June 2018

That's Classified! Inventing a New Patent Taxonomy[†]

Stephen D. Billington[‡]   Alan J. Hanna [**]

June 2018

Abstract

We investigate how patent classification influences the interpretation of patent statistics. Innovation researchers currently make use of various patent classification schemas. Their classification methodologies are hard to replicate. Using machine learning techniques, we construct a transparent, replicable patent taxonomy, and a new automated methodology for classifying patents. We then contrast our new schema with existing ones using a long-run patent dataset. In a quantitative analysis of patent characteristics, we find strong evidence of classification bias; our interpretation of regression coefficients is schema-dependant. We suggest that much of the innovation literature needs to be re-examined in light of our findings.

Keywords: Innovation, Invention, Machine Learning, Patents, Patent Classification, Taxonomy, Economic History.

JEL Codes: K11, N24, N74, 031, 033.

[†] Billington thanks Chris Colvin and Christopher Coyle for their detailed feedback and guidance in developing this paper, and Alessandro Nuvolari for hosting him on a visit to the Sant'Anna School of Advanced Studies, Pisa, in March 2018. Thanks also to Graham Brownlow, Alan de Bromhead, Nola Hewitt-Dundas and David Jordan for their helpful comments.

[‡] Corresponding author: Queen's Management School, Queen's University Belfast. Email: sbillington01@qub.ac.uk.

[**] Queen's Management School, Queen's University Belfast. Email: a.hanna@qub.ac.uk.

1 Introduction

Patent statistics are a widely used proxy for measuring technological change (Griliches, 1990).[1] Patentable inventions, however, have heterogeneous characteristics, which, if not accurately controlled for, have the potential to bias any interpretation of patent statistics. Biased statistics are likely to lead to ineffective policy measures. For example, the propensity to patent varies by industry, suggesting the decision to obtain a patent also varies by industry (Moser, 2005). A single system of classification is necessary to account for such characteristics consistently across studies. The innovation literature, however, does not have a standard, re-usable taxonomy. Some studies use the section headings of the International Patent Classification (IPC) schema (Nicholas, 2011c). Others employ various industrial classification taxonomies (Phillips, 1966; Rajan and Zingales, 1998; Aghion et al., 2002; Walsh et al., 2016). Still, others use historical classes derived from prize-giving institutions (Moser, 2005; Moser, 2012; Khan, 2013b; Khan, 2016). The inconsistency raises the following questions: how comparable are existing studies? And which, if any, of the prevailing taxonomies can and should be used in future studies?

Existing taxonomies can be divided into two types: "Official" and "Academic". Official taxonomies are produced by and for patent offices.[2] Academic schemas exist independent of patent offices, for conducting innovation research. Official taxonomies are limited in their scope: they either group too many unrelated patents because patents are classed by their technical functions, or too few related patents because of too many subclasses. Academic schemas, likewise, are limited because they are often not fully discussed or described. This makes academic taxonomies difficult to replicate or re-use consistently. At present, both types of taxonomies can complicate our ability to interpret the existing literature.

Patents can also be classified under "static" or "dynamic" schemas. Static schemas

[1] A patent is a temporary monopoly right granted to a particular novel and non-obvious invention or process; it provides the holder with the legal power to prevent replication or copying without express permission (Scotchmer, 1991; Scotchmer, 2004).

[2] Patent classification schemas have been developed to benefit patent examiners, rather than patent researchers (e.g., WIPO, 2016). The thousands of unique patent classes within schemas, such as the IPC, have greatly reduced the associated costs of patent examination.

consist of broad classes that do not change over time. Dynamic schemas encompass much more detailed classes, reflecting time- or country-specific innovation. This approach is highly useful for observing specific types of inventions, and also emerging technologies. By contrast, static taxonomies are useful for a comparative analysis as comparisons can be made over the long-run and between studies. Modern patent systems do not change much, having largely homogenised over time. The long-run, however, encompasses numerous periods of patent reform. Such events act as natural experiments, which can provide important insights concerning the optimal design of patent institutions. The ability to contrast patents throughout history is important for developing a complete understanding of how patent systems encourage innovation, and how they have developed over time. For this reason, we opt for developing a static taxonomy.

Our goal is to design a new, static patent taxonomy, for producing more consistent and comparable results within the innovation literature. We base our taxonomy upon the principle of transparency, so that future investigators can understand how our taxonomy is designed. In this way, investigators can either: reuse our schema, re-purpose it for their own needs, or even develop new schemas using our methods. We also propose a new methodology for automating patent classification to ensure patent data are grouped consistently. Our approach is to adopt machine learning techniques that can classify any patent data using any patent taxonomy. Machine learning techniques minimize the subjective element of classification, reducing the probability that some patents are classified incorrectly. Establishing a consistent approach to patent classification is likely to lead to increased comparability of innovation studies, which can only benefit policymakers in designing appropriate measures to encourage innovation.

The first half of this paper is concerned with developing a new taxonomy, and a method for classifying patent data. This methodology focuses on using text as data to derive our set of static patent classes. Because the literature abounds of competing patent taxonomies, we can observe which classes appear frequently. Frequent classes in contemporary and historical taxonomies reflect technology groups that exist independent of time or geography. These classes are then likely to represent static classes. Then, we apply machine learning techniques to the patent data to check if our static classes are valid. Patent data contain rich textual information in their titles (and in their abstracts). Using these titles, we elicit a set of common word associations, or “topics”. Topics capture specific technology groups, and can be used to observe whether we have omitted any potential classes; we derive topics from multiple patent datasets to check this. Finally, we use our machine learning techniques to automate the patent classification process.

The second half of our paper is focused on whether the choice of schema influences the results of examining patent characteristics – “classification bias”. To test this, we examine the population of British patents granted between 1700 and 1850. There are several advantages to using this dataset. First, taxonomies that have classified the data, such as Nuvolari and Tartari (2011), Bottomley (2014a), and Dowey (2017), can be replicated. Second, this data spans the period of the Industrial Revolution; any insights are important to our understanding of this phenomenon. Third, the dataset is relatively small, making manual assignments and comparisons of classes simpler, as well as reducing the time needed to run our machine learning techniques. Our results show that classification bias does exist; the magnitude, sign, and significance of coefficients in a regression analysis of patent characteristics against patent classes each depend on the taxonomy used.

This study contributes to the literature as follows. First, we present a new, well-defined, static patent taxonomy. We thoroughly describe the development of our schema to ensure future users understand how it was constructed and how it can be used. Second, we provide a new methodology for automating the classification of any patent dataset. This method classifies similar patents in a similar manner, leading to a classification process that is more consistent and has fewer errors. Our approach significantly reduces the subjective element associated with patent classification, and also drastically decreases the time needed to classify large patent datasets, reducing the opportunity cost of engaging in any large-scale analysis of patenting. Third, we identify classification bias. This bias makes it difficult for policymakers to develop measures that encourage innovation based on the existing literature. This is particularly important for research into the current industrial strategy of the UK Government

(HM Government, 2017). Any attempt to identify sector-specific innovation issues as part of the Government's "sector deals" policy may be hampered by the choice of classification schema.

Our paper is closely related to the seminal article of Lybbert and Zolas (2014). In their paper, the authors develop a concordance between the IPC and modern industrial classification schemas, to facilitate long-run analysis. They construct their concordance using a probabilistic algorithm, which matches keywords in modern industrial classes to patent titles from the European Patent Office's PATSTAT database. However, their method runs into the following difficulties. First, PATSTAT's historical collection is incomplete; any concordance using PATSTAT is less capable of contrasting patents over the long-run. Second, because their concordance uses keywords from modern industrial schemas, it cannot observe time- or region-specific terms, which then omits potentially useful text for classification. Our approach overcomes these drawbacks. The key difference is that we construct a new schema and not a concordance, so classification does not rely on any-pre-existing classification codes. This schema also accompanies a new methodology for classifying patent data. Our methodology uses only the text contained in patent titles for the purposes of classification. This allows us to exploit the entire set of unique words contained in patents titles from any patent dataset. In addition, we ask a new question: whether classification bias exists. Because existing studies inform policy measures, the implications from any bias for industrial policy are likely to be serious and therefore need to be acknowledged and understood.

The remainder of this paper is outlined as follows: Section 2 surveys the existing literature concerning patent classification. Section 3 discusses the machine learning techniques used in this present study. Section 4 details how we derived our static patent taxonomy. Section 5 outlines the data used in this study to test the efficacy of our new taxonomy. Section 6 provides the results of contrasting patent taxonomies in our analysis of patent classes upon patent characteristics. Section 7 discusses the implications of our findings for the study of innovation. Finally Section 8 concludes with some recommendations for future scholarship.

2 Patent Classification Literature

The development of official schemas has primarily been to aid patent examiners (WIPO, 1992). Examining patents usually requires examiners to engage in the time-consuming search for prior art: previous patents that are likely to influence or anticipate future ones. Having thousands of well-defined classes and subclasses facilitates a more efficient search process. Such classes make fine distinctions between seemingly similar types of inventions, allowing examiners to find the relevant art more effectively.

However, this approach is not universally appropriate. Academic studies do not need patent classes to search for prior art; they need them to control for any common patent characteristics likely to bias the study of patenting. Official schemas are not capable of doing this for two reasons. First, section headings are too broad for use, as these group together too many unrelated patents. The IPC has eight section headings, each of which captures many diverse types of invention. Second, official subclasses are too numerous, resulting in too few patents per class; the IPC has 61,397 subgroupings. For this reason, academic studies often develop their own taxonomies, which consist of fewer, broader classes. Replicating academic taxonomies, however, is not so easy: the absence of documentation means it is not always clear how authors constructed their schemas.

In 1830, John D. Craig, the US Superintendent of Patents, gave evidence to the US House of Representatives regarding the development of the US classification schema. In his evidence, Craig raised two points: the 'imperceptible shades of difference' of patent classes, and that 'a doubt frequently arose concerning the class to which some of the patents did properly belong' (cited in Bailey, 1946: p. 466). Craig's concern was that patented inventions have overlapping characteristics. Accurately pinpointing a particular class for a particular patent is then a difficult task. Modern schemas encounter this same difficulty, especially in instances where authors only assign one class per patent.

It is precisely because assigning classes is difficult that a standardised schema is necessary. Without a standard taxonomy, the inconsistent classification of patent data becomes highly likely. Pearce (1957) discusses the inconsistency problem in the context of industrial statistics. Prior to 1937, various agencies collected industrial statistics, but these agencies used different classifications to code their data. Any resulting comparisons of industrial data became 'difficult and often misleading' (Pearce, 1957: p. 1). This resulted in the creation of the Standard Industrial Classification (SIC) schema in the same year, for standardizing the classification of industrial data.

Such concerns are just as relevant for the current innovation literature that relies heavily on patent data to inform industrial policy. At present, different authors classify (often the same) patent data in different ways. Patent classes are supposed to account for common patent characteristics that could bias how we interpret such data. For example, "machine" inventions are often patented because they can be easily reverse-engineered (Moser, 2005). "Chemical" inventions, by comparison, are harder to reverse-engineer. In this case, inventors acquire patents for different reasons that need to be consistently accounted for. If scholars fail to take the same approach to identifying group-specific characteristics, different studies cannot be easily compared.

The innovation literature has been relatively quiet with regard to the development of patent taxonomies. Most studies use their classes as a set of industry controls for their econometrics. What is unclear, however, is how authors have constructed their taxonomies, defined their classes, and assigned patents to those classes. Without this information, replicating existing taxonomies is difficult. This leads to two possible outcomes. Either future investigators apply existing taxonomies incorrectly – leading to further inconsistencies – or they produce additional taxonomies, which may not be comparable with existing ones.

Table 1 details a sample of established taxonomies, both official and academic, which have been used in conjunction with patent data.[3] The table documents the source of the taxonomy, the number of classes within the taxonomy, if there is an accompanying description of its development, and if the classes receive definitions. As can be seen, the number of classes varies across studies, while few schemas are described in detail.[4] Official

[3] This table includes only unique taxonomies. The IPC, for example, has been used in multiple studies, but is included once. However, we do include studies that adapt existing codes because they produce a taxonomy different to the original.

[4] Not all of the provided taxonomies are for patents, some are for inventions submitted to prize-giving exhibitions. Since these act to group inventions much like a patent class, and crucially to compare with patents, we treat them as the same. Industrial taxonomies are also not intended for patent data, but

articles are more likely to discuss their schema and methodology in detail: 10 out of 13 articles explicitly describe their classes, while eight out of 13 provide their methodology. Such documentation is probably why seven out of the 23 listed academic articles adapt or adopt official schemas in some way.

Replicating taxonomies requires that they detail how to group patents. However, there are multiple methods for doing so. Table 2 provides an overview of the most common approaches to classification, which are relevant for developing a new patent taxonomy. In his seminal article, Griliches (1990) outlines three methods of classification: "Origin", "Production", and "Destination". Origin groups patents by the industry that produced them; this is suitable for examining R&D expenditure, as R&D occurs within a given industry. Production groups patents by the industry most likely to produce the invention, or use within the production of goods or services. Destination groups patents according to the industry most likely to make use of, or benefit from, the invention. Destination and Production overlap to some degree. The major difference is that the use of an invention does not intrinsically imply its use is in production, but an invention used within the production process constitutes Destination. Destination is likely to be the most suitable method for studying patenting within the wider economy, as it is easier to determine the intended industry of a patent.

Industrial taxonomies are often used when classifying patent data, for studying firm or industry innovation (e.g. Baten et al., 2007; Nicholas, 2011b; Schautschick, 2015). These taxonomies classify firms or industries by their "supply-side" or "demand-side" characteristics. The supply-side method groups firms according to their production process, or by their activities (Statistics Division, 2008; S&P Capital IQ and MSCI, 2015). For patents, this is not suitable. While this is useful for the study of firms and their acquisition of patents, it is less useful for studying the effects of patents in the wider economy. Firms may have similar production processes, but can produce entirely different output. This, in turn, suggests that the patents they obtain could be for widely different applications. Similar to official schemas, the supply-side method is

authors use them anyway. Because of this, they are also included and treated as patent schemas.

Table 1: *Classification Literature*

Authors	**Classes**	**Method**	**Definitions**
Academic Literature			
Bain (1954)	20	Census of Manufactures (1947)	No
Baten et al. (2007)	19	SIC Codes	No
Brunt et al. (2012)	10	Woodcroft Subject-Matter Index	No
Burhop and Wolf (2013)	9	None Provided	No
Galasso and Schankerman (2015)	6	None Provided	No
Hall et al. (2001)	6	USPTO codes	No
Khan (2013a)	12	Massachusetts Mechanics Association Fair classes	No
Khan (2013b)	4	None Provided	No
Khan (2017)	6	Royal Society for Arts	No
Khan (2015)	26	Annuaire de la Societe d'Encouragement pour l'Industrie Nationale	No
Krueger and Summers (1988)	7	CIC codes	No
Lampe and Moser (2016)	20	USPTO subclasses	No
Lehmann-Hasemeyer and Streb (2016)	5	None Provided	No
Moser (2012)	10	Historical Exhibitions	No
Moser (2005)	7	Crystal Palace Exhibition	No
Nanda and Nicholas (2014)	15	Mapping to SIC	No
Nicholas (2008)	3	Description of business activities	No
Nicholas (2011a)	30	Based on IPC	No
Nicholas (2011b)	16	Based on SIC	No
Nuvolari and Tartari (2011)	21	Expansion upon Moser Working Paper	No
Rajan and Zingales (1998)	36	SIC	No
Schautschick (2015)	8	Two-digit NACE codes	No
Sokoloff (1988)	4	None Provided	No
Official Literature			
British Patent Office-Austrian Scheme (1915)	89	None Provided	Yes
British Patent Office-French Scheme (1915)	20	None Provided	No
British Patent Office-German Scheme (1915)	89	None Provided	Yes
British Patent Office-Swiss Scheme (1915)	129	None Provided	Yes
British Patent Office (2007)	8	Supply-side methodology	No
SIC (2007)	21	Supply-side methodology	Yes
ISIC (2008)	21	Supply-side methodology	Yes
NACE (2008)	21	Supply-side methodology	Yes
GICS (2016)	67	Supply-side methodology	Yes
IPC (2016)	8	Supply-side methodology	Yes
NAICS (2017)	20	Supply-side methodology	Yes
Woodcroft (1860)	246	Patent titles	No
A Cradle of Inventions (2009)	15	None Provided	Yes

Notes: The table shows a sample of 36 published patent taxonomies. 'Classes' shows the number of individual patent classes in each taxonomy, at the broadest level. 'Method' details how the classes were constructed, or where they were adapted from. 'Definitions' states whether the article provided a list of definitions for their classes.

Sources: Official Industry Publications: SIC, ISIC, GICS, NAICS. Academic Literature: Bain (1954), Krueger and Summers (1988), Sokoloff (1988), Rajan and Zingales (1998), Hall et al. (2001), Moser (2005), Baten et al. (2007), Nicholas (2008), Nicholas (2011c), Nicholas (2011b), Nuvolari and Tartari (2011), Brunt et al. (2012), Moser (2012), Burhop and Wolf (2013), Khan (2013b), Khan (2013a), Bottomley (2014a), Nanda and Nicholas (2014), Schautschick (2015), Lampe and Moser (2016), Lehmann-Hasemeyer and Streb (2016), WIPO (2016), and Khan (2017). Historical Publications: Woodcroft (1860), and A Cradle of Inventions: British Patents from 1617 to 1894. All British Patent Office schemas: Franks (1915).

Table 2: *Methodological Considerations for Classifying Patents*

Approach	**Description**	**Advantages**	**Disadvantages**
Dynamic	Class patents by evolving classes	Identification of rising technologies; identification of periods of 'patent-mania'; consistency in use of IPC and industrial codes	Not appropriate for historical analysis; relies entirely on assigned IPC codes
Static	Class patents by fixed classes	Allows for historical comparison; identification of broad classes which rise and fall over time; comparability across countries and time; does not require IPC	Not useful for identifying niche technology fields; reliant on accurate identification of fixed classes
Destination	Class patents by the industry most likely to use them	Allows for analysis of how patenting activity influences economic indicators: GDP, Productivity; identification of fields inventor intended their invention for	Cannot be sure where inventor intended their patent to go; reliant on subjective classification of each patent
Origin	Class patents by the industry which invented them	Allows for analysis of which industries contribute most to technological progress and knowledge output	Reliant on data containing detailed information on the occupation of the inventor, or if the patentee is a firm; doesn't account for the influence inventions have, only industry output
Production	Class patents by the industry most likely to produce them, or use them in the production process	Allows for analysis of investment activity, and the relationship between output and investment	Limits observations to inventions requiring a production process, e.g. manufactured goods; different classes will be captured that share the same production process
Demand Side	Class patents along industry lines, where industry is defined by its close substitutes	Allows for comparison with the economists' definition of industry	Patents do not technically have substitutes; requires the accurate identification of related inventions
Supply Side	Class patents along industry lines, where industry is defined by its production process	Allows for analysis of investment activity, and the relationship between output and investment	Groups inventions by the firms production process, which is likely to group unrelated technologies together

Notes: The table details the most important aspects of constructing patent classes, and assigning patents to these classes.

Sources: Dynamic and Static are our own definitions. Industry of Origin, Production, and Destination (Griliches, 1990). Demand-side and Supply-side (ECPC, 1992; ECPC, 1993; ECPC, 1994; WIPO, 2016).

likely to group unrelated inventions, which then do not accurately account for group-specific characteristics of patents.

Alternatively, the demand-side approach groups firms by their competitors (e.g. Bain, 1951; Bain, 1954; ECPC, 1992; ECPC, 1993; ECPC, 1994; WIPO, 2016). Patents, however, are not the same as industries or firms, and should not be grouped in the same manner for the purpose of economic enquiry. Patents and inventions can be either "macro-inventions" or "micro-inventions" (Mokyr, 2009). A macro-invention is a substantially new technology, while a micro-invention complements this new technology by improving upon it with incremental advancements. Micro-inventions complement macro-inventions, as they improve upon minor aspects of the technology, increasing their cost-effectiveness or productivity. Patents, therefore, do not necessarily compete with each other. Studying the wider economy requires examining how an invention or innovation affects that economy. For this purpose, patents should be characterised and classified according to their applications.

The dynamic approach is another useful method for classification. This method classes patents by the specific details of new technologies, rather than by any application of it. For example, dynamic schemas classify inventions that improve the use of UV lighting in farming based on the use of UV lighting, but not on its use in agriculture. In this way, dynamic schemas are much like official schemas (WIPO, 2016). Accordingly, this method may be more appropriate for identifying emerging technologies, and also for observing the increasing modularisation of technology. The most common approach to producing a dynamic schema is to develop a 'concordance' between existing schemas (e.g. Verspagen et al., 1994; Kortum and Putnam, 1997; Johnson, 2002; Schmoch et al., 2003). Recent developments in this literature, however, have been to adopt statistical methods for classifying patents. Lybbert and Zolas (2014), for example, have pioneered a new approach of using probabilistic algorithms to match IPC classes to industrial schemas. By matching keywords contained within existing industrial schemas to keywords in patent titles, they attempt to reduce the subjective element of concordance mapping.

Such developments are of great value for examining the changing nature of technology, but are less applicable to producing a long-run static taxonomy for studying the wider economy. The importance of studying the long-run is that it contains numerous periods of technological change. Such periods can provide important lessons for directing policy measures to encourage innovation. The Industrial Revolution (1760-1830), for example, was the first period of major technological change historically, and one of the most significant events in human history. Comparisons over the long-run allow us to draw lessons from this period, which can then be used to encourage innovation, which may be useful to the ongoing "Fourth Industrial Revolution".

Dynamic schemas cannot make such comparisons. First, the dynamic method requires patent data to have pre-existing IPC codes. Patents without these codes – such as historical patents – are then effectively neglected, hampering the ability of dynamic schemas to observe the long-run. Second, studying patenting behaviour and its effects on the wider economy requires observing the Destination of patents, rather than their technical function. Technical functions reveal nothing about how a particular invention influences productivity levels, or R&D expenditure, or economic development, because this method groups together otherwise unrelated patents. Mapping to existing industrial taxonomies, which also use the supply-side methodology, further limits the capability of dynamic schemas to study the applications of patents.

Given the limitations of the official and dynamic classification schemas, this study opts to produce a new static taxonomy, based upon the application or Destination of patents. To effectively produce such a schema, we turn to machine learning techniques. The design of any schema should be replicable in future studies, and should minimize the subjective decision making of the investigator. Here, text analysis techniques are most useful: they can be easily replicated, and they do not require much human judgement as they rely strongly on the text contained in patent data.

3 Machine Learning Approach

Text analysis uses text as data, by attempting to extract subjective information from any natural language. One approach is to classify articles based on the common associations of particular words contained within them. The choice of words in a given article of text are conditional on the theme of the article; the words used can identify the relevant classification. Such associations can be derived from either "unsupervised" or "supervised" machine learning techniques. Unsupervised methodologies seek to find hidden associations between observations. Supervised techniques, by contrast, use known classifications to train a particular model. We opt for unsupervised techniques, so as to allow the data to derive latent patent groups.

Such techniques are particularly useful for classifying patent data. The data contain detailed titles, which are required to describe the nature of the invention that a patent covers. Under the European Patent Convention, for example, applications are checked by patent examiners to verify the accuracy of their titles (EPO, 2017). In instances where the title does not match the invention, the examiner can amend the title as they see fit. Therefore, patent titles provide a rich source of textual data, which can be used to identify patent classes.[5]

In text analysis, words that commonly appear together are called topics. One method to identify a set of topics is Non-Negative Matrix Factorization (NMF). Here, the dataset, or "corpus", is represented as a matrix composed of word frequencies for each article (row) and word (column). Frequencies can be simple term counts, but following O'Callaghan et al. (2015) we adopt a log-based term frequency-inverse document frequency (TF-IDF) representation, which helps to counter the influence of words that appear more frequently throughout the corpus. "Stop words" are entirely removed from the corpus.[6] The matrix is then approximately decomposed into the product of two non-negative matrices. Here,

[5] For the data used in this study, the titles are considered to reflect the invention: patents could be annulled if the invention was not properly described (Dutton, 1984; MacLeod, 2002; Bottomley, 2014b).

[6] The term stop words is used to describe words which are most commonly used in a particular language (for example the conjunctions like 'and', 'if', or 'when', and prepositions like 'to', 'with' or 'in'). Such words are unhelpful in understanding the content of the corpus and are therefore ignored. Stop words were sourced from http://www.ranks.nl/stopwords.

articles are represented in terms of scores relating to each topic, and each topic by scores relating to their use of words.

To understand how the NMF approach works, consider the following example. Suppose we have a corpus – a collection of patents in this instance – containing m patent titles, each composed from a set of n unique words. This corpus is represented by the matrix C, where $c_{i,j}$ represents, for each document i, the number of occurrences of word j. NMF attempts to factorize this matrix by approximating it as the product of two smaller non-negative matrices. This is represented as:

$$AT \approx C \tag{1}$$

where matrix T represents how often each word occurs within each topic. The weights in matrix A then reveal the extent to which a patent relates to each topic. Word associations define their topics, which allows them to be interpreted by the investigator for further classification.

Before deriving a set of topics, we first need to determine how many topics to produce. When using topic scores to classify patents, the number of topics influences where each patent is assigned.[7] Initially, we generated topics in multiples of 20, and examined the differences between them. Fewer topics were associated with less consistent word associations, while additional topics alleviated this inconsistency. Therefore, we adopt a set of objective measures to determine the number of topics to use. We rely on three separate measures: the Residual Sum of Squares (RSS); Entropy scores; and Coherence scores. For each measure, we derived a range of different numbers of topics: 10, 20, 30, 40, 50, 60, 70, 80, 100, 120, 150, and 200. These are displayed in Figure 1. Future investigators should reproduce these measures when determining how many topics to use; it is unlikely that the optimal number of topics would be the same for all studies.

Firstly, we compute RSS scores. The RSS measures the quality of the approximation to the original document term frequency matrix. This metric decreases with each

[7] We generate the optimal number of topics from our British dataset, described in Section 5.

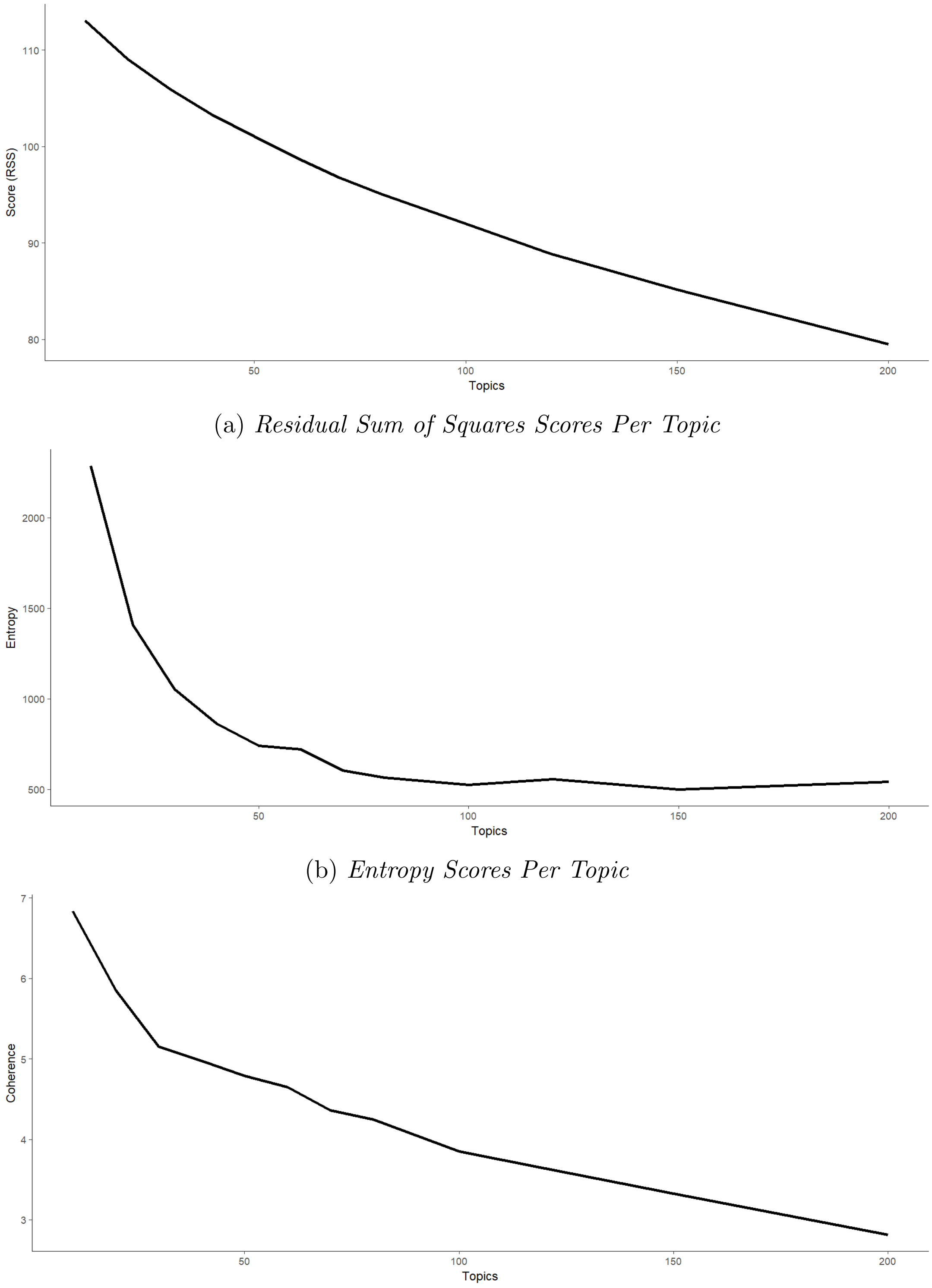
additional topic. In the case where there is a hidden number of groups, we may observe an improvement in the score once the number of topics reaches the number of these groups, with diminishing returns thereafter (Hutchins et al., 2008). Figure 1a shows the RSS scores to be decreasing in the number of topics, but at a marginal rate of decline. The slope of the curve becomes relatively flatter between 70 and 120 topics. For this reason, we suggest our optimal number lies within this range.

Secondly, we compute Entropy scores for our various numbers of topics. Entropy is a measure of unpredictability. Information theory shows that changes in entropy proxy as a measure of information gain. Following Stevens et al. (2012), for topic model M partitioning data into t groups, where t is the number of topics, entropy can be measured as:

$$H(M) = \sum_{i=1}^{t} -P(i)logP(i) \tag{2}$$

Entropy can therefore measure the amount of information gained from adding an additional topic. Figure 1b shows a negative association between the number of topics and information gain. A lower change in score suggests little information gain from one more topic. The figure shows that, for each additional topic, the extra information received is diminishing. Between 10 and 60 topics is when the greatest information gain occurs. This steadily falls between 50 and 100, getting flatter as the number of topics approaches 100. Information gain is relatively constant after 120 topics. Based on this measure, the optimal number of topics lies between 60 and 120, but closer to the upper bound.

Finally, we use Coherence-based scores. We can think of topics that make meaningful connections between words as being coherent. Measures of coherence are based on 'pairs of topic descriptor terms that co-occur frequently or are close to each other within a semantic space are likely to contribute to higher levels of coherence' (O'Callaghan et al., 2015: p. 1). Stevens et al. (2012) consider measures of topic coherence that align with judgements by human investigators. One such measure is the "UMass" measure of Mimno



(a) *Residual Sum of Squares Scores Per Topic*

(b) *Entropy Scores Per Topic*

(c) *Coherence Scores Per Topic*

Figure 1: *Measures for the Optimal Number of Topics*

Source: Author's calculations using *A Cradle of Inventions: British Patents from 1617 to 1894* (2009) et al. (2011). For topic T represented by the top n words t_i, the measure is defined as:

$$C(T) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} log \frac{D(t_i, t_j) + 1}{D(t_j)} \tag{3}$$

where $D(t_i)$ is the number of documents featuring word t_i, and $D(t_i, t_j)$ is the number of documents featuring both words t_i and t_j. For any given number of topics, we can then calculate the average topic coherence score.

Figure 1c displays the coherence scores. The overall trend suggests that additional topics lead to less coherent associations. The figure shows a sharp decline in coherence between 10 and 30 topics. The scores steadily fall until 100 topics, where the slope becomes flatter. There is also a small increase in Coherence between 70 and 80 topics. This measure suggests our optimal number lies between 70 and 100.

Based on the three metrics, we argue that the optimal number of topics for this instance is 100. Each score suggests that the range of 70-120 has the optimal amount. Collectively, the scores point to 100 as being the appropriate number. For the remainder of this discussion, we use 100 topics. It is important to note that this does not mean we will have 100 patent classes. The topic is a means to derive common word associations that we then classify according to a particular schema.

4 The Taxonomy

Our goal is to design a new, static patent taxonomy. This taxonomy should class patent data based on the text contained in patent titles. It should also classify patents according to their Destination, as described in Griliches (1990). This allows us to identify the relevant classification based on the information provided in patent titles. For example, the following patent title identifies the likely industry for this particular invention: *improvements in firearms*. This is a military improvement, but we cannot say for certain which industry would produce it, or which industry it came from.[8]

[8] The Production method is more subjective, as the investigator must decide the industry most likely to produce the new invention; it is doubtful the investigator has the requisite knowledge to do so. The Origin approach also requires the investigator to decide which industry an invention comes from; is this

Our methodology for developing a new taxonomy is two-fold. First, we undertake a number of counting exercises based on a sample of studies from the innovation literature. These exercises are, for the most part, subjective, but necessary, as human judgement is required to identify those existing classes that relate to each other. Second, we apply our machine learning techniques as an objective robustness check.[9]

Static classes are likely to be independent of both time and countries. Classes that appear frequently throughout the literature are likely to represent static classes, as their frequent usage indicates their value. Observing the entire population of established classes is not possible; not all studies publish their taxonomy. Instead, we use our sample of unique taxonomies presented in Table 1. This sample should be representative of the literature; it covers historical and contemporary taxonomies, as well as studies from different regions.

The first step to identifying our common classes is to decompose each taxonomy into a corpus of single words. This approach allows us to observe each word in isolation of its source. The majority of classes in our sample consist of a single word: we term them as short classes. Other classes are comprised of multiple keywords: we term these as long classes. However, determining whether a short class is similar to a long class is a difficult task. For example, consider the following long class taken from the historical German schema (Franks, 1915): 'Sheet Metal, Metal Pipe and Tube and Wire Manufacture and Working, Metal-Rolling'. This class focuses on the specific aspects related to metalworking and manufacturing. But, consider another, broader class, this time taken from Moser (2005): 'Manufactures'. The only relation between these classes comes from their use of the word 'manufacture'. But, the long class also relates to metalworking. Therefore, we could not consider these to be related. For long classes to be related they would have to consist of the same keywords, but this is highly unlikely. Instead, splitting the long class into a set of short classes means we can derive related

to be decided from the inventor's occupation or the industry of the firm the inventor works for, or some other criteria?

[9] It should be noted that no schema or methodology is capable of removing the subjective element of classification. However, our approach can significantly reduce this element, which then leads to a more consistent classification approach within the literature.

words much more easily.

Once each schema is decomposed, we initially tally how often each unique word appears. This result is displayed in the first half of Table 3. Our initial tally produced 1,600 unique words, the majority of which appeared only once or twice. This tally, however, is not very informative because it does not account for synonyms or related text.

The next step is to begin manually identifying related words and grouping them together. Here, human judgement has the advantage, as any machine learning algorithm would not necessarily identify, for example, that 'instruments' and 'accoutrements' are related terms. By grouping related words, we can derive a set of "word-groups" which act as preliminary classes. We start by observing an initial word – such as 'agriculture', for example – and then review the entire corpus for related terms – such as 'forestry', 'seeds', 'fishing', etc. Once our search of the corpus is complete, we then sum the tallies of related terms together to produce a score for each word-group. As a provisional check, both authors conducted this exercise independently, twice. Each exercise resulted in the same 24 word-groups. These are displayed in the second half of Table 3. The associated counts reflect how likely it is our word-group represents a static patent class. 'Lighting', for example, has the lowest count, as terms related to Lighting did not often appear as part of any class within our sample.[10]

However, certain taxonomies contain significantly more classes or words per class than others do, and may bias the results. The appearance of certain word-groups might result entirely from one taxonomy. Therefore, we repeat our counting exercise upon the broadest class levels within the sample. For example, the GICS has 'Sectors', which are divided into 'Industry Group', which are then further divided into 'Industries'; Sectors counts as the broadest level that we use in our second exercise instead of 'Industries. Consequently, we omit Woodcroft (1860) as it has too many classes and no broader level of classification. Our second counting method produced 536 unique words, compared to 1,600 from before.

[10] Higher scores suggest that a number of class contained related words. Some classes repeat the same words. However, they do not repeat such words too often. Therefore, lower scores possibly reflect this repetition, while high scores reflect the appearance of words in multiple class and schemas.

Table 3: *Industry Count Results*

Raw Count		Aggregate Count	
Words	**Count**	**Word-groups**	**Count**
Instruments	28	Commodities	263
Machinery	22	Machinery	249
Machines	22	Chemicals	244
Food	20	Instruments	213
Gas	20	Construction	194
Water	20	Textiles	170
Engines	19	Agriculture	169
Equipment	19	Transportation	116
Metal	19	Manufacture	98
Paper	17	Food	93
Agriculture	16	Apparel	89
Construction	16	Health	82
Electric	16	Paper	81
Mining	16	Engines	74
Manufacture	14	Metal	67
Materials	14	Electricity	60
Furniture	13	Gas	56
Printing	13	Water	54
Tools	13	Heating	52
Chemicals	12	Military	46
Appliances	11	Communications	43
Engineering	11	Mining	39
Glass	11	Utility	39
Leather	11	Lighting	23
Manufacturing	11		
Ships	11		
Textiles	11		
Carriages	10		
Chemicals	10		
Electrical	10		
Fabrics	10		
Lighting	10		
Steam	10		
Stone	10		
Transportation	10		

Notes: The 'Raw Count' columns represent our results from the initial frequency counts. 'Aggregate Count' displays the results for manually grouping certain words.

Source: See Table 1.

Of these words, we examined those with an initial tally of 10 or greater. The results are reported in Table 4. This allows us to identify whether new words have appeared compared to the initial count in Table 3. Of the 15 listed words, 'products', 'activities, and 'equipment' are new. Further investigation showed that 'products' and 'activities' were entirely from industrial schemas, ruling them out as common classes. Furthermore, we consider 'Equipment' to be related to 'Instruments' and group them accordingly. Words with a count of two or greater were then reviewed, with word-groups again being derived based on identifying related terms. The resulting word-groups remained identical to those from the prior exercise.

We next compare our word-groups against the sample literature. By doing so, we can check how often a word-group appears as a distinct class. Word-groups appearing

Table 4: *Industry Count Results Robustness Check*

Words	**Frequency**	**New**
products	24	Yes
activities	22	Yes
instruments	17	No
food	15	No
metal	15	No
agriculture	14	No
machinery	14	No
mining	14	No
paper	13	No
construction	12	No
chemicals	11	No
engines	10	No
equipment	10	Yes
machines	10	No
textiles	10	No

Notes: The table shows the words with an initial tally of 10 or greater. 'New' states whether the word appeared in the top 15 words of the first count shown in Table 3.

Source: Authors' calculations using data from Table 1.

frequently are then more likely to represent a static class. We check each word-group against each taxonomy, and then tabulate how often they appear verbatim. Table 5 shows the results of this matching process. For example, 24 out of the 36 sample taxonomies list Chemicals as a distinct class, while Utilities appears only five times. This suggests that Chemicals is representative of being a static class, while Utilities is less so.

To verify our derived word-groups, we next examine a selection of patent datasets with machine learning techniques. Specifically, we derive a set of topics from each dataset, and then match these topics to our list of word-groups. We base the strength of our proposed schema on whether it can suitably classify each topic. In particular, we are concerned with the spanning nature of the proposed classes: we wish to assign at least one word-group per topic, and are less concerned with instances where ambiguity arises. The ability to apply multiple classes mitigates concerns that might apply in the latter case. Large patent datasets will inevitably contain pioneering and niche inventions that are more difficult to classify. Such outlier patents are unlikely to undermine an entire classification schema. Nevertheless, if significant numbers of patents appear as distinct, unclassifiable

Table 5: *Results from Matching Word-groups to the Literature*

Word-groups	**Total**	**Percentage**
Chemicals	24	72.73
Machinery	21	63.64
Electricity	20	60.61
Food	19	57.58
Construction	18	54.55
Instruments	18	54.55
Textiles	18	54.55
Transportation	18	54.55
Agriculture	16	48.48
Health	16	48.48
Metal	15	45.45
Paper	14	42.42
Communications	13	39.39
Mining	13	39.39
Manufacturing	12	36.36
Apparel	10	30.30
Engines	8	24.24
Gas	7	21.21
Commodities	7	21.21
Military	7	21.21
Heating	6	18.18
Water	6	18.18
Lighting	5	15.15
Utilities	5	15.15

Notes: The table shows how often each word-group appears, verbatim, in our sample of 36 taxonomies from Table 1. Word-groups with higher scores are considered more robust and representative of the literature.

Source: Authors' calculations using data from Table 1.

topics then our schema is likely to be undermined.

To check the robustness of our word-groups, we apply the NMF topic analysis method to the patent datasets described in Table 6. We chose these datasets because they contain detailed patent titles, and span historical and contemporary periods collectively. For the USA and both UK datasets, we draw a random sample from each decade by extracting patents with an identification number ending in either one or six. By taking samples, we can ensure that each patent dataset is of a similar size, so that we can use the same number of topics.

To justify including an additional class, we would expect significant numbers of patents

Table 6: *Data Used for Topic Analysis*

Country	**Years**	**Source**
England	1617-1852	*A Cradle of Inventions (2009)*
Switzerland	1880-1930	PATSTAT Biblio
UK1	1893-1914	PATSTAT Biblio
USA	1950-1980	PATSTAT Biblio
UK2	1990-2016	PATSTAT Biblio

Notes: Datasets used to aid the development of our new patent taxonomy. For the USA and UK datasets, samples were taken from each decade.

Source: See Source column.

to arise that belong to a specific topic that we cannot map to an existing word-group. By extracting more topics from each dataset than word-groups within the proposed schema, we hope to expose any missing classes. Should such a distinct class exist, it follows that distinct language would be used to describe associated patents. If these patents appeared in significant numbers, we would expect a separate topic to appear. We can infer whether any omitted classes exist by reviewing the derived topics.

Our topic analysis confirms that the proposed schema is sufficient to capture patents from a number of diverse datasets.[11] We could readily assign each topic to at least one of our word-groups, supporting the word-groups as static classes. Similarly, the 'Switzerland' dataset was capable of classifying French patent titles in line with our proposed word-groups, providing further support that our methods can be used for any patent dataset, irrespective of the language. The results also suggest including an additional word-group, comprised of patents related to 'screws, nuts, bolts, nails, pins' etc. Consequently, we term this word-group as "Hardware" and append it to our set of word-groups.

To determine whether our proposed schema constitutes a static taxonomy, we compile the results from each of our exercises into Table 7. Here, each exercise is labelled as a "Step", and the respective tallies from each stage are shown. We also present a cut-off indicator to aid determine whether a word-group constitutes a patent class. This indicator provides an objective measure of how robust a word-group is in each step. For example,

[11] To prepare the patents for analysis, patent titles were stripped of non-printing characters and stop words. Suitable substitutions are applied to reduce all text to a standard character set.

Table 7: *Patent Class Methodology Scores*

Classes	Step One		Step Two		Step Three	Step Four					Frequency
	Raw	Aggregate	Raw	Aggregate		England	US	UK1	UK2	Swiss	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Chemicals	12	244	11	44	24	16	9	7	7	6	0
Construction	16	194	12	31	18	5	3	3	1	4	0
Electricity	16	60	8	24	20	1	10	2	2	7	0
Instruments	28	213	17	58	18	19	30	25	13	11	0
Machinery	22	249	14	66	21	28	8	6	8	5	0
Manufacturing	14	98	9	49	12	3	1	1	3	1	0
Transportation	10	116	8	35	18	7	6	3	7	5	0
Metal	19	67	15	31	15	4	1	1	4	**0**	1
Paper	17	81	13	22	14	3	1	2	3	**0**	1
Textiles	11	170	10	20	18	16	1	**0**	4	4	1
Agriculture	16	169	14	24	16	5	1	**0**	**0**	**0**	3
Communications	**7**	**43**	**7**	19	13	1	4	14	3	**0**	3
Engines	19	74	10	**16**	**8**	4	**0**	**0**	5	6	3
Food	20	93	15	34	19	**0**	**0**	1	1	**0**	3
Hardware	**4**	257	**2**	31	**4**	23	24	13	31	17	3
Apparel	**7**	89	**4**	19	10	1	**0**	**0**	1	1	4
Gas	20	**56**	8	**11**	**7**	3	1	2	**0**	1	4
Commodities	**8**	263	**7**	61	**7**	2	1	1	3	**0**	4
Water	20	**54**	8	**8**	**6**	3	2	3	**0**	1	4
Health	**8**	82	**6**	33	16	**0**	1	3	**0**	**0**	5
Heating	**8**	**52**	**4**	**12**	**6**	3	2	1	1	4	5
Lighting	10	**23**	**5**	**5**	**5**	2	**0**	1	1	1	5
Mining	16	**39**	14	**17**	13	2	1	**0**	**0**	**0**	5
Utilities	**6**	**39**	**3**	**3**	**5**	2	**0**	**0**	4	5	7
Military	**4**	**46**	**2**	**12**	**7**	**0**	**0**	**0**	**0**	**0**	10
Cut-off point	<10	Bottom Third	Bottom Third	Bottom Third	<10	>0	>0	>0	>0	>0	

Notes: 'Step One' refers to our first count of unique words. 'Step Two' refers to our second count of unique words, using the broadest level of classification available from our sample. 'Step Three' refers to counting the number of taxonomies in which each of our 24 word-groups appear. 'Step Four' shows how often each of our classes appeared in one of the 5 listed patent datasets using topic analysis. The "cut-off" points are defined as the threshold for determining whether a class is robust. A class which falls inside the cut-off point criteria is considered less robust. For each column they are as follows: Column 1 - classes with scores less than 10. Columns 2, 3, and 4 - the bottom third of classes. Column 5 - classes with scores less than 10. Columns 6, 7, 8, 9, and 10 - classes with a score of 0. In the final column, classes with higher scores are reviewed. Step Four is given a higher rating, as it is based on a greater number of empirical observations. The result is that the following classes are reviewed, then either modified, combined further, or kept: Apparel into Textiles; Engines becomes Power; Gas into Chemicals and Utilities; Water, Heating, Lighting into Utilities; Military into Instruments. Mining and Health are kept as separate classes because they do not readily fit into another existing class.

under 'Step One: Raw', the cut-off value states that any word-group with a score below 10 is less likely to be a static class. For most exercises, the cut-off value is intended to separate the bottom third of scores from the rest. An inspection of each column shows a greater separation between word-groups in the bottom third against the remainder. For Step Four, a word-group which can be readily assigned to a topic is considered a stronger indicator of whether it is a static class.[12] In each column, the text that appears in bold falls below our assigned cut-off values.

The final column of Table 7 is our measure to determine whether a word-group qualifies as a static class. A score of zero indicates a completely robust class. A score of ten, however, indicates that a word-group is not static. For any score between zero and ten, we review the associated word-group, with higher scores more likely to be removed or reformed.

Based on our review, we merge the following classes: Apparel into Textiles; Gas into Chemicals or Utility; Heating, Lighting and Water into Utility; and Military into Instruments. To avoid confusion, we reform Engines into Power (as Engines and Machines are very similar classes), which groups inventions that produced locomotion, energy, or force of any kind. Hardware, which initially arose from examining a selection of patent datasets, is also determined to be representative of a static class. We repeat each step to ascertain whether Hardware should be included in the taxonomy. This review led us to conclude that Hardware represents another usable class; Hardware appeared frequently under the topic analysis approach, which we consider a stronger indicator of robustness. Overall, our methodology produced a set of 19 static patent classes.

The final step was to use topic analysis techniques to aid the descriptions of our classes. Inadequate descriptions can lead to a subjective interpretation of how to apply our taxonomy. Such difficulties would deter adoption of the schema, and undermine results derived from its application. Since topics represent word clusters that tend to appear in combination with one another, where a topic directly relates to a class the

[12] Note that most topics under this Step were assigned at least two classes. We count both within the tally.

Table 8: *Patent Class Definitions*

Number	**Classification**	**Inventions Pertaining To:**
(1)	Agriculture	The growth of crops and raising of livestock; fishing, forestry and hunting; horticulture; unspecified use of land
(2)	Chemicals	The development of new chemicals, the applications of chemicals, or products developed by chemicals processes; organic and inorganic chemistry; gases; nuclear
(3)	Commodities	Consumable, durable, and non-durable goods which are not explicitly for industrial usage, with a focus on inventions to be sold in the market for private use; intangible services; recreational items
(4)	Communications	Facilitating communication between persons; signalling; digital inventions; software; media
(5)	Construction	Building; tools for building; civil engineering; construction and building related accessories; building of infrastructure; construction of items of a physical nature
(6)	Electricity	The creation, management, and application of electricity; of electrical appliances, components, and instruments; aspects of electricity which do not overlap with other utilities; combinations of electricity with galvanism, magnetism and the like
(7)	Food	The production, treatment, and management of foodstuffs and beverages for consumption; tobacco
(8)	Hardware	Devices, objects, or items which serve a purpose without requiring a direct application; Objects which do not require a direct action in order to function
(9)	Health	Improving the quality of life; life-saving medicines or apparatus; protection from ailments
(10)	Instruments	Measuring, gauging, weighing; general devices or objects which reduce the effort required to perform certain tasks; devices or objects which aid in productivity of labour; a tool or implement especially for precision work
(11)	Machinery	Machines which operate on mechanical power, and to their maintenance; processes conducted by machines
(12)	Manufacturing	The production of goods or items; large scale and small scale
(13)	Metal	Metallurgy; extracting metals from their ores; the application of chemical processes to metals, whether by producing, refining, galvanising or other such methods
(14)	Mining	The construction of mines, their excavation, management, flood management, and extraction of natural resources; the raising and lowering of heavy bodies
(15)	Paper	The use of paper; methods which improve paper; the process of printing; paper and cardboard production, and to other such related items; physical record keeping; bills, cheques
(16)	Power	Generating, regulating, and applying energy for power, speed, or such related uses
(17)	Textiles	The creation of fabrics from processes of weaving, spinning, knitting, felting, etc, and their bleaching or dyeing, and treatment ; clothing and clothing accessories
(18)	Transportation	Facilitating speedy, or easier, travel across distances; transport infrastructure; packaging and storage of items for easier transport
(19)	Utility	The management of public systems, such as sewerage; the creation, management, and application of gas, heat, light, and water; the regulation of water, light, heat, gas, and electricity as public goods; and to inventions which encompass combinations of water, light, heat, gas, and electricity; fireproofing structures

Notes: Definitions are constructed using the list of word associations derived from the topic analysis approach. Some classes could be further divided using these definitions, or further aggregated.

words comprising that topic act as descriptors. This uncovers the vocabulary used to connect a patent's description with its intended classification, upon which we build our class definitions. Table 8 presents our finalized classification schema.

5 Application of Taxonomy

The use of competing patent taxonomies within the literature is problematic: the results posited may be contingent on the choice of classification. How can we compare the results derived using different taxonomies, especially when we do not know how to replicate them? Before we test for classification bias, this section first discusses our dataset: the population of British patents granted during the period 1700-1850. We chose this dataset for the following reasons. First, this dataset has been used extensively within the historical innovation literature (Dutton, 1984; Sullivan, 1989; Sullivan, 1990; MacLeod, 2002; Nuvolari and Tartari, 2011; Meisenzahl and Mokyr, 2012; Bottomley, 2014a; Dowey, 2017; Khan, 2018). Second, prior studies have classified the patent data. For instance, Nuvolari and Tartari (2011) and *A Cradle of Inventions: British Patents from 1617 to 1894* have both assigned competing schemas to the data. Both schemas are obtainable for this present study, allowing for a simple comparison with our own. Third, this dataset covers the traditional period of the Industrial Revolution (1760-1830). Any insights from this era are vital to our understanding of this phenomenon, which transformed stagnant, agricultural economies into industrialised ones, birthing the modern age.

To prepare the dataset for comparison, we assign the three taxonomies. The Cradle of Invention (COI) schema had already been assigned when the data were extracted. The Nuvolari-Tartari (NT) taxonomy was provided by Nuvolari and Tartari (2011). We assign our schema using our machine learning methodology. After deriving our 100 topics, we assigned one class per topic. Where a topic was inconsistent in its word associations, we labelled it 'Unclear' and then manually reviewed any patents assigned to it.[13] Our method creates each topic and assigns patents to them simultaneously. We then assign the topic's

[13] Such occurrences, however, are relatively few: only four topics were labelled Unclear.

Table 9: *Comparison of Class Assignments*

Cradle of Invention				**Nuvolari-Tartari**				**TopicOne**				**TopicTwo**			
Class	**Count**	**Percentage**	**HHI**	**Class**	**Count**	**Percentage**	**HHI**	**Class**	**Count**	**Percentage**	**HHI**	**Class**	**Count**	**Percentage**	**HHI**
Agriculture	510	3.34	0.001	Agriculture	479	3.48	0.001	Agriculture	501	3.06	0.001	Agriculture	410	2.50	0.001
Beverages	310	2.25	0.001	Carriages	888	6.45	0.004	Chemicals	1,243	7.59	0.006	Chemicals	1,203	7.34	0.005
Clothing	302	2.19	0.000	Chemicals	1,236	8.97	0.008	Commodities	357	2.18	0.000	Commodities	178	1.09	0.000
Communications	102	0.74	0.000	Clothing	366	2.66	0.001	Communications	33	0.20	0.000	Communications	47	0.29	0.000
Domestic	1,747	12.68	0.016	Construction	692	5.02	0.003	Construction	1,167	7.12	0.005	Construction	1,549	9.46	0.009
Food	369	2.68	0.001	Engines	1,818	13.19	0.017	Electricity	141	0.86	0.000	Electricity	123	0.75	0.000
Industry	6,513	47.27	0.223	Food	784	5.69	0.003	Food	144	0.88	0.000	Food	73	0.45	0.000
Instruments	500	3.63	0.001	Furniture	716	5.20	0.003	Hardware	1,417	8.65	0.007	Hardware	1,296	7.91	0.006
Medicine	259	1.88	0.000	Glass	146	1.06	0.000	Health	428	2.61	0.001	Health	279	1.70	0.000
Military	235	1.71	0.000	Hardware	920	6.68	0.004	Instruments	1,013	6.18	0.004	Instruments	1,038	6.34	0.004
Mining	280	2.03	0.000	Instruments	651	4.72	0.002	Machinery	1,111	6.78	0.005	Machinery	1,370	8.36	0.007
Miscellaneous	18	0.13	0.000	Leather	237	1.72	0.000	Manufacture	1,431	8.74	0.008	Manufacture	1,830	11.17	0.012
Paper	580	4.21	0.002	Manufacturing	769	5.58	0.003	Metal	573	3.50	0.001	Metal	757	4.62	0.002
Textiles	1,865	13.54	0.018	Medicines	301	2.18	0.000	Mining	381	2.33	0.001	Mining	279	1.70	0.000
Transportation	1,667	12.10	0.015	Metallurgy	763	5.54	0.003	Paper	298	1.82	0.000	Paper	317	7.90	0.006
				Military	267	1.94	0.000	Power	973	5.94	0.004	Power	1,294	10.97	0.012
				Mining	94	0.68	0.000	Textiles	2,250	13.73	0.019	Textiles	1,797	7.00	0.005
				Paper	526	3.82	0.001	Transportation	1,658	10.12	0.010	Transportation	1,147	8.52	0.007
				Pottery	314	2.28	0.001	Utility	1,263	7.71	0.006	Utility	1,395	8.52	0.007
				Ships	648	4.70	0.002								
				Textiles	1,949	14.15	0.020								
HHI			**0.280**				**0.078**				**0.077**				**0.085**

Notes: The table displays the Herfindahl-Hirschman Concentration ratios for each taxonomy. Count represents the total number of patents related to that class. This is then represented as a percentage. The individual class HHI scores are represented. The bottom row displays the HHI ratio for each taxonomy as a whole.

Sources: Authors' calculations using data from *A Cradle of Inventions: British Patents from 1617 to 1894* and ServicesNuvolari and Tartari (2011). Both datasets cover 1700-1850.

associated class. By assigning the top two topic scores to each patent, we can account for any potential overlap across technology groups. We denote these as TopicOne and TopicTwo. In some instances, a patent has the same class assigned twice. We consider these patents to have no overlapping characteristics. We also manually classified the entire dataset, and compared our assignments with the machines. Both authors do this independently. In 90 per cent of cases, either of our manually assigned classes matched either of the assigned topics. The remaining 10 per cent either were the result of Unclear topics, or patents which had too few unique words.

Table 9 presents a comparison of the schemas used in this study. Several classes appear within each taxonomy: Agriculture, Food, Instruments, Medicines (or Health), Mining, Paper, and Textiles. For these commonly occurring classes, however, the number of assigned patents are not identical across taxonomies. The COI schema, for example, assigns 510 patents to Agriculture, while our own TopicTwo assigns only 410. At least 100 patents are prone to being classified inconsistently. Food patents suffer a similar inconsistency across existing schemas. COI lists 369 patents as Food, while NT lists 784 instead. The majority of patents also receive a different TopicTwo assignment, suggesting that the characteristics of many patented inventions spillover into multiple technology groups. This supports our assertion that patents require more than one classification.

We calculate Herfindahl-Hirschman (HHI) scores for each schema. HHI scores show how concentrated a particular taxonomy is. A higher score indicates a more skewed distribution of patents within a particular schema. For example, COI has the highest associated HHI score at 0.280, while TopicOne has the lowest at 0.077. Examining the COI schema shows that 'Industry' accounts for almost 50 per cent of all British patents. No other schema has such a 'catch-all' class.

6 Comparison of Taxonomies

The existing, competing schemas do not consistently classify patent data. Consequently, studies that use different schemas are likely to produce different results. To test for any potential bias, we observe each schema against two commonly examined patent characteristics: the citations of patented inventions, and the occupational status of patentee's. Because each taxonomy does not have the exact same patent classes, we present only those common to all schemas: Agriculture, Food, Instruments, Medicines (or Health), Mining, Paper, and Textiles.

6.1 The Citations of Patented Inventions

First, we examine how the chosen taxonomy affects an analysis of the citations of patented inventions. In the innovation literature, patent citations are a popular metric used to proxy for patent quality or value (Hall et al., 2001; Hall et al., 2005; Lach and Schankerman, 2008; Bernstein, 2015; Kogan et al., 2017). In place of citations, the historical literature has adopted the Woodcroft Reference Index (WRI), as pioneered by Nuvolari and Tartari (2011). This index lists how many contemporary scientific and trade journals referenced a particular patent within our dataset. The references are used to proxy for the technical and economic significance of a particular patented invention: more references signals a higher quality patent. Because the number of references artificially increases over time, we adopt the approach of Hall et al. (2005) and Nuvolari and Tartari (2011), by weighting the total sum of references on a patent by the average number of references on all patents within a given time period. To ensure comparability, our time periods are those of Nuvolari and Tartari (2011).[14]

The quality indicator is a count variable with a skewed distribution; many patents have few references, and few patents have many references. The negative binomial model accounts for this skewness by relaxing the assumption that the mean and the variance are equal (Greene, 2008).[15] Under this model, our dependent variable is the weighted number of references on a given patent. Our control variables constitute: whether the patentee had a prior patent; the patentee's occupation; whether the patentee's occupation matches the class of their invention; their nationality; and time controls. The explanatory

[14] These cohorts are as follows: 1700-1721; 1722-1741; 1742-1761; 1762-1781; 1782-1801; 1802-1811; 1812-1821; 1822-1831; 1832-1841; 1842-1850.

[15] We also test this relationship using the poisson model. The results from poisson are equivalent to the negative binomial approach.

Table 10: *Negative Binomial: Dependent Variable is the Weighted Number of References per Patent*

VARIABLES	(1) NT	(2) COI	(3) TopicOne	(4) TopicTwo	(5) TT
Food	0.018	0.022	0.017	-0.057	-0.070*
	(0.029)	(0.038)	(0.056)	(0.070)	(0.042)
Instrument	0.035	-0.022	-0.026	-0.037	-0.073***
	(0.030)	(0.032)	(0.029)	(0.040)	(0.017)
Medicines	-0.077**	-0.037	0.017	-0.039	-0.048**
	(0.036)	(0.041)	(0.034)	(0.050)	(0.023)
Mining	0.174***	0.188***	0.001	-0.042	-0.065***
	(0.060)	(0.044)	(0.039)	(0.047)	(0.024)
Paper	0.070*	0.053	0.022	-0.031	-0.046**
	(0.036)	(0.035)	(0.041)	(0.045)	(0.024)
Textiles	-0.067**	-0.037	-0.017	-0.065	-0.086***
	(0.028)	(0.028)	(0.028)	(0.040)	(0.017)
Constant	-0.131***	-0.111***	-0.134***	-0.074	-0.049
	(0.042)	(0.040)	(0.042)	(0.048)	(0.038)
Time	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y
Observations	13,286	13,286	13,286	13,286	13,286
Pseudo R-Squared	0.00376	0.00294	0.00317	0.00264	0.00303

Notes: The table shows how the quality of patented inventions varies by technology group. The dependent variable is the weighted number of references per patent. In each column, the omitted variable is the "Agriculture" class. Coefficients are interpreted as the difference in the logs of expected counts of the predictor variable. To translate this into a unit change, the coefficients need to be exponentiated. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Sources: Authors' calculations using data from *A Cradle of Inventions: British Patents from 1617 to 1894* and Nuvolari and Tartari (2011). Both datasets cover 1700-1850.

variables are the classes associated with each schema. We represent patent classes with dummy variables, where Agriculture is the chosen baseline category.

Table 10 provides the results of our approach. Column 1 uses the NT schema; column 2 then controls for the COI schema; column 3 examines our TopicOne taxonomy; column 4 represents the TopicTwo taxonomy; and column 5 controls for TopicOne and TopicTwo (henceforth known as "TT"). We argue that future investigators who employ our schema run three separate econometric specifications, using TopicOne, TopicTwo, and then both schemas together as a robustness check.

Classification bias exists. This bias affects all aspects related to interpreting regression coefficients. The magnitude of coefficients fluctuates considerably when comparing Mining inventions, for example. The COI schema suggests that Mining patents are likely to have 17-18 per cent more references per patent compared to Agricultural patents. One reasonable interpretation is that capital-intensive inventions are of a greater quality.[16] TopicOne, however, suggests that Mining patents have 0.1 per cent more references. Capital-intensive inventions, then, are of a similar quality to Agricultural ones.

Statistical significance also fluctuates considerably. Textile patents, for example, show a statistically significant association under the NT and TT schemas. However, this significance does not exist under the remaining schemas. Here, Textile patents are not statistically distinguishable from Agricultural patents, in terms of their respective number of references. Such a result is likely to lead investigators to consider Textile patents as being no different from Agricultural patents.

The direction of association of coefficients is also subject to bias. Most classes exhibit some variation in the direction of association; Textiles is the only class to show a consistently negative result. Furthermore, Food and Paper patents show the greatest variation, as both classes have an almost even split between positive and negative signs. For these classes, TopicTwo and TT schemas produce a negative result, suggesting they have fewer references than Agricultural patents. However, the remaining schemas produce a positive result, suggesting instead that these types of patents have more references.

6.2 Patentee Occupational Status

To ascertain whether classification bias is unique to examining the citations of patented inventions, we conduct an additional test by regressing patentee's occupations against patent classes. The innovation literature has examined the role of independent inventors

[16] Based on their titles, Mining patents were likely to be highly mechanised during this period. Such inventions are considered to be capital-intensive, as suggested by Khan (2005), because more capital than labour is required for their development.

Table 11: *Probit: Dependent Variable is a Dummy representing a Non-Manual Occupation*

VARIABLES	(1) NT	(2) COI	(3) TopicOne	(4) TopicTwo	(5) TT
Food	0.110***	0.167***	0.117**	0.021	0.009
	(0.025)	(0.035)	(0.046)	(0.066)	(0.035)
Instruments	0.035	-0.025	-0.018	-0.026	-0.061***
	(0.025)	(0.028)	(0.025)	(0.028)	(0.013)
Medicines	0.287***	0.292***	0.223***	0.070*	0.122***
	(0.037)	(0.040)	(0.033)	(0.038)	(0.021)
Mining	0.228***	0.236***	0.107***	0.048	0.036*
	(0.059)	(0.041)	(0.033)	(0.039)	(0.021)
Paper	0.131***	0.083***	-0.029	0.042	-0.027
	(0.028)	(0.028)	(0.034)	(0.037)	(0.021)
Textiles	-0.024	0.000	-0.024	-0.051*	-0.082***
	(0.022)	(0.022)	(0.023)	(0.027)	(0.012)
Time	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y
Observations	13,241	13,241	13,241	13,241	13,241
Pseudo R-Squared	0.0875	0.0690	0.0753	0.0585	0.0743

Notes: The table shows how the association between non-manual occupations and technology groups. The dependent variable is a dummy variable, where a value of 1 indicates a non-manual occupation. In each column, the omitted variable is the "Agriculture" class. Coefficients are interpreted as marginal effects at the means. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Sources: Authors' calculations using data from *A Cradle of Inventions: British Patents from 1617 to 1894* and Nuvolari and Tartari (2011). Both datasets cover the period 1700-1850.

and the types of industries they are likely to select into, or the types of inventions they are likely to produce (Schmookler, 1966; Khan and Sokoloff, 2004; Nicholas, 2010; Nicholas, 2011b; Khan, 2018). Our data allow us to conduct a similar examination. The patent data record the patentee's occupation alongside their name. This allows us to match occupations to a statistical measure of potential skills: the HISCLASS schema of Van Leeuwen and Maas (2011). The metric group occupations based on their skills, whether they are manual or non-manual labour, and the degree of supervision required. For simplicity, we break the HISCLASS codes into manual versus non-manual, following Klemp and Weisdorf (2012). Non-manual occupations are likely to be higher-skilled than their manual counterparts (Van Leeuwen and Maas, 2011).

We represent non-manual occupations using a dummy indicator variable. Consequently, a probit regression model is necessary to derive the probability of patent classes being associated with non-manual occupations. Our control variables constitute: whether the inventor had a prior patent; their nationality; and time controls. The explanatory variables are patent classes, with the baseline class being Agriculture. Table 11 reports our results.

Classification bias still exists, and all aspects related to interpreting coefficients are affected. Medicine patents show a significant range in terms of coefficient size. Under the COI schema, for example, an average Medicine patent is approximately 29 per cent more likely to be associated with a non-manual occupation, when compared to an Agricultural patent. The size of this result is large, and suggests that inventors of Medicine patents were skilled. However, the TopicTwo schema suggests that non-manual occupations were only seven per cent more likely to produce Medicine patents. While the conclusion remains similar, the reduced coefficient size suggests that the specific human capital and skills associated with elite occupations are less important for producing medicinal inventions.

Statistical significance also varies across taxonomies. The majority of patent classes present an almost even divide between significance and non-significance. For example, Food patents are statistically significant at the one per cent level under the NT, COI, and TopicOne schemas. From this, a reasonable interpretation is that Food patents are significantly different to Agricultural patents. However, the TopicTwo, and TT schemas are not statistically significant at conventional levels. This result undermines our initial interpretation.

The direction of association, likewise, fluctuates considerably. For Paper patents, there is an almost equal divide between positive and negative associations. The NT, COI, and TopicTwo schemas suggest Paper patents were more likely to be associated with non-manual occupations compared to Agricultural patents. TopicOne and TT, however, suggest the opposite: less skilled individuals were more likely to produce Paper patents.

7 Discussion

To our knowledge, the present study is the first to show that classification bias exists in the field of innovation studies. Prior studies have not explicitly addressed the potential consequences of their choice of taxonomy. This is a serious concern, as statistical significance, direction of influence, and coefficient magnitude are subject to bias when competing schemas are used. The extent of this bias is uncertain within the literature. Without a complete understanding of how authors construct their taxonomies, and for what purpose, we cannot determine how serious the bias is. Most academic studies do not provide such detail. Therefore, prior research articles that do not expressly describe their taxonomy should be interpreted with caution.

To illustrate the severity of our findings, consider the following example. Suppose there exists a policymaker tasked with designing appropriate measures for encouraging innovation. This policymaker bases her decisions upon the existing evidence presented to her. She is keen to promote innovation by directing resources toward particular high-value technologies. But, she must first discern which technology groups are associated with higher value inventions, and the types of skills associated with them. In her approach, the policymaker hires a number of academic investigators, one of which decides that long-run evidence on the subject would be useful. They classify patents using the COI taxonomy. In their analysis, they find that capital-intensive inventions – Mining, for example – were on average more valuable, and were more likely to be produced by higher skilled occupations. This result is consistent with the technologies that drove the Industrial Revolution: Mining was an important industry, and is arguably responsible for the advances in steam engine technology (Nuvolari, 2004; Allen, 2009; Mokyr, 2009).

Our policymaker may conclude that supporting capital-intensive innovation is the appropriate policy. She may then shape industrial policy to support, for example, the establishment of university research parks (URP).[17] Such parks would likely encourage

[17] URPs are intended to encourage innovation through knowledge diffusion from academic research to small, high-tech start-ups (Anselin et al., 1997; Siegel et al., 2003; Link and Scott, 2007).

capital-intensive innovation through access to highly skilled labour, scientific knowledge, and additional resources for R&D. Suppose, however, another investigator had the same idea to obtain historical evidence, except they use the TopicTwo schema. Based on their findings, the policymaker would likely conclude that capital-intensive inventions are not of great value nor produced by highly skilled individuals. She is likely, then, to question her initial policy measures. Had she implemented a policy to encourage URPs, then she may have misdirected important resources, with little effect on innovation.

Of course, this is but a very simplistic example. However, it does highlight the potential implications classification bias has for prescribing policy. Existing taxonomies are difficult to replicate, and may lead to the development of new taxonomies, which further compounds the inconsistency problem. The collective body of evidence on the economics of patents is then difficult to interpret. This could also be potentially problematic for the development of British industrial policy.

The UK Government recently published a White Paper on industrial strategy outlining its approach to increasing long-run economic growth and innovation in Britain (HM Government, 2017). Policymakers intend to encourage growth in the UK economy through "sector deals" between the government and key industries; the industries are chosen based on empirical evidence concerning the relative strengths and weaknesses of particular sectors (HM Government, 2017: p. 209). Initial deals have been arranged with Life Sciences, Construction, Artificial Intelligence, and the Automotive industries. The deals include directing financial resources toward these sectors, and establishing important links with research institutions (such as URPs).

However, the success of this initiative in encouraging innovation strongly depends upon how the sectors are classified, and identified. For example, the Construction sector could be grouped by its production process (supply-side) or by its competitors (demand-side). The former approach is likely to cluster potentially unrelated firms that produce the necessary materials for construction (such as steel, cement, and vehicles), while the latter would group related firms based on their activities, which may exclude useful supply-chains. In either case, potentially innovative firms or industries may not receive the benefits of this investment. If the UK Government intends to encourage productivity-enhancing innovation in particular sectors, then they need to understand who is capable of producing such valuable innovations. Useful construction innovations may be produced by individual inventors, unrelated firms, or unaffiliated research institutions. Therefore, by directing their efforts based on the existing evidence, policymakers are at risk of making sub-optimal decisions. Such decisions may have great economic cost, such as the opportunity cost of the resources used. Incorrect policy measures may even run the risk of inadvertently hindering rather than encouraging innovation.

Our recommendations for the literature are as follows. Firstly, creators of taxonomies should describe how they design them. This ensures that potential biases are identifiable and their methods replicable. Secondly, mitigating potential biases requires adopting a universal schema. The taxonomy produced in this study is a useful starting point, as it is readily adoptable and adaptable for future studies. Thirdly, descriptions need to accompany patent classes to ensure a consistent classification of patent data throughout the literature. Fourthly, subjectivity can be reduced by employing machine learning techniques to improve the consistency of patent classification. Finally, topic analysis provides a means to both identify appropriate classes and omitted classes, and to perform the classification of patent datasets in ways that are useful for economic analysis of innovation.

8 Conclusion

Our goal in this paper has been: to document methods of taxonomy construction; to design and develop a new, static patent taxonomy in a clear and transparent manner; to develop a new method for classifying all patent data consistently; and to show that classification bias exists. We recommend our methodology and taxonomy be used in future studies. We acknowledge, however, that our schema may not be applicable to every study. In such cases, future investigators should describe any new taxonomies they produce. The machine learning techniques described in this study are adaptable and adoptable for any future researchers, and could be used alongside other schemas, and could also be used for studies outside of the patent literature. The techniques presented here are capable of classifying any textual data, which may then increase the comparability in other research disciplines.

The implications of classification bias are likely to be profound for the patents and innovation literature. Classification bias exists, at least, in the long-run British patent data studied here. Whether this bias exists in other datasets necessitates a re-examination of the existing literature, for clarification. In the case that this bias is only moderate, interpreting the literature is then less problematic, and deriving appropriate policy measures would remain possible. However, in the extreme case, where all studies are biased, the literature becomes incomparable. If studies are not comparable, then appropriate policy measures cannot be readily prepared. We recommend, where possible, that existing studies be re-examined using our schema and methodology. This is not to say that our schema is "right", as there can be no objective measure of this. Our schema is transparent, however, making it straightforward for any subsequent studies to make use of it, or draw from it, as they see fit. Our methodology, likewise, is also not "right", but it is, at least, consistent. Human error is substantially minimised using our machine learning approach. Related patents will always be identified, and will always be grouped together. Compared to humans, the machine is much less likely to make mistakes.

9 References

A Cradle of Inventions: British Patents from 1617 to 1894 (2009). Stevenage, UK: Metal Finishing Information Services Ltd.

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2002). Competition and Innovation: An Inverted U Relationship. *National Bureau of Economic Research Working Paper Series* No. 9269.

Allen, R. C. (2009). *The British Industrial Revolution in Global Perspective.* Cambridge: Cambridge University Press.

Anselin, L., A. Varga, and Z. Acs (1997). Local Geographic Spillovers between University Research and High Technology Innovations. *Journal of Urban Economics* 42(3), pp. 422–448.

Bailey, M. F. (1946). History of Classification of Patents. *Journal of the Patent Office Society* 28(7), pp. 463–507.

Bain, J. S. (1951). Relation of Profit Rate to Industry Concentration: American Manufacturing, 1936-1940. *The Quarterly Journal of Economics* 65(3), pp. 293–324.

Bain, J. S. (1954). Economies of Scale, Concentration, and the Condition of Entry in Twenty Manufacturing Industries. *The American Economic Review* 44(1), pp. 15–39.

Baten, J., A. Spadavecchia, J. Streb, and S. Yin (2007). What made southwest German firms innovative around 1900? Assessing the importance of intra- and inter-industry externalities. *Oxford Economic Papers* 59(suppl 1), pp. i105–i126.

Bernstein, S. (2015). Does Going Public Affect Innovation? *The Journal of Finance* 70(4), pp. 1365–1403.

Bottomley, S. (2014a). Patenting in England, Scotland and Ireland during the Industrial Revolution, 1700-1852. *Explorations in Economic History* 54, pp. 48–63.

Bottomley, S. (2014b). *The British Patent System During the Industrial Revolution 1700-1852: From Privilege to Property.* Cambridge: Cambridge University Press.

Brunt, L., J. Lerner, and T. Nicholas (2012). Inducement Prizes and Innovation Inducement Prizes and Innovation. *Journal of Industrial Economics* 60(4), pp. 657–696.

Burhop, C. and N. Wolf (2013). The German Market for Patents during the "Second Industrialization", 1884-1913: A Gravity Approach. *Business History Review* 87(1), pp. 69–93.

Dowey, J. (2017). Mind over matter: access to knowledge and the British Industrial Revolution. PhD thesis. The London School of Economics and Political Science.

Dutton, H. I. (1984). *The patent system and inventive activity during the industrial revolution, 1750-1852*. Manchester University Press.

ECPC (1992). *Issue Paper No. 1: Conceptual Issues*. Tech. rep. U.S. Department of Commerce.

ECPC (1993). *Issue Paper No. 6: Services Classifications*. Tech. rep. U.S. Department of Commerce.

ECPC (1994). *Report No. 1: Economic Concepts Incorporated in the Standard Industrial Classification Industries of the United States*. Tech. rep. U.S. Department of Commerce.

EPO (2017). *Guidelines for Examination in the European Patent Office*. Tech. rep., p. 851.

Franks, W. T. (1915). *Key to the Classifications of the Patent Specifications of France, Germany, Austria, Netherlands, Norway, Denmark, Sweden, and Switzerland, in the Library of the Patent Office*. 3rd. London: Darling & Son Ltd.

Galasso, A. and M. Schankerman (2015). Patents and Cumulative Innovation: Causal Evidence from the Courts. *The Quarterly Journal of Economics* 130(1), pp. 317–369.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters* 99(3), pp. 585–590.

Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature* 28(4), pp. 1661–1707.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. *National Bureau of Economic Research Working Paper Series* No. 8498.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2005). Market Value and Patent Citations. *The RAND Journal of Economics* 36(1), pp. 16–38.

HM Government (2017). *Industrial Strategy: building a Britain fit for the future*. Tech. rep. Department for Business, Energy and Industrial Strategy, p. 256.

Hutchins, L. N., S. M. Murphy, P. Singh, and J. H. Graber (2008). Position-dependent motif characterization using non-negative matrix factorization. eng. *Bioinformatics (Oxford, England)* 24(23), pp. 2684–2690.

Johnson, D. K. N. (2002). The OECD Technology Concordance (OTC): Patents by Industry of Manufacture and Sector of Use. *OECD Science, Technology and Industry Working Papers.*

Khan, B. Z. (2005). *The Democratization of Invention: Patents and Copyrights in American Economic Development, 1790-1920.* Cambridge University Press.

Khan, B. Z. (2013a). Going for Gold. Industrial Fairs and Innovation in the Nineteenth-Century United States. *Revue économique* 64(1), pp. 89–113.

Khan, B. Z. (2013b). Selling Ideas: An International Perspective on Patenting and Markets for Technological Innovations, 1790-1930. *Business History Review* 87(1), pp. 39–68.

Khan, B. Z. (2015). Inventing prizes: a historical perspective on innovation awards and technology policy. *Business History Review* 89(4), pp. 631–660.

Khan, B. Z. (2016). Prestige and Profit: The Royal Society of Arts and Incentives for Innovation and Enterprise, 1750-1850. *LSE Economic History Working Papers* No. 248/20.

Khan, B. Z. (2017). Prestige and Profit: The Royal Society of Arts and Incentives for Innovation, 1750-1850. *National Bureau of Economic Research Working Paper Series* No. 23042.

Khan, B. Z. (2018). Human capital, knowledge and economic development: evidence from the British Industrial Revolution, 1750–1930. *Cliometrica* 12(2), pp. 313–341.

Khan, B. Z. and K. L. Sokoloff (2004). Institutions and Democratic Invention in 19th-Century America: Evidence from 'Great Inventors,' 1790-1930. *National Bureau of Economic Research Working Paper Series* No. 10966.

Klemp, M. and J. Weisdorf (2012). The lasting damage to mortality of early-life adversity: evidence from the English famine of the late 1720s. *European Review of Economic History* 16(3), pp. 233–246.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017). Technological Innovation, Resource Allocation, and Growth. *The Quarterly Journal of Economics* 132(2), pp. 665–712.

Kortum, S. and J. Putnam (1997). Assigning Patents to Industries: Tests of the Yale Technology Concordance. *Economic Systems Research* 9(2), pp. 161–176.

Krueger, A. B. and L. H. Summers (1988). Efficiency Wages and the Inter-Industry Wage Structure. *Econometrica* 56(2), pp. 259–293.

Lach, S. and M. Schankerman (2008). Incentives and invention in universities. *The RAND Journal of Economics* 39(2), pp. 403–433.

Lampe, R. and P. Moser (2016). Patent Pools, Competition, and Innovation-Evidence from 20 US Industries under the New Deal. *Journal of Law, Economics & Organization* 32(1), pp. 1–36.

Lehmann-Hasemeyer, S. and J. Streb (2016). The Berlin Stock Exchange in Imperial Germany: A Market for New Technology? *American Economic Review* 106(11), pp. 3558–3576.

Link, A. N. and J. T. Scott (2007). The economics of university research parks. *Oxford Review of Economic Policy* 23(4), pp. 661–674.

Lybbert, T. J. and N. J. Zolas (2014). Getting patents and economic data to speak to each other: An Algorithmic Links with Probabilities' approach for joint analyses of patenting and economic activity. *Research Policy* 43(3), pp. 530–542.

MacLeod, C. (2002). *Inventing the industrial revolution: The English patent system, 1660-1800.* Cambridge University Press.

Meisenzahl, R. and J. Mokyr (2012). The Rate and Direction of Invention in the British Industrial Revolution: Incentives and Institutions. *The Rate and Direction of Inventive Activity.* Ed. by J. Lerner and S. Stern. Chicago: University of Chicago Press, pp. 443–479.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum (2011). Optimizing semantic coherence in topic models. *Proceedings of the conference on empirical methods in natural language processing.* Association for Computational Linguistics, pp. 262–272.

Mokyr, J. (2009). *The Enlightened Economy: An Economic History of Britain 1700-1850.* Yale University Press.

Moser, P. (2005). How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs. *The American Economic Review* 95(4), pp. 1214–1236.

Moser, P. (2012). Innovation without Patents: Evidence from World's Fairs. *The Journal of Law & Economics* 55(1), pp. 43–74.

Nanda, R. and T. Nicholas (2014). Did bank distress stifle innovation during the Great Depression? *Journal of Financial Economics* 114(2), pp. 273–292.

Nicholas, T. (2008). Does Innovation Cause Stock Market Runups? Evidence from the Great Crash. *The American Economic Review* 98(4), pp. 1370–1396.

Nicholas, T. (2010). The Role of Independent Invention in U.S. Technological Development, 1880-1930. *The Journal of Economic History* 70(1), pp. 57–82.

Nicholas, T. (2011a). Did R&D Firms Used to Patent? Evidence from the First Innovation Surveys. *Journal of Economic History* 71(4), pp. 1032–1059.

Nicholas, T. (2011b). Independent invention during the rise of the corporate economy in Britain and Japan. *Economic History Review* 64(3), pp. 995–1023.

Nicholas, T. (2011c). The origins of Japanese technological modernization. *Explorations in Economic History* 48(2), pp. 272–291.

Nuvolari, A. (2004). Collective invention during the British Industrial Revolution: the case of the Cornish pumping engine. *Cambridge Journal of Economics* 28(3), pp. 347–363.

Nuvolari, A. and V. Tartari (2011). Bennet Woodcroft and the value of English patents, 1617-1841. *Explorations in Economic History* 48(1), pp. 97–115.

O'Callaghan, D., D. Greene, J. Carthy, and P. Cunningham (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications* 42(13), pp. 5645–5657.

Pearce, E. (1957). *History of the Standard Industrial Classification.* Tech. rep. Washington: Executive Office of the President. Office of Statistical Standards.

Phillips, A. (1966). Patents, Potential Competition, and Technical Progress. *The American Economic Review* 1(2), pp. 301–310.

Rajan, R. G. and L. Zingales (1998). Financial Dependence and Growth. *The American Economic Review* 88(3), pp. 559–586.

Schautschick, P. (2015). An Economic Investigation of the Use and Impact of Patents and Trade Marks in Germany. PhD Thesis. Ludwig Maximilians University.

Schmoch, U., F. Laville, P. Patel, and R. Frietsch (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research.*

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

Scotchmer, S. (1991). Standing on the Shoulders of Giants: Cumulative Research and the Patent Law. *The Journal of Economic Perspectives* 5(1), pp. 29–41.

Scotchmer, S. (2004). *Innovation and Incentives.* London: MIT Press.

Siegel, D. S., P. Westhead, and M. Wright (2003). Assessing the impact of university science parks on research productivity: exploratory firm-level evidence from the United Kingdom. *International Journal of Industrial Organization* 21(9), pp. 1357–1369.

Sokoloff, K. L. (1988). Inventive Activity in Early Industrial America: Evidence From Patent Records, 1790-1846. *The Journal of Economic History* 48(4), pp. 813–850.

S&P Capital IQ and MSCI (2015). Global Industry Classification Standard.

Statistics Division (2008). *International Standard Industrial Classification of All Economic Activities. Revision 4.* Tech. rep. Department of Economic and Social Affairs.

Stevens, K., P. Kegelmeyer, D. Andrzejewski, and D. Buttler (2012). Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* EMNLP-CoNLL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 952–961.

Sullivan, R. J. (1989). England's 'Age of Invention': The Acceleration of Patents and Patentable Invention During the Industrial Revolution. *Explorations in Economic History* 26(4), pp. 424–452.

Sullivan, R. J. (1990). The Revolution of Ideas: Widespread Patenting and Invention During the English Industrial Revolution. *The Journal of Economic History* 50(2), p. 349.

Van Leeuwen, M. H. and I. Maas (2011). *HISCLASS: A Historical International Social Class Scheme*. Leuven: Leuven University Press.

Verspagen, B., T. Van Moergastel, and M. Slabbers (1994). MERIT concordance table: IPC-ISIC (rev. 2). *MERIT Research Memorandum February*.

Walsh, J. P., Y.-N. Lee, and T. Jung (2016). Win, lose or draw? The fate of patented inventions. *Research Policy* 45(7), pp. 1362–1373.

WIPO (1992). The International Patent Classification (IPC). eng. *Journal of the Patent and Trademark Office Society* 74(7), pp. 481–483.

WIPO (2016). Guide to the International Patent Classification. Geneva.

Woodcroft, B. (1860). *Subject-Matter Index of Patents of Invention, From March 2, 1617 (14 James I.) to October 1, 1852 (16 Victoria)*. London: Queen's Printing Office.