Cover Page – MSc Business Analytics Consultancy Project/Dissertation 2020-21

Title of Project: An Unsupervised Learning Approach to Predicting Emerging Climate Change Mitigation Technologies (*Using Hierarchical Agglomerative Clustering and Patent Analysis to Forecast Emerging Technologies within the field of Renewable Energy Generation)*

Date: 15th August 2021

Word Count: 11410

Disclaimer:

I hereby declare that this dissertation is my individual work and to the best of my knowledge and confidence, it has not already been accepted in substance for the award of any other degree and is not concurrently submitted in candidature for any degree. It is the end product of my own independent study except where other acknowledgement has been stated in the text.

Marking Sheet – MSc Business Analytics Consultancy Project/Dissertation 2020-21

Criteria/Weight	Supervisor's comments
Topic, theoretical framework, literature, and methodology (35%): Topic is clearly identified and boundaries are asserted. Knowledge of relevant theories and their limitations. Current and relevant literature coming from reliable sources. Appropriate and adequate methodology for topics. Detailed methodology facilitating replication of project and reproducibility of results.	
Analysis and conclusions /recommendations (35%): Use of primary and/or secondary data. Rigorous analysis and interpretations. Alternative interpretations/arguments are considered. Limitations are identified and justified by reasonable arguments. Conclusions/recommendations are fully consistent with evidence presented.	
Structure, originality and presentation (10%): Provides a concise summary. Demonstrates an understanding of business context. Coherent and appropriate structure. Adequate presentation, language, style, graphs, tables, and referencing. Appropriate use of visualisation. Presents business recommendations.	
Complexity of project scope and progress made towards business goals (10%): Progress made towards overcoming technical and operational challenges encountered during the project. Progress made in overcoming problem framing and theoretical and data related problems encountered during the project.	
Project Management (10%): Good use of project management and communication tools. Use of Kanban board for structuring project work. Evidence of objectives being broken down in appropriate tasks and timely engagement with primary supervisor.	

General marking guidelines

85+ Outstanding work of publishable standard.

70-84 Excellent work showing mastery of the subject matter and excellent analytical skills.

60-69 Very good work. Interesting analysis with original insights. Some minor errors.

50-59 Good work which only covers a basic analysis. Some problems but no major omissions.

40-49 Inadequate work. Not sufficiently analytical. Some major omissions.

39- Work seriously flawed. Lack of clarity and argumentation. Too descriptive.

Mark: _____

An Unsupervised Learning Approach to Predicting Emerging Climate Change Mitigation Technologies

Using Hierarchical Agglomerative Clustering and Patent Analysis to Forecast Emerging Technologies within the field of Renewable Energy Generation

Innovation and Intellectual Property Management Laboratory,

University of Cambridge

Faculty of Engineering,

School of Management,

University College London

15th August 2021

Abstract

This research project aimed to predict the emergence of new technologies with the field of Climate Change Mitigation Technologies (CCMTs) related to the reduction in greenhouse gas (GHG) emissions in energy generation using patent data. We collected patent data from the USPTO, from 1980 – 2020, and utilised the predicting power of patent citations. We created a patent citation vector and employing hierarchical clustering algorithms to reflect the changing role a patented technology has over time within its technological field allowed us to identify the Y02E 10/549 CPC subclass of patents. We identified patented inventions related to organic photovoltaic (PV) cells prior to the technology's market fruition from the clustering approaches employed within our methodology. The results of this study can be a proof of concept that forecasting emerging technologies within CCMTs using patent data is possible and a tool to improve innovation efficiency.

Table of Contents

1. Introduction

The Kyoto protocol, signed in December 1997 and adopted in 2005, was a landmark moment in our efforts to tackle climate change. In short, the protocol marked the first agreement between nation states to mandate reductions in greenhouse gas emissions. As a result, developed nations that committed to cutting their emissions did so by an average of 5.2% by 2012.

However, 16 years on from the adoption the Kyoto protocol, the demand for innovation in the area of Climate Change Mitigation is encompassing economic, social and political agendas within the largest economies of the world. Worldwide, net emissions of greenhouse gases (GHGs) from human activities have increased the warming effect of climate change by 45% from 1990 to 2019 (Blair, 2021) and energy production today accounts for 72 percent of all emissions (UNFCC,2021). For example, with Africa's population is set to double by 2050, research and development within CCMTs (Climate Change Mitigation Technologies) needs to be conducted with greater efficiency to enable sustainable growth and supply energy to these rapidly increasing populations (Blair, 2021).

In this project, we take inspiration for improving the efficiency of innovation within CCMTs by building on the technological forecasting work of You et al.,(2017), Kim and Bae (2017) and Kyebambe et al., (2017) ; combining patent data and advanced machine learning clustering algorithms to identify technological clusters and predict emerging technologies. We utilise freely available and abundant patent data and apply it as a key indicator for innovation output, in conjunction with the computation of patent citation vectors to identify the technological field that characterises a patented invention. By tracking the evolution of technological clusters we can thus, predict the emergence of technological innovations. Ultimately, by exploring future technological landscapes, we can better direct the early stages of research and development.

The project is structured into five chapters. This first chapter provides a background into the field of research; USPTO and CPC classifications, citation networks and the need for R&D efficiency to stimulate innovation. We will also define our motivations and our research objectives. The second chapter will introduce the value of using patents as an innovation

metric and the previous predictive frameworks used to forecast emerging technologies within a literature review. The third chapter formulates the methodology applied to our patent data and the roadmap undertaken to induce technological clusters from our citation vector. The results of these are presented within the fourth chapter and the final chapter will synthesise these results, identify the limitations of our methodology and thus provide foundations for future work.

1.1 Innovation output of Research and Development

Despite the process of innovation being arguably as old as mankind itself, the modern-day obsession with innovation being a disruptive process is somewhat misleading. It is a gradual and iterative process, or as Ridley (2020) states; ‘ a process of constantly discovering ways of rearranging the world into forms that are unlikely to arise by chance – and that happen to be useful’.

Innovation has been key to economic growth and sustainability (Saviotti et al,. 2003) and is the fusion of state research expenditure and private actors; large corporations, research institutions or private individuals. Innovation requires huge investment in R&D. As of 2019 , worldwide R&D expenditures reached $1.7 trillion, and yet because innovation is unpredictable (Érdi et al., 2012) this investment is risky. Although R&D expenditure has increase in Europe to record highs, innovation has declined compared to in Asia and North America. Furthermore, national R&D budgets can be sensitive to political environments which greatly increases the cost of R&D failure (Kim and Bae, 2017). For governments and private entities alike, understanding and identifying the emergence of new technology and new technological fields can help direct public policy, investment and ultimately reduce risk and improve the efficiency of the investment (Érdi et al., 2012).

The recent digitisation of patent data and access to greater computing power have enabled researchers and to approach this innovation conundrum using predictive analytics and patent networks to increase the efficiency of R&D expenditure by forecasting emerging technologies.

1.2 Patents and the Classification of CCMTs

A patent is granted by patent offices to the inventor of a particular invention. It places the right to stop others from making, using or selling the invention without the permission of the inventor, for a limited time period. For the purposes of this project, we contextualised our use of patent data in the realm of CCMTs by looking at the classification framework adopted by the United States Patent and Trademark Office (USPTO) and the European Patent Office (EPO).

As climate technologies diffused to the forefront of technological innovation, the EPO and the USPTO developed a modified version of the IPC classification system dedicated for CCMTs. This is known as the Cooperative Patent Classification (CPC) system. CPC is divided into nine different sections ranging from Section A – H (Table 1.1), which are subsequently divided into subsections, classes, groups and subgroups. This hierarchical structure of the CPC can be demonstrated for the Y02E class. It is one of eight classes in the Y02 scheme (Table 1.2) and encapsulates 6 groups within Y02E. The subgroups within Y02E 10/00 can be visualised and indicate the classification for CCMTs within the field of energy; specifically technologies in the reduction of GHG emissions.

CPC Section	CPC Description
A	Human Necessities
B	Performing Operations; Transporting
C	Chemistry; Metallurgy
D	Textiles; Paper
E	Fixed Constructions
F	Mechanical Engineering; Lighting; Heating; Weapons; Blasting
G	Physics
H	Electricity
Y	General tagging of new technological developments; general tagging of cross-sectional technologies spanning over several sections of the IPC; technical subjects covered by former USPC cross-reference art collections [XRACs] and digests

Table 1.1 Sections by their technological CPC classification (A-Y)

Of the patents tagged within the Y section, the Y02 subsection exists for patents for which the technologies they represent control, reduce or prevent the emissions GHGs and those technologies that work to reverse the effects of climate change. There exist eight classes of Y02: Y02A, Y02B, Y02C, Y02D, Y02E, Y02P, Y02T and Y02W. The classes cover buildings, GHG capture, energy, industry and agriculture, transportation and waste management. For this project will we focus on patents tagged within the subclass Y02E, energy, in order to focus on emerging technologies in the realm of clean energy and reducing GHG emissions.

CPC Class	**CPC Description**
Y02A	Technologies for Adaption to Climate Change
Y02B	Climate Change Mitigation Technologies related to buildings, eg. housing, house appliances or related end-user applications
Y02C	Capture, storage, sequestration or disposal of greenhouse gases(GHG)
Y02D	Climate Change Mitigation Technologies in information and communication technologies [ICT], i.e. information and communications technologies aiming at the reduction of their own energy use
Y02E	Reduction of Greenhouse Gas Emissions, related to energy generation, transmission or distribution
Y02P	Climate Change Mitigation Technologies in the production or processing of goods
Y02T	Climate Change Mitigation Technologies related to transportation
Y02W	Climate Change Mitigation Technologies related to wastewater treatment or waste management

Table 1.2. Classes of Y02 by their CPC classification

This hierarchical structure of the CPC can be demonstrated for the Y02E class. It is one of eight classes in the Y02 scheme (Table 1.2) and encapsulates 6 groups within Y02E (Table 1.3). Looking deeper into the classification structure Y02E 10/00, for example, contains 26 sub- groups and the patents' technologies contained within this are classified as CCMTs within the field of energy; specifically technologies that relate to energy generation through renewable energy sources.

CPC Group	CPC Description
Y02E 10/00	Energy generation through renewable energy sources
Y02E 20/00	Combustion technologies with mitigation potential
Y02E 30/00	Energy generation of nuclear origin
Y02E 40/00	Technologies for an efficient electrical power generation, transmission or distribution
Y02E 50/00	Technologies for the production of fuel of non-fossil origin
Y02E 60/00	Enabling technologies; Technologies with a potential or indirect contribution to GHG emissions mitigation
Y02E 70/00	Other energy conversion or management systems reducing GHG emissions

Table 1.3. CPC groups of Y02E by their CPC classification

CCMT are unique as they do not align to any one single classification section. They belong to various technological domains of research due increased interoperability of innovation; new emerging technologies bridge software from one domain to hardware of another domain (Angelucci et al,. 2018). As a result the 'Y' classification section was introduced to encapsulate new technological developments that span over several sections of the IPC. This resulted in patent documents relating to CCMTs being identified and re-classified under the 'Y' section, in addition to their classification tags from non-Y sections. Prior to this, patents relating to CCMTs were scattered throughout the CPC and IPC and these patents fell under multiple different technological classification sections. (Angelucci et al,. 2018)

Using this background in the CPC scheme, we can induce that the patents classification systems can disseminate knowledge about technologies, but also allows us to structure and obtain information on the origins of patents published for CCMTs. In the next chapter, the literature review will outline how these patents, their classifications and their citations are valuable metrics for innovation and how vectors can be created from patent networks to reflect the formation of a new technological emergence.

2. Literature Review

The first section of this literature review summarises the theoretical background underpinning the use of patents citation analysis in academic research and its function as a metric for technological evolution and innovation. The second section will review the published literature with regards to the forecasting of technologies using patent data in recent history. We will examine the range of methodologies executed within this academic realm.

2.1 Patents as an innovation metric

Patents have existed for over a two centuries with patents being granted since the late 18th century and current patent databases dating back to the 1870's (Hall, Jaffe and Trajtenberg, 2001). Patents have become essential in describing something novel and not obvious as they provide an objective measure of new knowledge (Katila, 2000). For these reasons early studies hypothesised patents as a metric for innovation, a process that is notoriously difficult to quantify (Katila, 2000)(Dutta and Weiss, 1997)

For example, the United States patent system, contains 18,694,681 scholarly works. Although the US patent system, and other patent systems alike, can't provide a complete record of every technological progression and innovation, it remains a well-studied and valuable source of data (Érdi et al., 2012).

There is precedent for patents being used as metrics for innovative output (Katila, 2000)(Dutta and Weiss, 1997), however patent networks have curated much attention particularly in the last decade. There are numerous reasons for this. Firstly, the recent availability of patent data via electronic access has enabled a new generation of research and the use of patent-based measures of innovation (Katila, 2000). Patent data is open source, easy to access and most importantly; up to date (Kim and Lee, 2015). This new ubiquitous nature of patent data means that access to the EPO and USPTO has increased the use of patent data and in both academic and industrial research (Katila, 2000, Walker, 1995). Secondly, patents are full of valuable information about a technology as they have well-grounded descriptions of technology, additional information about future innovations; inventor names, time and date, possible applications within certain industries and a novelty

comparison to prior solutions (Kim and Lee, 2015). Furthermore patent data contains citations directed to patents published previously and to scientific literature that occurred before a patent is published (Hall, Jaffe and Trajtenberg, 2001). These citations open a 'window' into the process of knowledge flow (Jaffe and Trajtenberg, 1998) by using these citations to trace multiple linkages between new inventions, inventors, companies and researchers with those that came before them or are working in a different geographical region or area of expertise (Hall, Jaffe and Trajtenberg, 2001). This enables ' new' knowledge to be comprised of combinations of 'old' knowledge.

However, there are notable limitations of patent data. The most trivial being that not all inventions are patented. Firstly, many inventions don't satisfy the patentability criteria set up by patent offices such as the USPTO (Hall, Jaffe and Trajtenberg, 2001). The invention must be novel, non-trivial and have a commercial application (Hall, Jaffe and Trajtenberg, 2001) and whether an invention can be patented or not is subject to patent examiners and as a result, not all patents have the same value (Kim and Lee, 2015). Secondly, the inventor can face situations where patenting is strategically disadvantageous to their commercial aims. The propensity of patenting is specific to a particular company, institution or industry and changes in legislature also account as a factor (Fleming and Sorenson, 2001). Smaller enterprises may be restricted by budget and not publish patents accordingly: trade secrets mechanisms can be used as appose to patent publication or secrecy could be more appropriate (Kim and Lee, 2015). Furthermore there exist industries where patenting of product innovation is not the most efficient way to achieve a competitive advantage within the field (Kim and Lee, 2015). Finally, inventors are likely to limit their patent applications so that only their inventions with the highest chance of success will be filed for patent approval (Fleming and Sorenson, 2001).

2.2 Patent Citation Analysis

Patent citation analysis can be defined as the examination of citation links among patents (von Wartburg, Teichert and Rost, 2005). Citation counts have long been used to evaluate research performance (Érdi et al., 2012) and the long term value of an invention, in multiple studies, has been used as a foundation of technological value (Carpenter, Narin and Woolf,

1981)(Albert, Avery, Narin and McAllister, 1991) with positive correlations being found between citation count and the performance of a company or institution (von Wartburg, Teichert and Rost, 2005). Patent citations are either 'backward' or 'forward' measures of citations (von Wartburg, Teichert and Rost, 2005). Backward citations are those that appear within a patent and correspond to the 'technological antecedents of a particular patent (Depoorter, Menell and Schwartz, 2019) ie. the number of citations made by a patent. If a patent contains many backward citations that are technologically diverse, the patent is assumed to be derived from diverse set of previous inventions. Conversely, 'forward' citations are citations that are received by a patent from subsequent patents (Depoorter, Menell and Schwartz, 2019) ie. citations that a patent receives from other patents, after it has been originally published. Forward citations build up over time and therefore induce a citation lag; where the number of forward citations in the first few years of a patent being published do not appear until after this lag period. Forward citations for a highly cited patent within a selected technology category have been investigated to see if they follow the S-curve distribution as is seen during the life cycle of a new emerging technology (Fallah, Fishman and Reilly, 2011). It was concluded that forward citations followed a linear growth pattern and recommended being weighted in relation to the backward citations of a patent (Fallah, Fishman and Reilly, 2011).

Over the past 20 years, researchers have taken advantage of improvements in computing power and proposed multiple hypotheses for how best patent citations can be used. Lukach and Lukach(2007) introduced the PageRank score of patents which highlighted the symbiotic relationship between computing power and patent citation analysis (Depoorter, Menell and Schwartz, 2019). This inspired Google's "Random surfer" model, whereby patents were ranked based on the weighting of their forward and backwards citations (2)(Brin and page, 1998). Further studies concluded that forward citations were positively correlated with the rate of improvement for a technology over the subsequent ten-year period. In turn, the average age of backward citations was negatively correlated (Depoorter, Menell and Schwartz, 2019). Hall, Jaffe and Trajtenberg (2005), early on, confirmed the relationship between citation intensity and the value of patents by relating citation-weighted patents with the market value of the companies filing the patents(Depoorter, Menell and Schwartz, 2019) (Hall, Jaffe and Trajtenberg 2005).

Recently, two main uses of patent citations can be induced. The first is to use patent citations to investigate knowledge flows and spill overs, the second is measure the relatedness of different patent networks(Érdi et al., 2012) and to map innovation systems. The most notable studies around patent citations as a medium to investigate international knowledge flows were conducted by Jaffe and Trajtenberg (Jaffe and Trajtenberg, 1999). They demonstrated that patent citations are localised geographically and presented a clear path in which, for a patent published in the US, an inventor in the US has a higher probability of citation in the early years of the invention being made than elsewhere (Jaffe and Trajtenberg, 1999)(Jaffe and Trajtenberg 1998). A year later they went further as to prove that patents filed from the same company are more likely to cite each other, patents in the same CPC class are 100 times more likely to cite each other as patents from different classes and patents whose inventors live in the same country are 30-80% more likely to site domestically as opposed to internationally (Jaffe and Trajtenberg, 1999).

Most relevant to our work within this project is the use of patent citations to structure knowledge from different technological classes using the patent citation network. This involves a measure of similarity between patents (Érdi et al., 2012). An important approach domain of analysis includes co-citation analysis which defines the frequency with which two patents are cited together (Érdi et al., 2012). It has been used to provide insights into the development of technology in biotech, semiconductors and nanotechnology (Érdi et al., 2012) and assumes that if patents are frequently cited together, they would in turn cover closely related technologies (Érdi et al., 2012). This has recently been taken up as a methodological approach by Wallace et al. 2009 to use co-citation networks to detect clusters from a citation-weighted network. The case has been made that that co-citation should be used to develop patent classification systems to assist understanding the landscape, and evolution of a technological class of patents (Érdi et al., 2012). In the case of nanotechnology, Huang et al 2004 (Li, Chen, Huang and Roco, 2007) created various patent citation networks and then applied core network, critical node and topological analysis on these citation networks to identify the influential players and subfields within nanotechnology (Li, Chen, Huang and Roco, 2007). Cho et al 2021(6) recently used patent data to uncover the crucial companies in the field of autonomous vehicles using cross-citation analysis and main path analysis (Cho, Liu and Ho, 2021). This revealed the

knowledge flow through the industry while also highlighting the development trajectory of technologies in production (Cho, Liu and Ho, 2021).

Ultimately, these previous works using patent citation analysis enable us to understand how patent citation exploration is not just an indicator of patent value, but can also be used to structure a technology's developmental path. This will form an essential part of our methodology and inspire the use of a citation vector as a predictor.

2.3 Predictive tools for forecasting emerging technologies and innovation

In recent times there has been a large pool of literature covering the field of technology forecasting from patent data from which our methodology could take inspiration from. A range of methods have been employed and we are able to detect the evolving predictive methodologies being employed; with earlier work encompassing the use of growth curves and forward citations, and as research progressed we see the use of text mining, supervised learning, deep learning and unsupervised learning techniques.

For example, an early study conducted by Fallah, Fishman and Reilly, 2009 employed growth curves to measure the rate of technological innovation across three technological classes using forward citations (Fallah, Fishman and Reilly, 2009) . Their theoretical base was inspired by Everett Roger's Diffusion of Innovation and Technology Life Cycle, and thus hypothesised that patent forward citations for emerging technologies would follow a classic sigmoid or S- curve distribution(Fallah, Fishman and Reilly, 2009). Their methodology involved selecting groups of patents issued over 20 years ago, so that there were sufficient forward citations to extract and fit the cumulative forward citations to a variety of growth models; with the linear model fitting best (Fallah, Fishman and Reilly, 2009).

In recent times, supervised learning methods have been employed to generate forecasting results by introducing foreign knowledge into the model, in the form of labels (Y.Zhou 2020). In particular work by Kyebambe et al,. (2017) used patent data from the USPTO, a novel labelling algorithm and predictors from patent citations to predict the emergence of new technologies a year before they emerge (Kyebambe et al., 2017). The methodology relied on backward citations as oppose to forward citations to avoid the citation lag and

applied bibliographic coupling based on those backward citations (Kyebambe et al., 2017). This measured the extent to which two patents cite the same patents; a similar concept to co-citation analysis which we have seen is fundamental to forming patent citation networks (Kyebambe et al., 2017). This enabled patents to be grouped into technological clusters and using conventional k-means algorithms and the novel labelling algorithm (Kyebambe et al., 2017). However, limitations to this method arise as the number of clusters needed to be specified when the clustering was applied and for this external expertise of the patent landscape is required (Kyebambe et al., 2017). Furthermore new technologies will fail to be labelled appropriately the longer the period of prediction due to the lack of training samples available.

Y Zhou, 2020 attempted to combine data augmentation and deep learning methods to overcoming the lack of training samples and applied deep learning methods to forecast emerging technology (Zhou et al,. 2020). A sample patent data set was constructed and multiple patent features were created; incorporating both forward and backward citations, technology cycle time and IPC code. However, a generative adversarial network (GAN) was also used to generate synthetic samples to increase the scale of the data set and a deep neural network classifier was trained to forecast emerging technology based on the augmented data (Zhou et al,. 2020). The results supported previous claims that forward citations had a positive correlation with backward citations (Zhou et al,. 2020). However, this method still required the addition of external industry knowledge within the deep learning classifier meaning the method cannot be applied without expertise of a technological field.

The most significant work in forecasting emerging technologies has been conducted using unsupervised clustering; specifically clustering based on co-citation and citation networks. The work of Erdi. 2012, which will form the basis of our methodology, searched for emerging and evolving technology clusters based on a citation network over various time periods(Érdi et al., 2012). Erdi identified the emergence of a new CPC class of technology within the field of agriculture, food and textile years before the class had been identified by the USPTO classification system, by applying hierarchical clustering to a feature engineered ‘citation vector’. This served as a measure of similarity between patents, to which the hierarchical agglomerative clustering algorithm was applied, to capture the evolution of

technological fields over various time periods (Érdi et al., 2012) based on potential cluster evolution dynamics outlined by palla et al,. (2007). The citation vector was the sum of forward citations received by patents in the 36 pre-defined technological subcategories from USPTO classifications and was weighted by the overall number of backward citations made by the sender patent. Repeated over multiple time periods, temporal changes within the cluster evolution were identified from structural changes in dendrograms and validated using the qualitative method of back testing (Érdi et al., 2012). The methodology was executed on USPTO's and NBER's classification of patents, but it is thought that this could be applied to any technological space within any classification system. Due to the CCMTs being classed in 2010, with this methodology, we would not be able to identify the formation of new classes over time due to the all the patents undergoing a cross tagging process where patents were re classified as CCMTs in retrospection . Thus no new class was introduced or identified, however the methodology could still predict the emergence of new technology, albeit without identification of a new class.

Further unsupervised learning methodology techniques were presented in 2017 by Kim and Bae who used patent data from the USPTO to forecast technology in the wellness and care industry within the US (Kim and Bae, 2017). Kim and Bae constructed patent matrices using Pearson's correlation coefficient between documents, applied K-means algorithm and defined the clusters produced based on the top 10 CPC codes (Kim and Bae, 2017). In this methodology, forward citations were not used to build the vector for which the clustering algorithms were applied to, but to validate the clusters and identify promising technology (Kim and Bae, 2017).

Additional methods using text mining have been used to complement clustering approaches. Kim et al 2019 used text mining to induce patents with similar topic semantics to form clusters of patents within the field of wireless power transfer (Kim et al., 2019). The topic semantics were extracted and a matrix of word frequencies was created and weighted with TFIDF prior to clustering, with the results of the clustering being validated by time series and growth cycle analysis (Kim et al., 2019). Kim et al,. in 2016 (Kim, Lee, Jang and Park, 2016)) conducted a similar text mining approach when conducting patent analysis on Korean car manufacturers but encounter the same limitations of text mining; its subjectivity (Kim, Lee, Jang and Park, 2016). Both methodologies required the results of clustering

needed to be validated by qualitative judgement from experts in the field over the interpretation of the technology clusters using the extracted key words (Kim et al., 2019). Ideally the methodology we seek to follow will be able to forecast emerging technology, without the need of expert industry knowledge to validate the clusters formed.

To conclude, we can induce various findings and trends from the literature that proceeds this project. The first is that there is a clear demand and capability for R+D decision making to be made objectively on data; unburdened by emotion, external pressures or political influence. Although many of these studies have limitations, in terms of scale and validation requirements, it is clear that machine learning techniques should continue to direct technology forecasting research. We can also identify the trade of between supervised and unsupervised learning techniques; the former requires the input of labels and the latter requires qualitative validation. Both techniques, require external expert knowledge of the industry within which the forecasts are being made. Finally, we see that these forecasting methodologies are being used in a wide variety of technological industries, which supports our methodology of using clustering to forecast emerging technologies within the realm of CCMTs.

3. Methodology

3.1 Research Framework

The patent analysis in this project is divided into four parts and will largely follow the framework laid out by Erdi et al,.(2012). In the first part, we will conduct bibliographic extraction of features needed to create our 'predictor' or citation vector. Second, we will summarise the steps required to construct the citation vector which will be our predictor of technological development. In the third part, we discuss the application and theoretical significance of our agglomerative clustering that will be applied to our predictor. The final part will involve detecting structural changes in the dendrograms produced from our hierarchical clustering methods, the identification of clusters that require further clustering and an overview of the qualitative validation process we will be using; back testing.

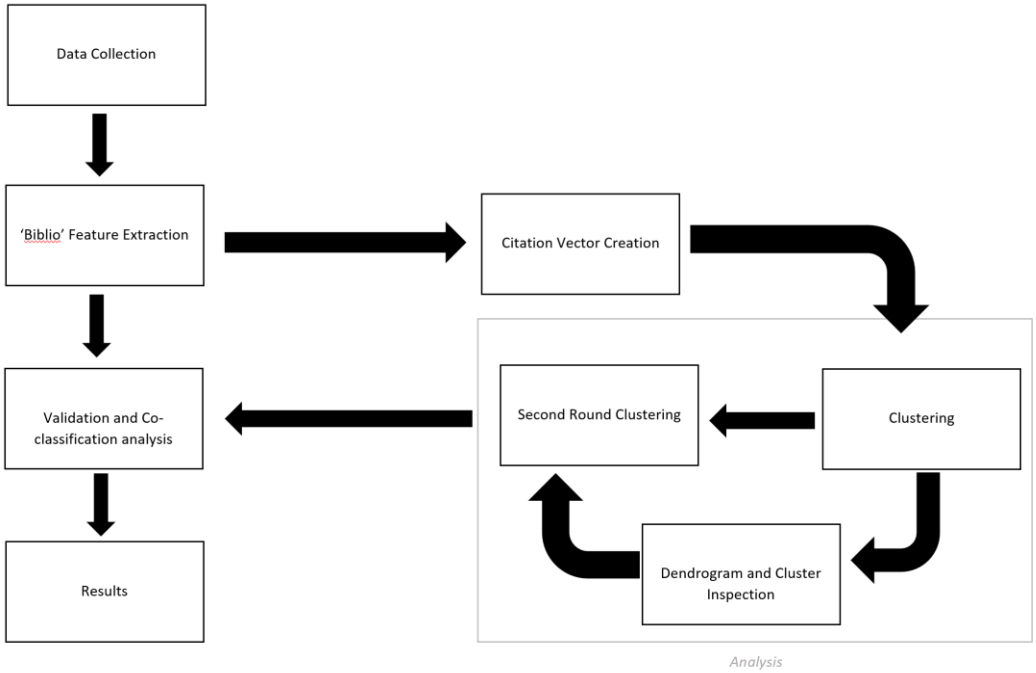


Figure 3.1. Research Framework for each time period

3.2 Hypothesis Formulation

The overall orientation of our research methodology is similar to that of Erdi et al,.(2012) as we are searching for emerging and evolving technology clusters based on a patent citation network. This methodology should provide a framework to use the predicting power of a patent citation vector visualised through temporal changes in cluster evolution to identify emerging technologies before they come to market fruition and without the need for expert knowledge within the domain of interest. While Erdi et al,.(2012) limited his research to the fields of agriculture, textiles and food, we will conduct ours within the field of CCMTs as previously justified. However regardless of the industry, identifying changes within the structure of the dendrograms that are produced from our hierarchical clustering approach, should be similar practise.

Our hypothesis is consistent with the work conducted in Érdi et al., (2012) as we use the citation vector as the predictor, across multiple time periods, because changes in this citation vector will reflect the evolving influence that a patented technology has on the technological development in the wider patent network. Furthermore we predict the clustering algorithm we apply, will group patents based on the measure of similarity determined within the citation vector. When analysing dendrograms produced from our hierarchical clustering techniques, we will describe systematically the dynamics of cluster evolution analogous to the elementary events laid out by Palla et al. (2007). Erdi's hypothesis differs from ours, as he was looking to identify a new class of technology prior it being introduced to the CPC classification using the method of 'back testing'(Érdi et al., 2012). In our case the Y section of the CPC classification system was introduced in 2013 and patents existing patents underwent a re-classification programme. As a result when the Y classification tags were introduced, all classes, groups and subgroups existed. However, we still rely on the predictive methodology that we should be able to 'predict evolution from the more distant past to the more-recent past'. Our back testing methods will demonstrate if this criteria is met and we hope to identify technological clusters prior to their adaption as commonplace in CCMT research and development and co-classification analysis to determine why and how clusters of technology evolved over time in relation to other industries.

3.3 Data Collection and Biblio Feature Extraction

3.3.1 Patent Data Confines

Using patent data induced a number of confines that had to be predetermined before the process of data collection. The three major variables were the jurisdiction, the time period and the subsection of patent classifications that we would focus our analysis towards.

Jurisdiction

In this project, we use patents from the USPTO. Of all the patent offices that exist globally, the USPTO holds the largest volume of patents and is accordingly considered to be the main market for securing patents for new innovation and technology. The world's leading institutions and companies consistently publish patent applications to the USPTO, as the US is a marketplace with the largest demand for technology and has the fastest technology development in the area of CCMTs. Patents in the US are cited and referenced with greater frequency than any other patent office and the USPTO has the least biased patent ratio with more than half of the patents granted in the US going to non-US entities (Kim and Lee, 2015).

Patent subgroup

Our objectives were identify emerging technology in the realm of CCMTs, and so accordingly we focused on patents that fell within the Y02E class of the cross tagged Y section of the CPC classification scheme. Working with patents that are classified within the Y02E class, enabled us to focus on emerging technology from patents that are related to the reduction of GHGs via energy generation through renewable resources within the Y02E 10/00 subgroup, as their classification states. However, the nature of co-classification will also allow us to see how non Y02E 10/00 groups contribute to the formation of the emerging technology in question. A table of all the subgroups of Y02E 10/549 is available in appendix A.

Time period

Patent data from the USPTO is available as far back as 1976, however for the purposes of our project we used patent data from 1980-2020. Although research and development into

CCMTs started to increase exponentially at the turn of the millennia, it is important for our purposes to use patents from as far back as 1980 in order capture clusters of technological evolution across multiple time periods. Furthermore, there were no limitations placed on the number of patents used in our analysis.

3.3.2 Patent Scraping and Feature Extraction

The patent data was scraped from the USPTO via a LENS.org API. The data was formatted within six data packs which only contained patents that contained classified within the Y02E class. The data was merged and each contained the following features;

1. *Lens_id*
2. *Jurisdiction*
3. *kind*
4. *date_published*
5. *doc_key*
6. *docdb_id*
7. *lang*
8. *biblio*
9. *families*
10. *legal_status*
11. *claims*
12. *description*
13. *publication_type*

All features were dropped from the patent data set apart from 'lens_id', 'biblio' and 'date_published'. The date of publication was required in order to sort the patents into our soon to be defined time periods prior to the creation of the citation vector. 'Lens_id' contains an individual identification number for each patent which is crucial in creation of the citation vector. 'Biblio' contains all the bibliographic information for each patent, and is the most crucial variable from the scraped patent data. Crucially for our objectives, it contains the multiple CPC classifications for each patent, the number of backward citations that the patent has made, and the lens_ids of the patents that make up its forward citations.

Thus, prior to the creation of the citation vector our patent data is comprised of a patent's unique 15-digit lens id, the date of publication. Furthermore, the CPC codes that each individual patent has been classified as, the lens ids of the patents that comprise this particular patent's forward citations and the count of the backward citations are extracted from the 'biblio' variable for each patent. An example of this feature extraction is shown in table 3.1. The patent in table 3.1 has a unique 15-digit lens id number of '000-329-891-057-952' and its publication date was 14/10/1986. The patent falls into multiple CPC sections and this is reflected in the 14 CPC tags extracted from 'biblio' variable. This is commonplace for patents within the Y-section of the CPC classification scheme, including the Y02E tagged patents that we scraped, due to the cross tagging scheme employed to create the Y-classification section.

Table 3.1: The features extracted from the original data set for an individual patent. The beginning of the biblio is included. The patents unique lens id number, date published and CPC classifications are included. The number of backward citations the patent makes are shown as 'patent count' and the unique lens id numbers of the forward citations the patent receives are represented by 'cited by'.

Lens_Id: 000-329-891-057-952	**Biblio(Shortened) :** "{'publication_reference': {'jurisdiction': 'US', 'doc_number': '4616390', 'kind': 'A', 'date': '1986-10-14'}, 'application_reference': {'jurisdiction': 'US', 'doc_number': '66238784', 'kind': 'A', 'date': '1984-10-18'}, 'priority_claims': {'claims': [{'jurisdiction': 'US', 'doc_number': '66238784', 'kind': 'A', 'date': '1984-10-18', 'sequence': 1}], 'earliest_claim': {'jurisdiction': 'US', 'doc_number': '66238784', 'kind': 'A', 'date': '1984-10-18', 'sequence': 1}}, 'invention_title': [{'text': 'Superdensity assembly method and system for plastic heat exchanger resists large buoyancy forces and provides fast melt down in phase change thermal storage', 'lang': 'en'}], 'parties': {'applicants': [{'residence': 'US', 'extracted_name': {'value': 'MACCRACKEN CALVIN D'}}], 'inventors': [{'residence': 'US', 'sequence': 1, 'extracted_name': {'value': 'MACCRACKEN CALVIN D'}}], 'agents': [{'extracted_name': {}}], 'owners_all': [{'recorded_date': '1984-10-18', 'execution_date': '1984-10-18', 'extracted_name': {'value': 'CALMAC MAUFACTURING CORPORATION A CORP. OF NEW YORK'}, 'extracted_address': '150 SUTH VAN BRUNT STREET, ENGLEWOOD, NEW JERSEY, 07631'}]}, 'classifications_ipcr': {'classifications': [{'symbol': 'F28D20/02'}, {'symbol': 'F28F9/013'}]}, 'classifications_cpc': {'classifications': [{'symbol': 'F28F9/0132'}, {'symbol': 'F28F9/0132'}, {'symbol': 'F28D20/021'}, {'symbol': 'F28D20/021'}, {'symbol': 'Y02E60 …
Date_published : 1986-10-14	
CPC_Codes : 'F28F9/0132', 'F28F9/0132', 'F28D20/021', 'F28D20/021', 'Y02E60/14', 'Y02E60/14', 'Y10T29/4938', 'Y10T29/4938', 'Y10T29/49872', 'Y10T29/49872', 'Y10T29/49876', 'Y10T29/49876', 'Y10T29/49945', 'Y10T29/49945'	
Cited_By : '029-849-420-274-91X', '170-437-458-499-648', '034-177-630-805-928', '099-753-864-085-338', '132-459-753-533-617', '179-955-625-726-375', '122-905-627-572-564', '054-517-510-541-108', '104-188-690-476-848', '033-236-758-798-123', '037-617-456-078-539', '067-311-312-699-279', '062-187-610-388-340', '056-675-590-626-983'	
Patent_Count: 27	

3.4 Creation of Citation Vector as a Predictor

Once the data scraping and feature extraction processes had been completed for the entire patent data set spanning 1980 – 2020, we began the process of citation vector creation.

3.4.1 Filtration and determination of time periods

First, due to our objectives being centred around identifying emerging technology clusters within the Y02E 10/00 subclass, we removed all patents from the data set that did not include a 'Y02E 10/00' tag within its CPC classification. This was done by creating a Y02E 10/00 'mask' by which any patent that did not possess a Y02E 10/00 subgroup within its CPC_codes variable being removed from the data set. This reduced the number of patents within the data set from 349,468 to 115,718 patents and limited our citation network to include technologies related to energy generation through renewable energy sources.

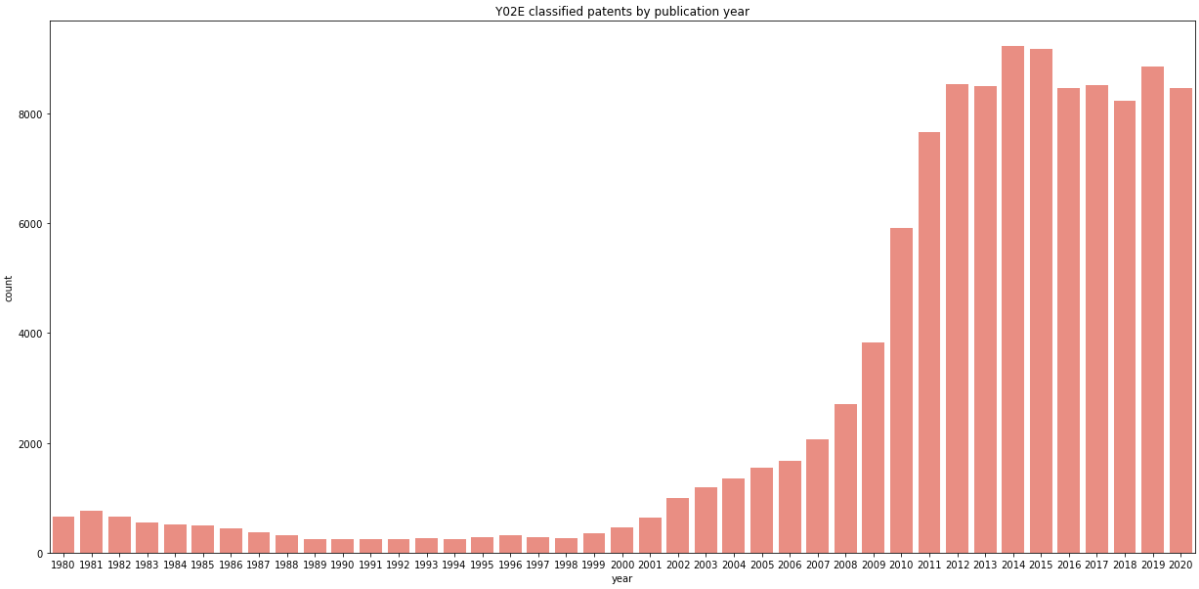
The second stage was to determine the various time periods between the years 1980 and 2020 at which different citation vectors would be created. The citation vectors contained different values at different time periods and allowed us to create vectors that reflect the evolving relationship of patent citations with one another. This allowed 'snapshots' to be taken in a systematic fashion to observe the evolving cluster formation and disappearance that appears over time. The time periods were set as in table 3.2.

Time Period	**Years of Patent Data Included in the Citation Vector**
TP1	1980 – 1995
TP2	1980 – 2008
TP3	1980 – 2011
TP4	1980 – 2014
TP5	1980 - 2017
TP6	1980 - 2020

Table 3.2: The years of patent data included within each time period of citation vector creation

The time periods were defined as such due exploratory analysis conducted on the number of Y02E patents published by year, as seen in figure 3.2. We see that the number of patents classified within the Y02E class, from our data set, increase exponentially after the turn of the millennia and thus it is most relevant to our research to investigate the evolution of technological clusters during and after this period of CCMT growth.

Figure 3.2. Count plot of the number of Patents within the Y02E class granted from the year 1980-2020



3.4.2 Construction of predictor

With the patents filtered to include only those with a Y02E 10/00 subgroup within one of its CPC classifications, and six data sets created to reflect patents only published within the defined years, we created our citation vectors. The citation vector created for each patent will act as its ‘ predictor’ (Erdi et al,.2012), for each of the six time periods (*TP1, TP2, TP3, TP4, TP5, TP6*) using the following steps:

1. We scanned through the CPC tags within the CPC_codes variable for each patent and extract the first 3 values of each tag. For example if a patent possessed the tag ‘H01F 1/0054’, it would have ‘H01’ extracted and inserted into the data frame as a column. As each patent has multiple tags, for the whole data set we created a 127 component vector. The size of the vector naturally decreased the smaller the time period.

2. Next, we then calculated the measure of similarity between patents. For any given patent, we iterated through each one of its forward citations. By taking the lens IDs from the forward citations of the original patent, we used those lens ID to identify the sender patents and the 3 letter CPC codes that the sender patent are classified by. The CPC tags that the sender patents are classified by were then used to indicate if a 1 should be added to the column with the matching 3-letter CPC code within the components of the original patent's vector values.
3. We then weighed the vector values from the forward citations by the number of backward citations (patents that the sender patent has cited in total) by dividing the '1' value placed within the original patent vector by the patent count of the sender patent. As a result, each value within the citation vector was proportional to the frequency that a patent has been cited by other patents within a technological realm at particular time period (Érdi et al., 2012).
4. Any patents that did not receive and forward citations, naturally gained vector values with all zero entities. These were dropped after vector creation. The vector was finally normalised using *sci-kit learn's* 'preprocessing' module.

Each value or coordinate of the citation vector helped quantify the similarity between patents and across various time periods, these values changed. The change in the values reflected the changing role that a patent was contributing to the development of a technology at a specific time (Érdi et al., 2012) as, naturally, patents will influence innovation at different rates over time.

In the vector creation steps, we deviated from the methodology of Erdi et al,.(2012), firstly by not limiting the number of components in our vector to 36 components. This was due to Erdi et al,.(2012) defining 36 technological categories denied by Hall et al, however we had no limit on the number of components within the vector as we added every 3 letter code from every set of CPC codes from all patents. Secondly, in Erdi's methodology, if a patent was cited by a patent within the same CPC class, the original patent did not place a '1' value within the vector co-ordinates of the matched CPC code, so to focus on the combination of different fields. However, in our methodology, since CCMTs are scattered throughout the CPC classifications scheme, we made exception to this rule.

3.5 Application of Clustering Algorithms

The next stage in our methodology was to group the patents within our citation vector into clusters according to their technological relevance, but also identified the evolution of technological clusters over time to identify the development and emergence of new technology.

For this we considered the various clustering algorithms that could be applied to our predictor. Previously, we identified the use of K-means clustering which is one of the commonly use clustering algorithms. It requires a pre-specified number of clusters, k, as an input for the algorithm and thus, an advanced knowledge of how many technological clusters of patents we want to produce. This would be impractical for our research purposes as it would require an advanced understanding of the technological landscape of CCMTs across 30 years of innovation. As a result, K-means was inappropriate for our applications although it is computationally inexpensive and well suited to large data sets such as ours.

Hierarchical Agglomerative Clustering (HAC)

Although our data set is significantly larger than the data set that is used by Erdi et al,. (2012), we also employed Agglomerative hierarchical clustering to best reflect the structural relationships between clusters patents. Agglomerative hierarchical clustering reflects a 'bottom-up' partitioning approach where each data point or observation is individually grouped at first. These observations are successively grouped until every observation forms one large cluster (Bunge and Judson, 2005). This method is computationally less expensive than the 'top' down approach of divisive hierarchical clustering, in which all observations are clustered into one cluster and successively split into smaller clusters and ultimately end as individual observations (Bunge and Judson, 2005).

When applying hierarchical clustering algorithms, the linkage criteria or similarity measure that determines how the distance between different clusters is calculated needs to be determined. We employed Ward linkage, as previously conducted by Erdi et al,. (2012), in which the distance between clusters is the sum of squared differences with all clusters. The main advantage of using Ward is that it induces the most evenly sized clusters. The distance

metric used to calculate the distance between data points is the Euclidean distance, as in Erdi et al ,.(2012) and measures the shortest distance between data points.

Clustering Steps

1. HAC algorithm was applied to the citation vectors created for based on patents within each of our six pre-defined time periods (*TP1 …..TP6*).
2. After clustering was applied to the citation vectors, the resulting clustering methods were visualised in dendrograms. They were essential in determining the optimal number of clusters for each time period. The clustering hierarchy was reflected by the root node which encompassed the entirety of the clustered data and the height of each binary branching point represented the distance between the two branches of clusters.
3. Once the optimal number of clusters for each time period was identified for each time period, from the resultant dendrograms, we were able to predict the contents of each cluster and direct which technological clusters at which points in time look to show the emergence of novel technology. We investigated which tags within the Y02E 10/00 subgroup were most dominant within each of the clusters produced and looked for structural changes in the clusters across time periods. This directed our second round of clustering; intra-cluster clustering.
4. Having identified the contents of the clusters for each of the six time periods, we produced another set of dendrograms focusing on one cluster of patents and again identifed the contents of each of the new, more granular clusters.

3.6 Validation & Detecting structural changes in clusters & co-classification analysis

Validation of our results required us to systematically identify structural changes and the emergence or disappearance of technological clusters. We ultimately identified these changes in dendrograms, in quantitative insights into which CPC classification tags were most prevalent in clusters of interest and once particular technologies were identified, we used ‘back testing’ to confirm the this technology was a candidate for an emerging field of interest (Érdi et al., 2012). Finally, co-classification analysis was used to further understand

the evolution of a technological cluster over time by analysing the technologies in close proximity within the patent citation network.

3.6.1 Detecting Structural Changes and Evolving clusters

The interpretation of our cluster formation and dynamics was based on the work of Palla et al,. (2007); this described the dynamics with regards to cluster death, birth, growth, shrinking and merging. Figure 3.3. As carried out by Erdi et al,. (2012) we divided these temporal changes seen in our dendrograms over time into four types of events; increased or decreased in height of a branching point, formation of a new cluster or fusion of two clusters.

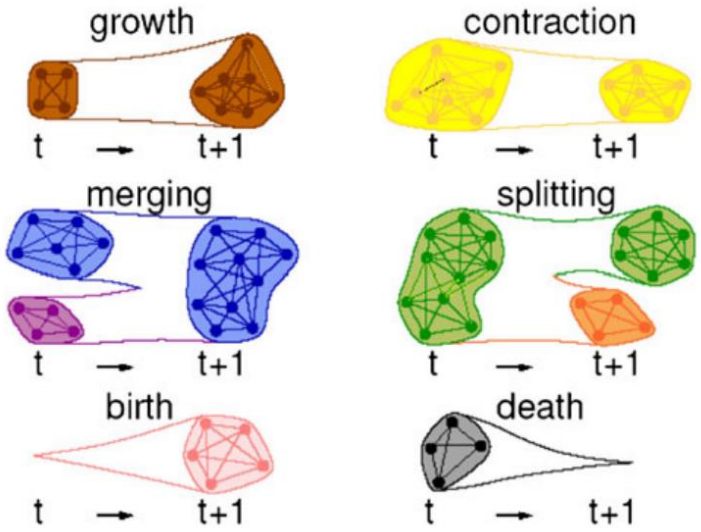


Figure 3.3. Cluster dynamics as set out in Palla et al,. (2007)

3.6.2 Back testing and Co-classification Analysis

To test whether a cluster is of significance in our aims of predicting emerging technologies, once the cluster's were differentiated by their composition of Y02E 10/00 tags, we employed a process of back testing. This involved us comparing the time period at which a technology cluster first appeared and grew to dominate multiple clusters, to market analysis and outlooks of that technology at the same point in time. This confirmed if our methodology correctly predicted the emergence of a technology, prior to it reaching market fruition.

We also employed co-classification analysis, which involved us looking at clusters of significance and analysed the top 10 CPC tags that made up that cluster of patents, not limited to tags of the Y02E

10/00 subgroups. This helped us understand the evolving relationship a technology has over time with other technological fields throughout its evolution.

4. Results & Discussion

Within this chapter, we present the results of our post vector creation clustering, cluster insights and present a case for our work to be a framework for forecasting of emerging technology. The first section will highlight the visual changes seen in the dendrograms from the first round of clustering and the key cluster insights in relation to the Y02E 10/00 subgroup tags. The second part will encompass the second round of clustering where we identify a technology, within a Y02E 10/00 subgroup, that is shown to evolve in significance over time from our clustering and cluster insights. The final part will present our use co-classification analysis and back testing to show the identified technology's interaction with other technology classifications over time and the market uptake of the emerging technology over time.

4.1 Vector Creation

Table 4.1, as seen below, highlights the number of patents used to construct the citation vector for each time period. Within the dendrograms that visualise our citation vector, these Y02E 10/00 patents will be the *x*-axis and the *y*-axis will denote the distance between the patent clusters created.

Year	**1995**	**2008**	**2011**	**2014**	**2017**	**2020**
Number of patents used to create citation vector	6552	20,373	37,774	64,037	90,186	115,718

Table 4.1. Count of number of patents used to create the citation vector for each time period

4.2 First Round Clustering Results

The results of our first round of clustering can be see across all six time periods (figure *) as previously defined and represent the results of the hierarchical clustering of patents within the Y02E 10/00 subgroup. We were first able to visualise the temporal changes in the patent cluster structure.

4.2.1 First Round Clustering Dendrograms

The first observation to make is that for all six time periods, the clustering algorithm determines that there are three significant technology clusters within the Y02E 10/00 subgroup (Figure 4.1). The main three branches are extremely stable and identifiable throughout all six time periods, despite the exponential increase in the number of patents, from 1995 to 2020, that are input into the citation vectors and thus the algorithm. Due to the sensitivity of hierarchical clustering we can induce that the largest technological clusters have been correctly identified; the citations vectors and correctly differentiation patents from different technological fields. At this point in our analysis, the quantitative changes that we can observe are that for each time period, with the number of patents incorporated into the citation vector increasing (Table 4.1) we see that the heights of the branching points for the three main clusters identified in the dendrograms increases from 1995 to 2017. Furthermore, between graph A and graph B in figure 4.1, we see that the cluster in green (cluster 1) and the cluster in red (cluster 2) proportionally become more distant in similarity as the branching point is lowered. This change is resumed through the remaining time periods.

The stable nature of the three main clusters identified by our clustering methodology does not transfer to the cluster evolution of the smaller patent clusters identified within the larger clusters. The orange cluster (cluster 0), between 1995 (graph A) and 2008 (graph B) appears to contract in size, before merging between 2011 (graph C) and 2014 (graph D) and splitting in 2017 (graph E). The central cluster (cluster 1) appears to grow and contract multiple times from 1995 to 2011 before splitting into two defined sub clusters in 2014 as seen in graph D. The red coloured cluster (cluster 2) from 2008 to 2017 contracts nut interestingly splits into 4 subclusters by 2017 (graph E) and then merges into 3 subclusters by 2020 (graph F).

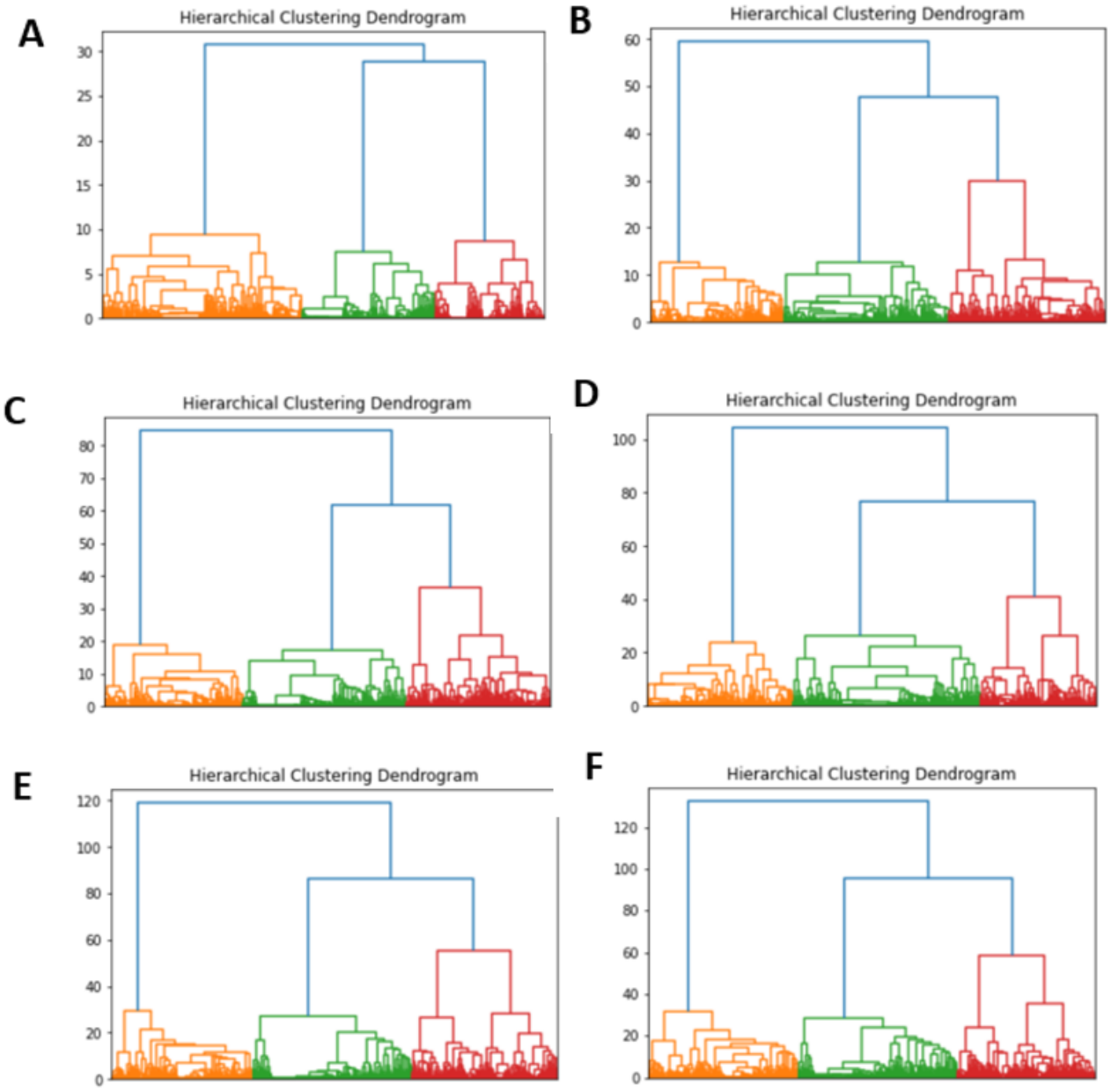


Figure 4.1 Dendrograms produced from the first round of clustering. Time periods are defined from 1980 to as follows; Graph A – 1995, B – 2008, C – 2011, D – 2014, E – 2017, F – 2020.

4.2.2 First Round Cluster Insights

Due to all three clusters expressing interesting subcluster evolution, it is appropriate we present more qualitive analysis of the results of the cluster insights. Table 4.2 compiles the top ten Y02E 10/00 tags for each of the three main clusters for 1995, 2014 and 2020, as

reflected in the dendrograms A, C and F in figure 4.1.

From the cluster insights we see that the top 10 Y02E 10/00 tags for the orange cluster (cluster 0) and the green cluster (cluster 1) are remarkably stable from 1995 through to 2020. The top 10 tags in 1995 still comprise the top 10 tags in 2020 and the dominant tags have not changed significantly. However, the composition of dominant tags in cluster 2 includes a noticeable change. The tag Y02E 10/00 which represents patents which are related to Organic PV cells (Appendix A) dominates the cluster by 2020, yet in 1995 is insignificant. The composition of tags in 2014 shows that Organic PV cells increase in prevalence over time and overtake the other forms of PV cells.

Year	**1995 (graph A)**	**2014 (graph D)**	**2020 (graph F)**
Cluster 0	Y02E10/44 922 Y02E10/40 917 Y02E10/47 542 Y02E10/10 154 Y02E10/46 133 Y02E10/52 124 Y02E10/50 70 Y02E10/56 64 Y02E10/60 49 Y02E10/72 17	Y02E10/47 4761 Y02E10/50 3067 Y02E10/40 3001 Y02E10/44 2797 Y02E10/56 2533 Y02E10/10 1072 Y02E10/52 984 Y02E10/46 968 Y02E10/76 742 Y02E10/72 551	Y02E10/47 8366 Y02E10/50 7878 Y02E10/56 6380 Y02E10/40 4488 Y02E10/44 3785 Y02E10/52 2761 Y02E10/76 1530 Y02E10/10 1366 Y02E10/46 1346 Y02E10/60 817
Cluster 1	Y02E10/72 524 Y02E10/30 452 Y02E10/20 360 Y02E10/74 196 Y02E10/728 138 Y02E10/46 64 Y02E10/70 34 Y02E10/44 14 Y02E10/40 12 Y02E10/727 11	Y02E10/72 14001 Y02E10/30 3901 Y02E10/20 3301 Y02E10/728 3119 Y02E10/74 2137 Y02E10/76 1237 Y02E10/46 852 Y02E10/727 675 Y02E10/70 470 Y02E10/50 297	Y02E10/72 23612 Y02E10/30 6031 Y02E10/728 5183 Y02E10/20 5095 Y02E10/74 3262 Y02E10/76 2551 Y02E10/46 1411 Y02E10/727 1229 Y02E10/70 960 Y02E10/56 584
Cluster 2	Y02E10/50 570 Y02E10/548 350 Y02E10/547 250 Y02E10/52 146 Y02E10/544 115 Y02E10/543 72 Y02E10/546 55 Y02E10/541 44 **Y02E10/549 30**** Y02E10/40 26	Y02E10/50 5798 Y02E10/547 4346 **Y02E10/549 4119**** Y02E10/52 2835 Y02E10/541 2275 Y02E10/548 2194 Y02E10/542 1849 Y02E10/544 1284 Y02E10/546 770 Y02E10/40 665	**Y02E10/549 10524**** Y02E10/50 7775 Y02E10/547 6580 Y02E10/52 3433 Y02E10/541 3245 Y02E10/548 2752 Y02E10/542 2679 Y02E10/544 2152 Y02E10/546 1046 Y02E10/543 971

Table 4.2: Top 10 tags of the Y02E 10/00 subgroup in each cluster over three time periods based on the first round of clustering

We can also induce the three different technological clusters that our algorithm has differentiated the patents in our citation vector by. Cluster 0 appears to be dominated by patents related to mounting or tracking (Y02E 10/47), solar thermal energy (Y02E 10/40) and photovoltaic (PV) energy. Cluster 1 is dominated by patented technologies relating to wind turbines (Y02E 10/72), onshore wind turbines (Y02E 10/728) and energy from the sea (Y02E 10/30). Cluster 2 is completely dominated by patents related to Photovoltaic energy and PV cells (Y02E 10/50, Y02E 10/547 and Y02E 10/549).

Of all the subgroups of Y02E 10/00, Y02E 10/549 was best suited at this stage of our methodology to hypothesise as an emerging technology identified by our clustering methods. Due to the rapid rise to cluster dominance of organic PV cell within the qualitative analysis of cluster 2, this was chosen for our second round of clustering to further understand the emergence and evolution of the Y02E 10/549 subclass and confirm our hypothesis.

4.3 Second Round Clustering

The results of our second round of clustering (clustering within cluster two) can be see across five time periods (figure 4.2) ; 2008 – 2017. We did not include 1995 because, as table 4.2 shows, the Y02E 10/549 subgroup was not a significant part of the composition of cluster 2 in the first round of clustering. We also excluded 2020 because the Y02E 10/549 subgroup because we want to track the emergence of the subgroup prior to it reaching its dominance within the patent citation network. This would mean clustering further into cluster 2 in 1995 would not give us any meaningful insight into the evolution of that particular subgroup.

4.3.1 Second Round Clustering Dendrograms

The dendrograms in figure 4.2 represent the results of the intra-cluster hierarchical clustering of patents classified under the Y02E 10/00 subgroup, within cluster two of the first round. Our dendrograms show that our algorithm at first identified four different technological clusters of patents in 2008 and 2011 (graph A and B). In 2014, cluster 2 (red)

and 3 (purple) merge and result in only 3 clusters of patents being reflected in the dendrogram graphs. These main clusters remain relatively stable with all three clusters being detected up to 2017. Interestingly, when looking at a more granular level, cluster 1 (green) in between 2011 (graph B) and 2014 (graph C) splits into two sets of subclusters which could be a consequence of clusters 2 and 3 merging in 2014. Furthermore, following the merging of clusters 2 and 3 in 2014, cluster 0 (orange) contracts in patent size and cluster 2 (red) grows in size.

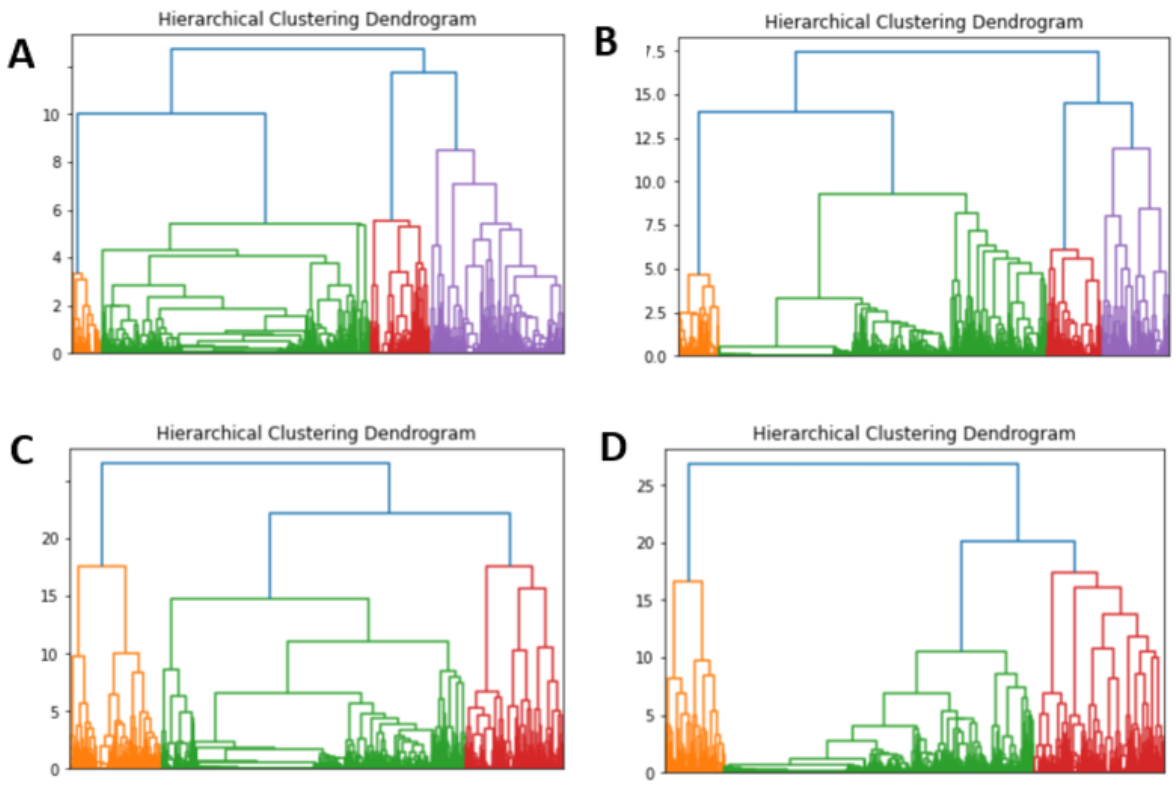


Figure 4.2. Dendrograms produced from the second round of clustering. Time periods are defined from 1980 to as follows; Graph A – 2008, B – 2011, C – 2014, D – 2017,

4.3.2 Second Round Cluster Insights

A more qualitative look into the top five Y02E 10/00 tags that comprise each of the clusters, provides greater insight into the significance of Y02E 10/549 tagged patents as an emerging technology (Table 4.3).

Year	2008	2011	2014	2017
Cluster 0	Y02E10/50 589 **Y02E10/549 335** Y02E10/547 327 Y02E10/548 207 Y02E10/52 122	**Y02E10/549 810** Y02E10/50 453 Y02E10/542 198 Y02E10/541 90 Y02E10/52 86	**Y02E10/549 2352** Y02E10/50 847 Y02E10/542 559 Y02E10/547 379 Y02E10/52 271	**Y02E10/549 2132** Y02E10/50 1707 Y02E10/547 979 Y02E10/52 604 Y02E10/542 493
Cluster 1	Y02E10/547 860 Y02E10/50 800 Y02E10/548 717 Y02E10/52 440 **Y02E10/549 395**	Y02E10/547 2239 Y02E10/50 2082 Y02E10/548 1221 Y02E10/52 1022 **Y02E10/549 84**	Y02E10/50 1567 Y02E10/52 1130 Y02E10/40 506 Y02E10/56 503 Y02E10/47 502	**Y02E10/549 2264** Y02E10/50 519 Y02E10/542 332 Y02E10/541 139 Y02E10/52 127
Cluster 2	**Y02E10/549 489** Y02E10/542 149 Y02E10/547 85 Y02E10/50 68 Y02E10/546 42	**Y02E10/549 822** Y02E10/542 261 Y02E10/547 203 Y02E10/50 101 Y02E10/544 81	Y02E10/547 3746 Y02E10/50 3384 Y02E10/541 1933 Y02E10/548 1915 **Y02E10/549 1586**	Y02E10/547 4710 Y02E10/50 3929 **Y02E10/549 2561** Y02E10/541 2290 Y02E10/548 2132
Cluster 3	Y02E10/541 151 Y02E10/50 120 Y02E10/548 110 Y02E10/543 30 Y02E10/547 29	Y02E10/541 362 Y02E10/548 275 Y02E10/50 230 Y02E10/547 132 Y02E10/543 98		

Table 4.3: Top 5 tags of the Y02E 10/00 subgroup in each cluster over three time periods based on the second round of clustering

The cluster insights show that the Y02E 10/549 tag was dominant in two of the four main clusters as far back as 2008. This, shows that our methodology is capable of identifying technology clusters based on their citation network years before they dominate clusters in 2020. Although the tag loses the dominance over a cluster in 2014, due to the algorithms determination of a merging between clusters 2 and 3, by 2017 Y02E 10/549 tagged patents are dominating two clusters again. We are thus able to confirm that organic PV cells, by the

citation network of patents within its CPC classification, is a potential candidate for the forecasting on an emerging technology.

4.4 Back testing

The key part of our validation process was to predict the emergence of a new technology prior to its market fruition, without prior expertise of the field in question. Thus far we have identified the technology class of 'Organic PV cells' within the field of 'Photovoltaic (PV) cells' as an emerging technology as far back as 2008.

Firstly, by analysing the market activity and outlook of PV cells we can confirm that Organic PV cells as a technological category, were correctly identified by our methodology as an emerging technology. According to the International Renewable Energy Agency report on the future of Solar PV cells, the worldwide capacity of installed PV power increased from the year 2000 by 45% year over year up to 2018 (Market Report Series: Renewables 2017 – Analysis - IEA, 2021). In 2018 it was forecast that the global installed capacity of PV cells is predicted to rise sixfold by 2030, with organic PV cells becoming a dominant proponent of this growth (Market Report Series: Renewables 2017 – Analysis - IEA, 2021). Solar PV power's share of electricity supply is forecast to reach 25% and would imply a ten-fold increase in PV cell activity since 2016 (Market Report Series: Renewables 2017 – Analysis - IEA, 2021). Although at this moment, conventional PV cells dominate the existing market, the near future looks to be dominated by organic PV cells, as identified in our clustering insights, due to recent advances in molecular engineering and nanotechnology (Bhanvase et al., 2018).

Secondly, we analyse the market sentiment around PV cells and organic PV cells in 2008 when our algorithm first detected clusters of patents dominated by Y02E 10/549 tags. In 2008 PV cells produced 36.1 GW of power worldwide, by 2016 it was 259.7 GW and in 2022 predicted to produce 438.3 GW (Market Report Series: Renewables 2017 – Analysis - IEA, 2021). Together with PV cells being a relatively insignificant feature of feature of renewable energy in 2008 when organic PV cells were first identified as a emerging technology, and with organic cells being forecast to be a dominant technology of the future; we can confirm our hypothesis that our methodology identified emerging technologies prior to market fruition and without industry expertise.

4.5 Co-classification Analysis

The co-classification analysis conducted looks into the clusters produced in our dendrograms from our second round of clustering and compiles the top 10 CPC tags within those main clusters. The cluster insights were not limited to those within the Y02E 10/00 subgroup, in order to demonstrate the evolving relationship patents related to organic PV cells (Y02E 10/549) have had since the year it was identified as an emerging technology from our results in 2008. Table 4.4 highlights the technological fields that the non - Y02E 10/00 tagged patents that dominated their respective clusters have been classified as within CPC. Only those clusters dominated by Y02E 10/549 tags have been included. A full breakdown of the top ten tags in each cluster is in Appendix B.

Year	Cluster 0	Cluster 1	Cluster 2	Cluster 3
2008	Semiconductor devices adapted for the conversion of energy.	*No analysis as Y02E 10/549 not dominant in this cluster*	Nanotechnology for information processing, materials and for surface science.	*No analysis as Y02E 10/549 not dominant in this cluster*
2011	Semiconductor devices adapted for the conversion of energy & Solid State Devices of Organic Materials.	*No analysis as Y02E 10/549 not dominant in this cluster*	Nanotechnology for information processing, materials and for surface science.	*No analysis as Y02E 10/549 not dominant in this cluster*
2014	Nanotechnology for information processing, materials and for surface science & Solid State Devices of Organic Materials.	*No analysis as Y02E 10/549 not dominant in this cluster*	Semiconductor devices adapted for the conversion of energy.	*Cluster is merged (Figure *)*
2017	Nanotechnology for information processing & Solid State Devices of Organic Materials.	Solid State Devices of Organic Materials.	Semiconductor devices adapted for the conversion of energy.	*Cluster is merged (Figure *)*

Table 4.4. The most prevalent technological fields of the dominant tags across the CPC classification system within the clusters formed from the results of the second round of clustering

The CPC descriptions of the classes of the dominant tags of each cluster within the clusters from the second round of clustering show how the clusters dominated by Y02E 10/549 tagged patents develop with certain technologies over time. In 2008, when we first identified organic PV cells as an emerging technology, two technological clusters of patents existed: semiconductor related technologies and nanotechnology related technologies. These two technologies are consistently dominating at least one cluster throughout the emergence of organic PV cells up to 2017. In 2014, we can see the addition of patents within the field of Solid State Devices comprise of Organic materials

added to the cluster's dominated by organic PV cell technology. We can induce that the development of organic PV cells is inextricably correlated to developments in nanotechnology and semiconductor technologies, without any expertise in the field of organic PV cells.

5 Conclusion

5.1 Research Conclusions

Overview of findings and project

To conclude, this project utilised patent data from the USPTO from 1980 – 2020, specifically patents classified under the remit of technologies related to the reduction of GHG emissions from energy generation. We aimed to extract citations and CPC classifications from these patents to develop a predictive framework with which governments, policy makers and companies can predict the emergence of new technologies using clustering algorithms.

We were able to fulfil out objectives, firstly because we utilised backward and forward citations to construct citation vectors that quantified the similarity patents within the field of CCM had with each, and indeed the other technological fields that these patents are co-classified with.

Secondly, using our hierarchical clustering methodology, we were able to successfully predict the emergence of the technological field that encapsulates organic PV cells as far back as 2008, based solely on the patent citation vector we created. This emergence of the Y02E 10/549 subgroup was identified, before the technology reach market fruition, in concordance market activity in later years and without any external industry expertise. Furthermore, the evolution this technology had, over time, could be captured by using co-classification analysis of our second round clustering results to identify which technological fields are closely related to the emerging technology in question. It confirmed the premise that new technologies are made of different confirms of old technologies.

Business Implications and Recommendations

Our results are a proof of concept for governments, policy makers and companies that are focused on increasing the efficiency of R&D to induce investment within CCMTs. The first recommendation to make is that governments and policy makers globally ensure a robust, reliable and digital patent data base is in place which can ensure that citation networks are easily accessible for further research within the field of forecasting emerging technologies.

Secondly, with more research into forecasting emerging technologies being conducted, governments globally should harness this new tool to induce more investment and innovation into fields identified as candidates. The frameworks should be advertised and should help direct a roadmap for innovation and public policy agendas. Finally, we encourage companies to use the identification of emerging technologies to enhance the investment into R&D that could support not only the large companies, but the individuals and small business that operate within that technological field to promote the iterative nature of innovation.

5.2 Limitations and Future Work

Future work in forecasting emerging CCMTs should address the limitations that most affected our research. The first limitation we encountered is the time lag between the emergence of a new technology and its accumulation of forward citations. Research ahs shown that for a recently granted panted it can take approximately 15 months to accumulate forward citations from sender patents granted after its publication (Csa´rdi et al. (2009). This citation lag, which is applicable to patents in the field of CCMTs, inevitably means there is a delay between a new technological emergence and it being detected by our methodology.

Furthermore, we have a lot of the patent citation network left to quantify. For example, we weigh our citation vector values with the total number of backward citations that the sender patent made. However our citation vector does not weigh our vector in terms of the size technological that the sender patent belongs to. Nor does it take into account the recency of a patent, or the company or institution that publishes the sender patent. Future work should incorporate this into their citation vectors.

Finally a more precise framework for identifying cluster evolution should be defined prior to cluster analysis within the dendrograms produced. This will make the process more quantitative as appose to observing cluster structure evolution with the naked eye.

Bibliography

Albert, M., Avery, D., Narin, F. and McAllister, P., 1991. Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20(3), pp.251-259.

Angelucci, S., Hurtado-Albir, F. and Volpe, A., 2018. Supporting global initiatives on climate change: The EPO's "Y02-Y04S" tagging scheme. *World Patent Information*, 54, pp.S85-S92.

Bhanvase, B., Pawade, V., Dhoble, S., Sonawane, S. and Ashokkumar, M., 2018. *Nanomaterials for green energy*. pp.325-350.

Blair, T., 2021. *Tony Blair: A Bold, Progressive Agenda for Climate Change*. [online] Institute for Global Change. Available at: <https://institute.global/tony-blair/tony-blair-bold-progressive-agenda-climate-change> [Accessed 10 June 2021].

Bunge, J. and Judson, D., 2005. Encyclopedia of Social Measurement.

Carpenter, M., Narin, F. and Woolf, P., 1981. Citation rates to technologically important patents. *World Patent Information*, 3(4), pp.160-163.

Cho, R., Liu, J. and Ho, M., 2021. The development of autonomous driving technology: perspectives from patent citation analysis. *Transport Reviews*, pp.1-27.

Csa´rdi, G., Strandburg, K., Tobochnik, J., E´rdi, P. (2009). Chapter 10. the inverse problem of evolving networks – with application to social nets. In: B. Bolloba´s, R. Kozma, D. Miklo´s (Eds.) Handbook of Large-Scale Random Networks (pp. 409–443). Berlin: Heidelberg.

Depoorter, B., Menell, P. and Schwartz, D., n.d. *Research handbook on the economics of intellectual property law*.

Dutta, S. and Weiss, A., 1997. The Relationship Between a Firm's Level of Technological Innovativeness and Its Pattern of Partnership Agreements. *Management Science*, 43(3), pp.343-356.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P. and Zalányi, L., 2012. Prediction of emerging technologies based on analysis of the US patent citation network. *Scientometrics*, 95(1), pp.225-242.

M. H. Fallah, E. Fishman and R. R. Reilly, "Forward patent citations as predictive measures for diffusion of emerging technologies," *PICMET '09 - 2009 Portland International Conference on Management of Engineering & Technology*, 2009, pp. 420-427, doi: 10.1109/PICMET.2009.5262201.

Fallah, M., Fishman, E. and Reilly, R., 2011. Forward patent citations as predictors for patent valuation. *International Journal of Intellectual Property Management*, 4(3), p.165.

Fleming, L. and Sorenson, O., 2001. Technology as a complex adaptive system: evidence from patent data. *Research Policy*, 30(7), pp.1019-1039.

Hall, B., Jaffe, A. and Trajtenberg, M., 2001. *The NBER patent citations data file*. Centre for Economic Policy Research.

IEA. 2021. *Market Report Series: Renewables 2017 – Analysis - IEA*. [online] Available at: <https://www.iea.org/reports/renewables-2017> [Accessed 16 August 2021].

Jaffe, A. and Trajtenberg, M., 1998. *International knowledge flows*. Tel-Aviv: Foerder Inst. for Economic Research.

Jaffe, A. and Trajtenberg, M., 1999. International Knowledge Flows: Evidence From Patent Citations. *Economics of Innovation and New Technology*, 8(1-2), pp.105-136.

Katila, R., 2000. Using patent data to measure innovation performance. *International Journal of Business Performance Management*, 2(1/2/3), p.180.

Kim, G. and Bae, J., 2017. A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, 117, pp.228-237.

Kim, G., Lee, J., Jang, D. and Park, S., 2016. Technology Clusters Exploration for Patent Portfolio through Patent Abstract Analysis. *Sustainability*, 8(12), p.1252.

Kim, J. and Lee, S., 2015. Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. *Technological Forecasting and Social Change*, 92, pp.332-345.

Kim, K., Han, Y., Lee, S., Cho, S. and Lee, C., 2019. Text Mining for Patent Analysis to Forecast Emerging Technologies in Wireless Power Transfer. *Sustainability*, 11(22), p.6240.

Kyebambe, M., Cheng, G., Huang, Y., He, C. and Zhang, Z., 2017. Forecasting emerging technologies: A supervised learning approach through patent analysis. *Technological Forecasting and Social Change*, 125, pp.236-244.

Li, X., Chen, H., Huang, Z. and Roco, M., 2007. Patent citation network in nanotechnology (1976–2004). *Journal of Nanoparticle Research*, 9(3), pp.337-352.

Ridley, M., 2020. *How innovation works*.

Saviotti, P., Looze, M., Nesta, L. and Maupertuis, M., 2003. Knowledge dynamics and the mergers of firms in the biotechnology based sectors. *International Journal of Biotechnology*, 5(3/4), p.371.

Unfccc.int. 2021. [online] Available at: <https://unfccc.int/kyoto_protocol> [Accessed 10 August 2021].

von Wartburg, I., Teichert, T. and Rost, K., 2005. Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, 34(10), pp.1591-1607.

Walker, R., 1995. *Patents as scientific and technical literature*. Metuchen, N.J.: Scarecrow Press.

You, H., Li, M., Hipel, K., Jiang, J., Ge, B. and Duan, H., 2017. Development trend forecasting for coherent light generator technology based on patent citation network analysis. *Scientometrics*, 111(1), pp.297-315.

Zhou, Y., Dong, F., Liu, Y. *et al.* Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics* **123,** 1–29 (2020). https://doi.org/10.1007/s11192-020-03351-6

Appendix A: Table of Y02E 10/00 subgroups and their CPC descriptions

CPC Sub Group	CPC Description
Y02E 10/00	Energy generation through renewable energy sources
Y02E 10/10	Geothermal energy
Y02E 10/20	Hydro energy
Y02E 10/30	Energy from the sea, eg. Using wave energy or salinity gradient
Y02E 10/40	Solar thermal energy eg. Solar towers
Y02E 10/44	Heat Exchange systems
Y02E 10/46	Conversion of thermal power into mechanical power
Y02E 10/47	Mountings or tracking
Y02E 10/50	Photovoltaic (PV) Energy
Y02E 10/52	PV Systems with concentrators
Y02E 10/541	CulnSe2 material PV cells
Y02E 10/542	Dye sensitized solar cells
Y02E 10/543	Solar Cells from Group II-IV materials
Y02E 10/544	Solar Cells from Group III-V materials
Y02E 10/545	Microcrystalline silicon PV Cells
Y02E 10/546	Polycrystalline silicon PV Cells
Y02E 10/547	Microcrystalline silicon PV Cells
Y02E 10/548	Amorphous silicon PV cells
Y02E 10/549	Organic PV cells
Y02E 10/56	Power Conversion systems
Y02E 10/60	Thermal-PV hybrids
Y02E 10/70	Wind Energy
Y02E 10/72	Wind turbines with rotation axis in wind direction
Y02E 10/727	Offshore wind turbines
Y02E 10/728	Onshore wind turbines
Y02E 10/74	Wind turbines with rotation axis perpendicular to the wind direction
Y02E 10/76	Power conversion electric or electronic aspects

Appendix B: Co-classification analysis table of CPC tags

Year	Cluster 0	Cluster 1	Cluster 2	Cluster 3
2008	Y02P70/50 972 Y02E10/50 589 Y02E10/549 335 Y02E10/547 327 H01L31/1804 246 H01L31/048 216 Y02E10/548 207 H01L31/202 150 H01L31/046 144 H01L31/022425 137	*No analysis as Y02E 10/549 not dominant in this cluster*	Y02E10/549 489 Y02P70/50 398 B82Y10/00 307 B82Y30/00 202 H01L51/4253 158 Y02E10/542 149 H01L51/0037 149 H01L51/0036 142 H01L51/0078 108 H01L51/0053 104	*No analysis as Y02E 10/549 not dominant in this cluster*
2011	Y02E10/549 810 Y02P70/50 637 Y02E10/50 453 H01L31/048 217 Y02E10/542 198 C09K11/06 191 H01L51/5012 165 H01L51/0036 164 H01L51/0043 155 H01L51/0059 147	*No analysis as Y02E 10/549 not dominant in this cluster*	Y02E10/549 822 Y02P70/50 756 B82Y10/00 580 B82Y30/00 362 H01L51/4253 290 Y02E10/542 261 H01L51/0036 251 H01L51/0037 218 Y02E10/547 203 B82Y20/00 186	*No analysis as Y02E 10/549 not dominant in this cluster*
2014	Y02E10/549 2352 Y02P70/50 1938 B82Y10/00 913 Y02E10/50 847 H01L51/0036 671 Y02E10/542 559 H01L51/4253 550 B82Y30/00 514 H01L51/0037 447 C09K11/06 408	*No analysis as Y02E 10/549 not dominant in this cluster*	Y02P70/50 8282 Y02E10/547 3746 Y02E10/50 3384 H01L31/1804 2629 H01L31/022425 2300 Y02E10/541 1933 Y02E10/548 1915 Y02E10/549 1586 Y02E10/52 1434 H01L31/0322 1336	*Cluster is merged (Figure *)*
2017	Y02P70/50 3138 Y02E10/549 2132 Y02E10/50 1707 Y02E10/547 979 B82Y10/00 847 H01L51/0097 806 H01L31/1804 724 H01L2251/5338 623 H01L31/048 620 Y02E10/52 604	Y02E10/549 2264 Y02P70/50 912 C09K11/06 856 H01L51/0036 699 H01L51/0043 589 Y02E10/50 519 H01L51/501 474 C08G61/126 460 H01L51/0072 456 H01L51/4253 451 .	Y02P70/50 9658 Y02E10/547 4710 Y02E10/50 3929 H01L31/1804 3227 H01L31/022425 3043 Y02E10/549 2561 Y02E10/541 2290 Y02E10/548 2132 H01L31/068 1770 Y02E10/52 1722	*Cluster is merged (Figure *)*

Appendix C: Code Repository and Project Management

Code Repository

All the code used throughout this project is found in the a GitHub repository titled ‘ *MSIN0114 – Dissertation’.*

https://github.com/Ayushj14/MSIN0114-Dissertation-

Ayushj14/MSIN0114-Dissertation

The code is attached within a series of Jupyter notebooks. These can be downloaded and or visualised in Github.

Project Management

The project management for this dissertation was conducted using a minutes folder, to keep track of weekly meetings and activities with my supervisor and a Notion Board. The Notion Board kept track of weekly activities, immediate and long term tasks and also contains the minutes of the meetings with my supervisor. A link to this board is attached below:

https://diligent-tracker-067.notion.site/ME2-Cambridge-Project-2ef73d9826cf472094e4f58f26f2dbab