Measuring innovation in electric secondary batteries: trends in technologies and geographic origins

Philipp Metzger[1,*], Author 2[2], Author 3[1,2,+], and Author 4[2,+]

[1]NOVA IMS Information Management School, Lisbon, 1070-312, Portugal
[2]Affiliation, department, city, postcode, country
[*]philipp.metzger.bat.pat@gmail.com
[+]these authors contributed equally to this work

ABSTRACT

This study is a descriptive analysis of patent data on electric secondary battery technologies from the years of 2000 to 2019. Its goal is to enrich the existing literature on the topic by providing a deeper insight into where innovation in this field has been created, which technologies and concepts are emerging, declining, or becoming established, both in absolute and in relative terms, and how geographic locations can be characterized in terms of their position in the technology space. For this purpose, worldwide battery patent counts were created and broken down alongside their geographic, temporal, and technological dimensions.

1 Introduction

The importance of batteries has been growing and with their new main fields of deployment being electric mobility and the storage of energy generated by fluctuating, non-dispatchable sources, their importance is expected to grow further in the coming decades. The authors of a report about the recent developments in electricity storage technologies created by the International Energy Agency (IEA) in cooperation with the European Patent Office (EPO)[1] claim that under the Sustainable Development Scenario (SDS) defined by the IEA, "the level of deployment and the range of applicability of batteries [...] expands dramatically" and "charging batteries in electric vehicles will become the largest single source of electricity demand, accounting for around 5% of global demand by 2050." Furthermore, "the use of batteries in stationary energy storage applications is [already] growing exponentially."[1] Given this dynamic, it is worthwhile to study where and in which technological sub-areas innovation in the field of electric batteries is being and has been created.

For this study a dataset containing 92,700 international electric secondary battery patent families from the years of 2000 to 2019 was compiled and analyzed. The raw data was extracted from PATSTAT Online (edition: Autumn 2021), the web interface of the PATSTAT database[2] maintained by the European Patent Office containing a vast collection of worldwide patent data for the purpose of statistical analysis.

2 Foundations and concepts

In this chapter, literature and concepts that are relevant for this study are presented. The following section provides a short overview on the advantages and limitations of using patents as a source of information.

2.1 Patents as an indicator

Patents contain information such as geographic locations associated with inventors, descriptions and classifications of the respective inventions, and temporal features like their publication date. This allows for the aggregation of patent counts alongside geographic, temporal, and technological dimensions. Furthermore, patents have proven to be a good proxy for measuring innovation. Griliches[3] explains that patents "are available; they are by definition related to inventiveness, and they are based on what appears to be an objective and only slowly changing standard."

Even though patents are a good proxy for innovation it must be noted that they are not per se a precise measure, since not all inventions are patented (or even patentable) and the inventions that are patented differ greatly in quality[3]. The first point of concern has to be accepted as a limitation of patent analysis but the latter one can be addressed by restricting the patent data analysed to a subset that can be expected to be more homogeneous. This is the approach followed for this study and will be explained in more detail later on. The following section presents literature that makes use of battery patents for their analyses.

2.2 Analysis of battery patents

Battery patent data has played a central role in other work, such as Aaldering et al.,[4] an analysis built upon battery patent data highlighting developments in post-Lithium-ion battery technologies, Malhotra et al.,[5] a citation network analysis combining knowledge extracted from patent data with results from interviews conducted with Lithium-ion battery experts, and Stephan et al.,[6] a Lithium-ion battery patents data-based study investigating how sectoral diversity and sectoral distance of prior knowledge affect certain features of subsequent knowledge.

The current study extends the existing literature on this matter, specifically it aims to expand the findings presented in the report created by the IEA in cooperation with the EPO.[1] It can be understood as a continuation of their methodological approach, enriched by some reasonable additions, which allow for a more granular perspective on some aspects of this topic.

Unsure whether or not to keep the following paragraph:

The report by IPA and EPO presents information extracted from patents related to batteries and electricity storage and was published in September 2020. The information that is presented in it unveils how the yearly number of new patents in electricity storage has risen both robustly and disproportionally to the total number of patents, how patent counts are distributed across some geographic locations, several technologies, and the most relevant applicants and how applicants can be characterized.

The current study—in contrast to the just mentioned report—focuses on battery technology only. The research gaps that were identified by the authors and which this study aims to fill are how patent counts are distributed across continents, how scaling them by the sizes of the respective labor forces affects the outcome of the analysis, what their distribution across another technological classification scheme looks like, how countries can be characterized based on their position in a resulting technology space, and what information can be extracted from patent titles and abstracts.

The authors of the report by IEA and EPO use the concept of *international patent families (IPF)* for aggregating and counting patent applications. They claim that an IPF "is a reliable proxy for inventive activity because it provides a degree of control for patent quality by only representing inventions for which the inventor considers the value sufficient to seek protection internationally." The following section explains the differences between patent applications, patent families, and international patent families and highlights some advantages and limitations of the latter concept.

2.3 International patent families

A patent application is a formal request made by a patent applicant (or multiple applicants) at one patent office of their choice. This could be the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), or any other national or regional patent office. The applicants' goal is to obtain legal protection for their invention that they deem (1) directed to patentable subject matter, (2) novel, (3) inventive, and (4) capable of industrial application, which are the four conditions for patentability.[7] The term *patent family* refers to the whole set of patent applications covering the same invention.[8] By counting patent families instead of applications, double counting of inventions is avoided. Now, restricting one's scope to only patent families that contain applications filed in two or more countries, one obtains international patent families. The benefit of this restriction is that only patents of higher value are assessed, resulting in a more homogeneous dataset with better comparability between elements.

Three limitations regarding the concept of international patent families should be considered that are discussed in Schmoch and Gehrke[9]: Firstly, "the propensity to patent in foreign countries differs between countries of origin," meaning that for example an applicant from a European country might be more inclined to seek protection in another European country than an applicant from China might be inclined to seek protection in the USA. This can be problematic because both situations would imply that the respective patent is filed in two countries, making their patent family an international patent family. Secondly, "in specific analyses for technologies, the patent numbers of Japan appear to be overestimated."[10] Thirdly, the model [of international patent families] is influenced by the limited PCT transfer of China, as it is possible that a family with seemingly two members at the stage of applications is reduced to one member later on." For more information on the limited PCT transfer of China, we refer to Frietsch and Kroll[11]. Schmoch and Gehrke[9] discuss several other concepts that exist parallel to IPFs, highlighting their advantages and limitations. For the reason of comparability to the report by IEA and EPO[1], the authors of this study decided to apply the concept of international patent families, so all depicted counts refer to IPFs. The following section briefly explains what electric secondary batteries, the subject-matter of this study, are and their relevance in today's economy.

2.4 Electric secondary batteries

Electric secondary batteries are able to receive energy in the form of electricity, store it, and at a later point in time—and with a certain energy loss due to the energy conversion processes taking place—release it again, feeding electricity back to the grid or powering a given application. Secondary batteries are rechargeable, unlike primary batteries that can only discharge once and then need to be discarded. In the context of the ongoing energy transition away from dispatchable sources of electricity such as coal-fired power plants and towards electricity sources such as wind and solar power, whose input is not controllable and hardly synchronous with the population's and the industry's needs, batteries and other means of energy storage are the bridge

that conjoins the temporal gap between supply and demand. Furthermore, accelerated electrification in the transporting sector, especially in individual mobility, bids for more batteries providing higher capacities and longer lifespans.

When speaking of batteries, one has to differentiate between the terms “battery”, “module”, and “cell”. Whilst an entire battery pack potentially consists of multiple modules that are "wired in series and/or (less often) parallel",[12] a module itself consists of multiple cells that "are connected in series or parallel".[12] For simplicity’s sake, electric secondary batteries, meaning battery packs in their entirety, will henceforth be referred to as *batteries*. The following section introduces the classification scheme used for identifying relevant battery patent data and technological sub-areas in the field of batteries that were relevant for this study.

2.5 International Patent Classification (IPC)

The international patent classification system (IPC) provides a hierarchical classification scheme that is used for categorizing patents according to different technological areas. This study’s analysis is build on data related to patents that can roughly be characterised as four groups: (1) innovations related to casing, wrapping, or covering, i.e. non-active parts of batteries, (2) innovations in battery electrode manufacturing, (3) innovations related to the manufacturing process of secondary cells, and (4) innovations related to charging of batteries. Patents belonging to these four groups were identified using the IPC classification scheme, which is a constituent part of the data provided at the PATSTAT database. The following chapter describes the results obtained by counting battery IPFs both on global and national levels, disaggregating the dataset into different technologies and analyzing how countries position themselves in the resulting technology space. Furthermore, the results of a text mining approach deployed on patent titles and abstracts are presented.

3 Results

3.1 Some introductory figures

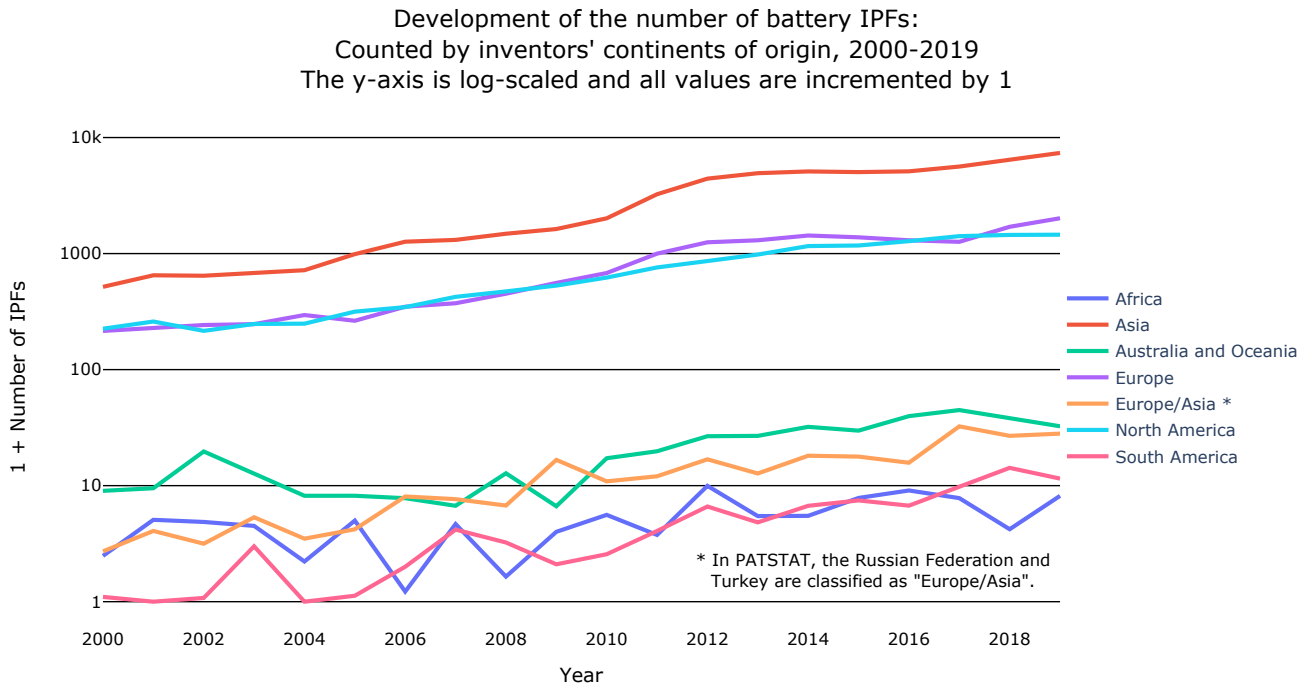

Figure 1

To ease the reader’s way into the topic and to demonstrate the relationship between this work and the report by IEA and EPO[1], this section contains some introductory numbers and figures highlighting key points about worldwide battery patenting trends.

The global yearly number of battery IPFs has mostly increased between adjacent years in the 20-years time frame assessed in this study. Only from 2001 to 2002 and from 2014 to 2015 slight decreases in IPFs are measured. The whole time period’s mean year-over-year increase in IPFs for the given time frame is 14.30% (all percentages are rounded to two decimal places).

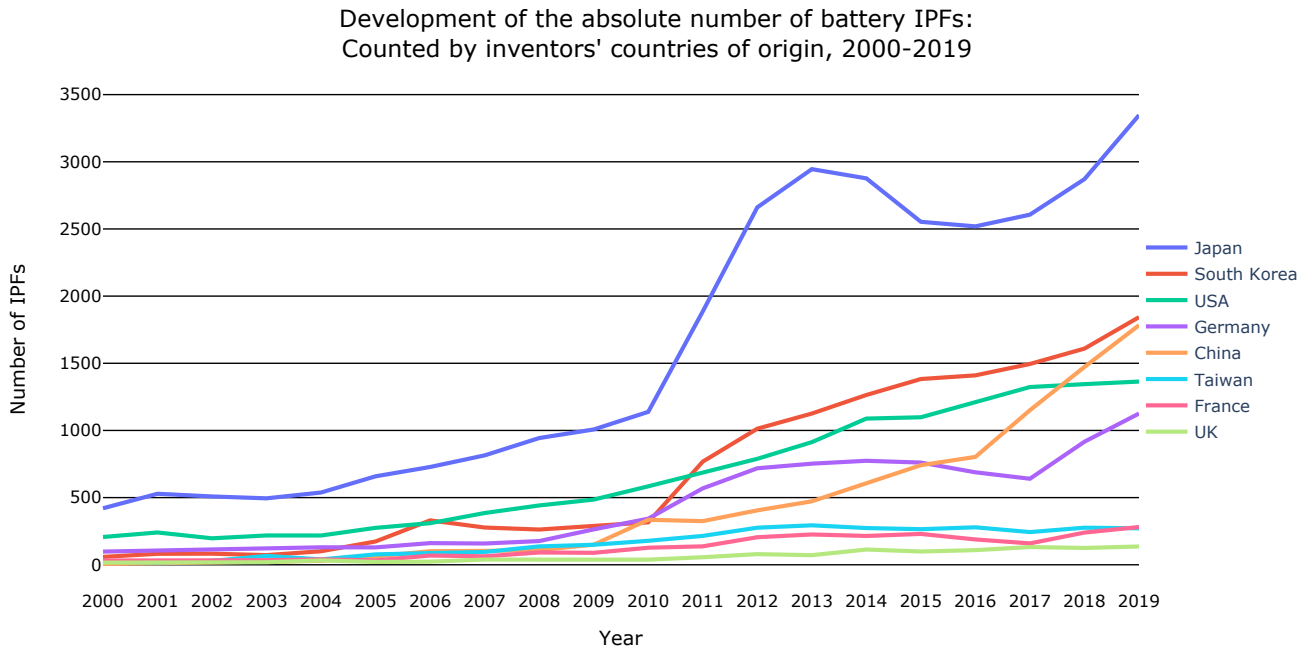

Figure 2

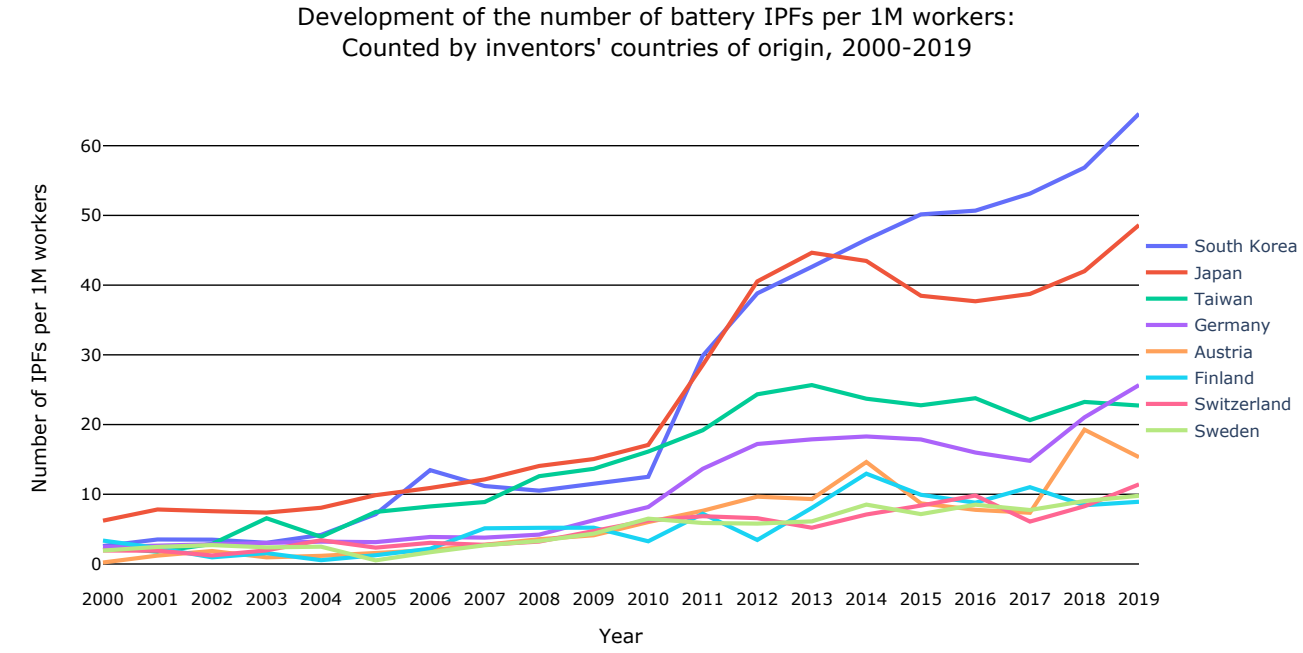

Figure 3

Asian countries dominate the battery market: The Asian continent's mean annual battery IPF output is approximately four times higher than Europe's and North America's (factor 3.57 and 4.09, respectively). Furthermore, the number of IPFs from Asia has increased by 15.97% on average each year in 2000-2019. The average increase values for Europe and North America are 13.46% and 10.78%, respectively. A comparison of continents' battery IPF output is presented in Fig. 1. Please note that the y-axis is log-scaled.

Breaking down battery IPF counts by inventors' countries of origin (Fig. 2), the dominance of Asia becomes even more apparent: In 2019 the three top countries in terms of battery IPF output were Japan, South Korea and China, followed by the USA, Germany, France, Taiwan and the UK. Japan, historically the undisputed leader in battery innovation, is displaying a sharp increase in inventive output since 2016, which was preceded by a downswing after 2012. China is catching up with South Korea, which has held the second place in battery IPF output since 2011. Germany also displays an upward trend in battery IPF output. Please note the similarity of the trajectories presented in this plot to the ones depicted in Fig. 6.2 and 6.3 of the report by IEA and EPO[1]. The higher numbers in this study result from the underlying data being defined somewhat differently, with the largest difference being that IPFs related to the field of battery charging were included in the dataset of the current study.

By scaling the numbers shown in the previous plot by each country's and year's labor force count, one obtains patent intensities.[13] This measure gives the viewer a different perspective on the IPF counts, allowing for assessment of a country's innovative output relative to the size of its working population. Figure 3 shows the eight countries with the highest scaled total battery IPF output over the whole time span and it can be seen that in contrast to Fig. 2, Austria, Finland, Switzerland, and Sweden are part of the top eight. In this light, South Korea overtook Japan in 2014, establishing itself as the global leader of battery patent intensities.

3.2 Battery technologies

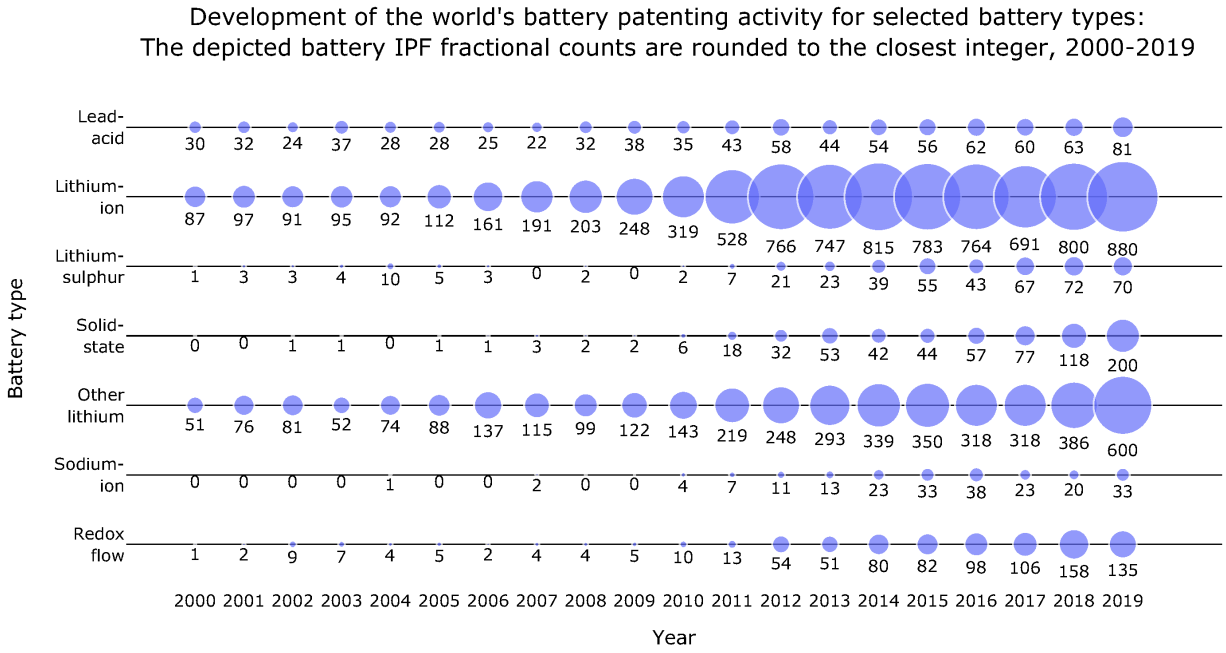

Figure 4

By assigning battery technology sub-areas to patent families a disaggregation of the dataset into 19 battery cell technologies was obtained. This process is described in detail in 5.1. The technology classes used in this study are Lead-acid, Lithium-air, Lithium-ion, Lithium-sulphur, Solid-state, Other Lithium, Magnesium-ion, Nickel-cadmium, Nickel-iron, Nickel-zinc, Nickel-metal hydride, Rechargeable alkaline, Sodium-sulphur, Sodium-ion, Aluminium-ion, Calcium(-ion), Organic radical, Redox flow, and Nickel–hydrogen.

Figure 4 presents the development of IPF counts of the six major categories. They were selected based on their total IPF count in the given timeframe of 2000-2019. While the number of IPFs related to Lead-acid batteries has been relatively

stable over the depicted 20 years, which resulted in its overall share in battery IPFs to decrease steadily over this time period, Lithium-ion batteries and other Lithium-based battery technologies have soared drastically. Still less relevant today than Lithium-ion batteries, but with considerably higher counts than other smaller battery technologies are the four remaining categories presented in Fig. 4: Patenting activity related to Lithium-Sulphur, solid-state, Sodium-ion, and redox flow batteries have seen a notable increase in IPF numbers in 2010-2019. In 2019 solid-state batteries reached an all time maximum of 200 IPFs and also redox flow battery IPF counts have trended upwards since 2010.

The observation that the recent decade displayed increased patenting activity in these four emerging technologies motivated the way the next part of the analysis is set up: The following section describes the results obtained by clustering countries based on their position in a technology space computed using their technology distribution of the years of 2010-2019.

3.3 Clustering

The most suitable technology space for clustering was found to be spanned by the countries' distribution values over the following technologies: The four emerging technologies Lithium-Sulphur, solid-state, Sodium-ion, and redox flow alongside the older lead-acid technology. Clustering 36 countries, k-means was found to be the clustering algorithm with a better R^2 value for all relevant numbers of clusters (for more details on this metric see section 5.4). Setting the numbers of clusters to two, a clear separation of the dataset between countries with a high focus on lead-acid batteries (82.61% of IPFs are related to lead-acid batteries in this cluster) and countries with comparatively high shares of IPFs related to the four emerging technologies and consequently a relatively low share of lead-acid related IPFs (19.57%) was obtained.

Setting the number of clusters to three, in order to achieve a more granular separation, one finds that the lead-acid focused cluster from the previous stage is still fairly intact, whilst the "emerging technologies" cluster has been separated in two: One cluster that displays a stronger focus on redox flow and solid-state batteries and another that, relative to the other two clusters, has a higher focus on Sodium-ion and Lithium-Sulphur-related IPFs. Figure 5 shows the distribution profiles of the three-clusters solution generated with the k-means variable "random_state" set to zero. The variable "random_state" determines the centroid initialization of k-means and results in deterministic runs of the algorithm when a value is assigned to it.

For this particular clustering solution the countries' affiliations to their clusters are as follows. Inside each cluster, countries are ordered by their total IPF count in the five categories:

- Cluster 1 (12 countries):

  Japan, USA, South Korea, Germany, Italy, Taiwan, Belgium, Austria, Netherlands, Australia, Thailand, Switzerland.

- Cluster 2 (15 countries):

  India, Russia, Turkey, Bulgaria, New Zealand, Luxembourg, Poland, Sweden, Malta, Mexico, North Korea, Kazakhstan, Hungary, Serbia, Greece.

- Cluster 3 (9 countries):

  China, UK, France, Canada, Spain, Israel, Norway, Hong Kong, Ukraine.

Whilst the approximate shape of the clustering profile depicted in Fig. 5 is fairly insensitive to alterations or non-assignation of "random_state", the affiliation of the countries to their clusters varied enough to motivate running k-means a higher number of times (with the variable "random_state" undefined) in order to compute each country's cluster affiliation distribution and using that result to assess which cluster each country belongs to in the majority of events. Running k-means 10,000 times results in the following most probable cluster affiliations:

- Cluster 1 (12 countries):

  USA, Germany, Taiwan, Austria, Netherlands, Thailand, Switzerland, South Korea, Japan, Belgium, Italy, Australia.

- Cluster 2 (17 countries):

  India, Russia, Turkey, Bulgaria, New Zealand, Luxembourg, Poland, Sweden, Malta, Mexico, North Korea, Serbia, Greece, Hungary, Kazakhstan, Israel, Norway.

- Cluster 3 (7 countries):

  Canada, Spain, Ukraine, UK, France, China, Hong Kong.

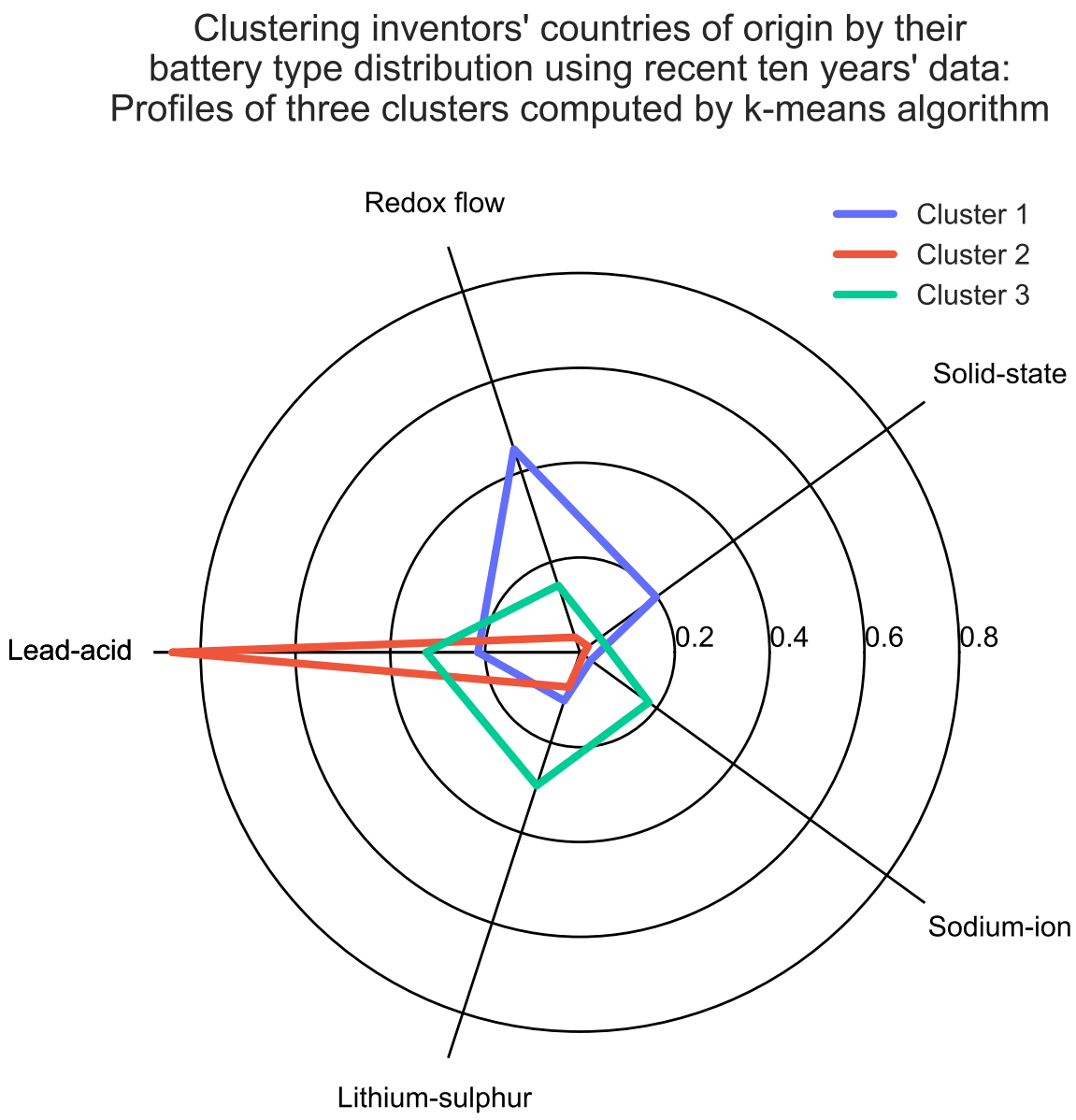
Clustering inventors' countries of origin by their battery type distribution using recent ten years' data: Profiles of three clusters computed by k-means algorithm
Redox flow
Solid-state
Sodium-ion
Lithium-sulphur
Lead-acid
0.2
0.4
0.6
0.8
Cluster 1
Cluster 2
Cluster 3

Figure 5

Inside each cluster, countries are ordered by (1) their probability p to be in this cluster, and (2) their total IPF count in the five categories. Each country's name is coloured according to the following schema, indicating its probability p for belonging to the respective cluster:

$p = 1$ $p \in [0.95,\ 1)$ $p \in [0.9,\ 0.95)$ $p \in [0.85,\ 0.9)$

A value of $p = 1$ indicates that a country was assigned to this cluster during every of the 10,000 runs, meaning that it's cluster affiliation can be expected to be independent from the centroid initialization of the algorithm. The following section describes the results obtained by mining the titles and abstracts of the patent applications contained in this study's dataset for key phrases.

3.4 Title and abstract mining

Two methods were deployed in order to scan patents' titles and abstracts for meaningful phrases: The first one, henceforth called *n-gram counts* was to simply count occurrences of n-grams in patent abstracts and titles for each year. The second method, from now on referred to as *n-gram intensities* was to additionally scale these counts by the respective year's number of abstracts or titles, respectively. The resulting unit of measure for n-gram intensities is occurrences per 1,000 abstracts or titles, respectively, and all depicted n-gram intensities are rounded to the closest integer. Unigrams, bigrams, and trigrams were counted. The resulting n-gram counts and n-gram intensities were sorted in three different ways, which are described in detail in section 5.5. The top 50 increasing trigrams extracted from battery patent abstracts are displayed in descending order of total increase over the given 20-years time period in Fig. 6 (trigram counts) and 7 (trigram intensities). A robust analysis is provided by considering the two figures jointly.

Both the trigram counts and intensities yield several expectable trends like the surge of "lithium secondary battery" or "lithium ion battery". The increase of the term "energy storage system", which is also confirmed in the intensities, might hint at an increase in the importance of increasingly complex systems for managing the storage of energy. This is affirmed by the term "battery management system" that also occurs in both counts and intensities. As already established by Fig. 4, solid-state batteries have been growing in relevance, especially in the past decade. This is confirmed by the increasing counts and intensities for the terms "solid electrolyte layer" and "solid state battery" during that timeframe. Notable trigrams in the subfields of battery charging and electric vehicles are "wireless power transmission" and "electric vehicle charging", which have both increased considerably in both counts and intensities. The surge in relevancy for redox flow batteries (see Fig. 4) is also confirmed by both counts and intensities ("redox flow battery"). The trigrams "plurality battery cell" (results from "plurality of battery cells" due to stop word removal and lemmatization) and "battery module plurality" (both are present in both counts and intensities) hint at a substantial increase in innovative output related to compositions of cells and modules inside battery packs. An unexpected yet reasonable appearance in the top 50 trigram intensities is the term "unmanned aerial vehicle", that exhibited 4, 13, 16, and 8 occurrences per 1,000 abstracts in the years of 2016, 2017, 2018, and 2019, respectively. It indicates increased innovation related to the deployment of battery technology in drones.

4 Discussion

In regards to the results presented in Fig. 2 and Fig. 3, is is worthwhile to mention that comprehensive analyses that were undertaken before defining the final dataset for this study have resulted in the observation that the majority of battery patent applications from China in the considered timeframe of 2000-2019 are filed only nationally. Given the IPF constraint deployed for this study and the report by IEA and EPO[1], these solely-nationally-filed applications are not taken into account in either one. It is reasonable to define the data for the current study and the analysis undertaken by IEA and EPO like this, because it can be expected that patents filed in only one country are of considerably lesser value than international patent families, thus including them would result in a rather inhomogeneous dataset. But nonetheless, the authors of this study find it necessary to mention that if the IPF restriction was to be discarded and one-country patent families were to be considered, China would take the first place in battery patent counts in the majority of years of the recent decade. As a resulting thought, it would be worthwhile to study the battery patenting dynamics of China in detail within the context of future research in order to shed light on why China's battery patenting behaviour is so nationally-focused and what implications this has for studies in this field.

Mention the percentage of battery patents being IPFs? For this I need to rerun create_dataset without restricting the dataset to IPFs. The step of restricting the dataset to IPFs would go to the beginnings of counts_technologies_clustering and title_and_abstract_mining.

Interpreting the clustering solution presented in section 3.3, the three resulting clusters could be characterized as follows:

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
electrode active material	89	82	72	121	136	157	241	278	294	356	427	991	1,253	1,288	1,323	1,386	1,273	1,308	1,382	1,990
active material layer	17	9	29	27	37	96	114	143	134	188	186	364	419	553	499	383	336	510	619	883
lithium secondary battery	46	91	72	65	112	128	192	140	124	178	222	340	451	487	490	410	492	374	540	844
lithium ion battery	15	26	41	31	16	49	58	83	99	121	197	321	476	421	449	520	618	568	702	685
energy storage device	5	41	27	39	33	20	44	70	109	94	147	183	224	367	367	277	370	538	576	648
secondary battery electrode	29	27	21	25	55	82	113	109	95	122	146	286	417	422	384	400	419	441	435	593
electrode current collector	7	7	10	12	15	15	24	46	57	104	96	135	206	171	155	172	215	226	412	539
power storage device	11	3	7	10	2	8	4	30	69	37	183	166	317	310	361	311	311	520	293	394
electrolyte secondary battery	25	27	41	36	65	74	110	118	180	129	113	214	301	269	437	425	412	286	264	405
aqueous electrolyte secondary	24	38	45	46	67	75	111	116	182	135	121	216	307	289	456	427	421	284	248	390
plurality battery cell	3	1	3	1	7	9	10	28	26	29	32	135	175	229	177	206	212	232	314	351
ion secondary battery	10	25	15	23	32	58	68	73	75	81	162	261	331	409	428	468	388	337	300	357
power supply device	11	3	25	7	20	28	35	28	84	60	80	140	227	259	285	222	228	256	281	348
lithium ion secondary	12	29	18	30	33	63	72	74	76	84	164	270	366	418	442	506	379	342	291	345
current collector electrode	5	8	6	4	8	11	12	27	37	45	56	127	113	114	138	97	137	138	217	323
active material lithium	17	24	16	28	27	36	36	43	48	65	100	168	256	183	225	255	222	177	160	266
cathode active material	6	10	23	31	9	28	59	49	44	64	92	79	146	168	199	182	168	198	211	247
power supply system	7	20	25	18	27	21	32	52	47	59	82	150	187	162	165	145	166	185	218	241
energy storage system	0	7	0	8	2	11	6	25	35	32	46	93	80	116	121	106	138	165	200	222
electrode mixture layer	0	0	0	3	0	0	4	14	25	38	26	56	87	93	89	111	151	98	170	219
solid electrolyte layer	5	2	1	0	3	0	3	3	19	11	21	48	46	92	43	70	68	105	126	208
solid state battery	0	0	0	1	0	0	1	2	5	4	3	11	29	58	49	29	33	53	95	198
layer electrode active	3	1	2	3	3	5	6	17	12	20	30	60	71	93	85	112	102	105	120	200
battery management system	0	3	3	1	2	3	11	30	39	20	22	61	91	113	97	147	107	139	163	185
material layer electrode	2	0	1	3	3	11	15	19	9	30	34	76	79	89	82	86	74	101	91	178
energy storage unit	1	1	21	1	4	7	20	16	26	47	43	81	78	78	44	94	92	101	201	176
secondary battery lithium	6	15	4	10	8	13	10	13	21	25	40	46	69	69	90	96	106	70	125	177
anode active material	6	2	4	15	14	20	58	61	65	55	77	80	92	113	211	108	74	117	166	175
transition metal oxide	3	7	3	9	3	10	20	18	13	28	38	49	54	68	91	132	106	80	83	167
active material particle	4	6	12	6	14	19	41	32	32	36	38	57	94	93	93	150	161	213	199	164
collector electrode active	1	2	4	1	3	6	4	15	13	15	28	68	44	92	109	67	67	72	82	157
active material electrode	27	23	13	15	20	20	32	34	37	40	52	107	119	126	118	161	128	131	123	182
electrical energy storage	1	3	10	15	18	3	16	38	15	28	52	32	89	116	74	77	79	100	125	153
wireless power transmission	0	0	0	0	0	0	0	0	2	13	23	33	115	105	183	172	154	223	143	148
redox flow battery	0	0	8	12	2	1	0	0	0	2	2	22	44	30	50	93	86	102	120	146
power storage element	1	0	0	0	0	0	0	0	13	11	10	4	46	58	93	112	160	162	170	143
aqueous electrolyte solution	4	11	5	0	11	13	12	12	22	14	31	42	64	62	70	47	55	38	89	143
material electrode active	7	4	4	5	5	14	11	18	17	33	27	50	82	77	69	84	89	70	84	146
material lithium ion	5	10	3	13	5	17	12	18	23	33	63	106	162	134	153	181	142	125	126	140
power transmission device	0	0	0	2	0	0	0	0	2	30	27	41	53	92	121	173	96	113	136	134
lithium transition metal	8	16	6	19	24	31	20	19	36	29	42	74	75	82	106	138	112	87	112	138
battery module plurality	1	4	0	2	0	5	17	3	15	15	18	52	62	71	65	71	60	64	117	127
material lithium secondary	11	22	21	21	33	32	27	27	26	50	54	81	125	105	134	99	102	80	84	130
electric vehicle charging	0	0	0	0	0	0	1	0	5	9	25	48	93	113	65	53	38	50	64	118
battery cell electrode	0	1	4	1	1	2	2	8	2	11	14	24	30	59	42	43	53	47	72	117
battery electrode active	5	7	9	13	21	14	22	44	25	36	39	77	128	120	115	133	111	133	102	122
power supply circuit	10	12	27	14	11	25	12	44	38	26	31	59	71	59	73	37	63	101	100	127
control unit configured	0	0	0	0	0	2	1	2	6	6	9	13	34	34	42	46	50	62	61	114
state secondary battery	0	4	2	1	3	3	4	5	9	9	16	23	31	41	28	27	57	96	45	112
electrode lithium secondary	4	8	11	10	17	19	8	13	17	19	29	39	31	48	35	28	38	39	74	114

Figure 6. Occurrence counts of trigrams in battery patent abstracts.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
electrode active material	91	70	64	102	106	100	123	129	119	127	126	192	185	173	167	179	163	155	143	181
active material layer	17	8	26	23	29	61	58	67	54	67	55	71	62	74	63	50	43	60	64	80
energy storage device	5	35	24	33	26	13	23	33	44	33	44	36	33	49	46	36	47	64	60	59
lithium ion battery	15	22	36	26	12	31	30	39	40	43	58	62	70	56	57	67	79	67	73	62
electrode current collector	7	6	9	10	12	10	12	21	23	37	28	26	30	23	20	22	27	27	43	49
lithium secondary battery	47	78	64	55	87	82	98	65	50	63	66	66	67	65	62	53	63	44	56	77
plurality battery cell	3	1	3	1	5	6	5	13	10	10	9	26	26	31	22	27	27	27	32	32
power storage device	11	3	6	8	2	5	2	14	28	13	54	32	47	42	45	40	40	62	30	36
current collector electrode	5	7	5	3	6	7	6	13	15	16	17	25	17	15	17	13	17	16	22	29
secondary battery electrode	30	23	19	21	43	52	58	51	38	43	43	56	62	57	48	52	54	52	45	54
ion secondary battery	10	21	13	19	25	37	35	34	30	29	48	51	49	55	54	61	50	40	31	32
power supply device	11	3	22	6	16	18	18	13	34	21	24	27	34	35	36	29	29	30	29	32
energy storage system	0	6	0	7	2	7	3	12	14	11	14	18	12	16	15	14	18	20	21	20
electrode mixture layer	0	0	0	3	0	0	2	7	10	14	8	11	13	12	11	14	19	12	18	20
lithium ion secondary	12	25	16	25	26	40	37	34	31	30	49	52	54	56	56	65	48	41	30	31
solid state battery	0	0	0	1	0	0	1	1	2	1	1	2	4	8	6	4	4	6	10	18
battery management system	0	3	3	1	2	2	6	14	16	7	7	12	13	15	12	19	14	16	17	17
cathode active material	6	9	20	26	7	18	30	23	18	23	27	15	22	23	25	24	21	23	22	22
layer electrode active	3	1	2	3	2	3	3	8	5	7	9	12	11	12	11	14	13	12	12	18
energy storage unit	1	1	19	1	3	4	10	7	10	17	13	16	12	10	6	12	12	12	21	16
power supply system	7	17	22	15	21	13	16	24	19	21	24	29	28	22	21	19	21	22	23	22
material layer electrode	2	0	1	3	2	7	8	9	4	11	10	15	12	12	10	11	9	12	9	16
solid electrolyte layer	5	2	1	0	2	0	2	1	8	4	6	9	7	12	5	9	9	12	13	19
wireless power transmission	0	0	0	0	0	0	0	0	1	5	7	6	17	14	23	22	20	26	15	13
redox flow battery	0	0	7	10	2	1	0	0	0	1	1	4	7	4	6	12	11	12	12	13
collector electrode active	1	2	4	1	2	4	2	7	5	5	8	13	7	12	14	9	9	9	8	14
electrical energy storage	1	3	9	13	14	2	8	18	6	10	15	6	13	16	9	10	10	12	13	14
power transmission device	0	0	0	2	0	0	0	0	1	11	8	8	8	12	15	22	12	13	14	12
transition metal oxide	3	6	3	8	2	6	10	8	5	10	11	10	8	9	11	17	14	9	9	15
power storage element	1	0	0	0	0	0	0	0	5	4	3	1	7	8	12	14	20	19	18	13
electrolyte secondary battery	26	23	36	30	50	47	56	55	73	46	33	42	45	36	55	55	53	34	27	37
aqueous electrolyte secondary	25	33	40	39	52	48	57	54	73	48	36	42	45	39	57	55	54	34	26	35
active material particle	4	5	11	5	11	12	21	15	13	13	11	11	14	12	12	19	21	25	21	15
electric vehicle charging	0	0	0	0	0	0	1	0	2	3	7	9	14	15	8	7	5	6	7	11
battery cell electrode	0	1	4	1	1	1	1	4	1	4	4	5	4	8	5	6	7	6	7	11
battery module plurality	1	3	0	2	0	3	9	1	6	5	5	10	9	10	8	9	8	8	12	12
control unit configured	0	0	0	0	0	1	1	1	2	2	3	3	5	5	5	6	6	7	6	10
state secondary battery	0	3	2	1	2	2	2	2	4	3	5	4	5	5	4	3	7	11	5	10
secondary battery lithium	6	13	4	8	6	8	5	6	8	9	12	9	10	9	11	12	14	8	13	16
anode active material	6	2	4	13	11	13	30	28	26	20	23	16	14	15	27	14	9	14	17	16
current collector layer	0	0	1	2	2	1	2	0	0	0	1	1	1	5	2	1	2	4	10	9
aqueous electrolyte solution	4	9	4	0	9	8	6	6	9	5	9	8	9	8	9	6	7	5	9	13
unmanned aerial vehicle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	4	13	16	8
electrode active substance	1	2	9	0	14	10	3	1	1	2	4	4	7	6	11	7	8	4	7	9
plurality battery module	1	4	1	6	1	2	4	6	1	4	3	10	7	7	6	4	8	5	11	9
present electrode active	0	2	2	2	1	1	1	2	1	2	2	7	5	3	7	7	9	8	7	8
solid state secondary	0	0	0	0	0	0	1	1	0	0	1	2	1	2	1	1	6	8	3	8
power receiving device	0	0	0	0	2	3	0	3	1	12	6	8	11	16	22	14	10	10	12	8
solid electrolyte material	0	0	0	0	0	0	0	0	2	0	2	3	4	5	3	1	2	4	4	8
power storage system	0	0	1	0	0	1	0	4	0	1	1	4	7	7	10	6	9	4	8	8

Figure 7. Occurrence intensities of trigrams in battery patent abstracts. The unit of measure is occurrences per 1,000 abstracts and all values are rounded to the closest integer.

- Cluster 1–"Technological forefront 1":

  These countries are putting an increased focus on the two emerging technologies of solid-state and redox flow batteries. Their patent output related to lead-acid batteries is the lowest out of the three clusters and their Sodium-ion-related IPF share is close to zero. This cluster contains high-tech industrial nations like the USA, Germany, and Taiwan that are known to have explicitly expressed their ambitions in the field of battery technology.

- Cluster 2–"Laggard countries":

  A large portion of these countries' battery innovation results are still made up of lead-acid battery patents. Their share of battery patents related to the four analyzed emerging technologies are close to zero, except for their Lithium-Sulphur component, which accounts for approximately 10% of their IPF output in 2010-2019. This cluster contains countries like India, Russia, and Turkey that are considerably industrialized but are not known for their innovative impact in the worlds technology sector.

- Cluster 3–"Technological forefront 2":

  These countries' focus lays on lead-acid and Lithium-Sulphur batteries, which accounts for about 30% each. They have almost no innovative output in solid-state batteries and exhibit a considerably greater share in Sodium-ion batteries than the other two clusters. This cluster is made up of countries like Canada, Spain, and China; countries that have a considerable economic impact (China especially), but have not (yet) acquired a reputation for their inventiveness in battery technology.

Any (more) discussion points regarding country counts, technologies, clustering, or abstract mining?

5 Methods and Data

The foundation of this study is the PATSTAT database[2] provided by the European Patent Office, more precisely the 2021 Autumn edition of PATSTAT Online. Transact-SQL or T-SQL is the language used for querying it. The query used for selecting and downloading the data used for this study is defined in the text file "PATSTAT_Online_query.txt", which is included in the GitHub repository associated with this work, which can be found by following this link:

`https://github.com/ph1001/battery_patents.git`.

The patents that were downloaded from PATSTAT and that made up the raw dataset for this study were all patent applications that are part of patent families whose intra-family value for the feature "earliest publication date" lies in the time frame of 1999-2019 (the timeframe was later reduced to 2000-2019) and who contain at least one IPC entry matching one of the following codes:

- H01M... (Processes or means, e.g. batteries, for the direct conversion of chemical energy into electrical energy)
- H02J 3/32 (Circuit arrangements for AC mains or AC distribution networks using batteries with converting means)
- H02J 7... (Circuit arrangements for charging or depolarising batteries or for supplying loads from batteries)
- B60L 53... (Methods of charging batteries, specially adapted for electric vehicles; Charging stations or on-board charging equipment therefor; Exchange of energy storage elements in electric vehicles)

PATSTAT Online has the restriction that all SQL queries must begin with a "SELECT" statement, a fact that makes analyses of a higher complexity impossible to achieve inside PATSTAT Online itself. Consequently, data has to be queried, downloaded and then processed in a different environment. The programming language used for all steps after querying the database and downloading the data was Python (Version 3.9.7)[14], more precisely the web application Jupyter Notebook (Version 6.4.3)[15], the data processing libraries pandas (Version 1.3.3)[16] and Numpy (Version 1.20.3)[17], visualisation tools like Plotly (Version 5.1.0)[18] and Seaborn (Version 0.11.2)[19], the text mining suite Natural Language Toolkit (NLTK) (Version 3.6.5)[20] and the analytics toolboxes Scikit-learn (Version 0.24.2)[21] and SciPy (Version 1.7.1).[22]

The labor force counts used for scaling were downloaded from the world bank's website[23] and for the specific case of Taiwan from the website of "National Statistics: Republic of China (Taiwan)".[24]

5.1 Preprocessing

Preprocessing steps undertaken in order to reduce the raw data downloaded from PATSTAT to the final dataset used for this study and for making some sensible adjustments to it are defined in the Jupyter Notebook "create_dataset.ipynb", which is included in the GitHub repository linked above. The following paragraphs contain a summary of these preprocessing steps.

First, the raw data downloaded from PATSTAT Online was loaded and checked for its integrity. Then each patent family's earliest intra-family value for the feature "earliest publication date" was determined and added as a new column to every row of the dataset (i.e. it was harmonized on patent family level). Like this, patent families can easily be assigned to their respective year later during the analyses. Next, all patent families were classified and tagged as either "IPF", "singleton", or "-" (meaning "neither"). The resulting tags are stored in the newly created column "tag". Next, more tags for further data selection were created. This process took place in five steps, which are described in the following:

- First, every patent family was scanned for the IPC codes related to non-active battery parts, electrodes, or secondary cells (IPC codes H01M 2..., H01M 50..., H01M 4..., and H01M 10...). Patent families that contained any of these code were added in their entirety, except if they contained any of the IPC codes H01M 6..., H01M 8..., H01M 12..., H01M 14..., or H01M 16..., which are related to primary cells, fuel cells, hybrid cells, electrochemical current or voltage generators not provided for in groups H01M 6/00-H01M 12/00, and structural combinations of different types of electrochemical generators, which were hereby explicitly excluded from the analysis. The patent families passing this stage were tagged as "non active parts, electrodes, secondary cells".

- In a second step, every patent family was scanned for the IPC codes related to "circuit arrangements for ac mains or ac distribution networks using batteries with converting means" (H02J 3/32), "circuit arrangements for charging or depolarising batteries or for supplying loads from batteries" (H02J 7...), "methods of charging batteries, specially adapted for electric vehicles" (B60L 53...), or "secondary cells; methods for charging or discharging" (H01M 10/44). Patent families that contained any of these codes were added in their entirety, except if they contained any of the IPC codes listed for exception in the above step or any of the codes B60L 53/54, B60L 53/55, or B60L 53/56 that refer to charging stations using fuel cells, capacitors, or mechanical storage means, respectively. Patent families that passed this stage were tagged as "charging".

- As a third step, in order to identify affiliations of the resulting patent families to a set of technological categories, each patent family's titles and abstracts were scanned using individual sets of regular expressions for each technology. These regular expressions are defined in the Jupyter notebook "create_dataset.ipynb". Titles and abstracts of all languages were considered and a patent family was selected in its entirety if any substring of its titles or abstracts matched any of the respective regular expressions. Please note that—in order to decrease the risk of false positives—before scanning abstracts for these regular expressions, they were cut off at the beginning of any appearance of the string "independent claims are also included for". The selected patent families were one-hot tagged in the newly created columns with the column name "is x", with $x \in$ {Lead-acid, Lithium-air, Lithium-ion, Lithium-Sulphur, Other Lithium, Magnesium-ion, Nickel-cadmium, Nickel-iron, Nickel-zinc, Nickel-metal hydride, Rechargeable alkaline, Sodium-Sulphur, Sodium-ion, Solid-state, Aluminium-ion, Calcium(-ion), Organic radical} being the name of the respective technology. Please note that due to the considerable overlap of the concept of solid-state batteries with other technologies, especially Lithium-ion batteries, all patent families that were classified as patents related to solid-state batteries were untagged in any other category in which they acquired tags through the process described here. To be very clear: This especially means that the Lithium-ion battery category does not contain any patent families that are tagged as solid-state battery inventions.

- The fourth step's purpose was to add patent data related to redox flow and Nickel–hydrogen batteries to the dataset. For this purpose, a combination of IPC classes queries and text queries was deployed. Redox flow and Nickel–hydrogen batteries are closely related to fuel cells and, consequently, patents associated with them are often included in IPC classes that were excluded by the above steps. Analogous to the above steps, the IPC classes qualifying for potential inclusion were H01M 2..., H01M 50..., H01M 4..., H01M 8..., and H01M 10... and the IPC classes demanding exclusion were H01M 6..., H01M 12..., H01M 14..., and H01M 16.... Analogous to the above step, these patent families' titles and abstract were then scanned using one set of regular expressions for redox flow and another for Nickel–hydrogen batteries. These regular expressions can be reviewed in the Jupyter notebook "create_dataset.ipynb". All patent families that passed this stage were one-hot tagged in the newly created columns with the names "is Redox flow" or "is Nickel–hydrogen", respectively.

- As a last step, another additional column was computed: The dataset column "technologies one hot sum" contains the sum across each row's "is <technology name>" values. This sum is needed in the rare cases where technology classifications overlap. The share of patent families that had more than one technology associated to them was 0.61% in the final dataset.

The counts resulting from these overlapping technologies were not counted multiple times but, using the respective "technologies one hot sum" value, distributed as equal fractions across the overlapping classes.

The tags created in the above steps were used for selecting the appropriate data for each analysis. As already hinted earlier in this paper, all patent families not having the "IPF" tag were filtered out before all analyses. That the rest was kept in the unfiltered dataset was only for completeness, having the potential of future analyses with a broader scope in mind. The data selection method that was applied before each analysis that is based on the labels whose creation was described above is presented in Fig. 8.

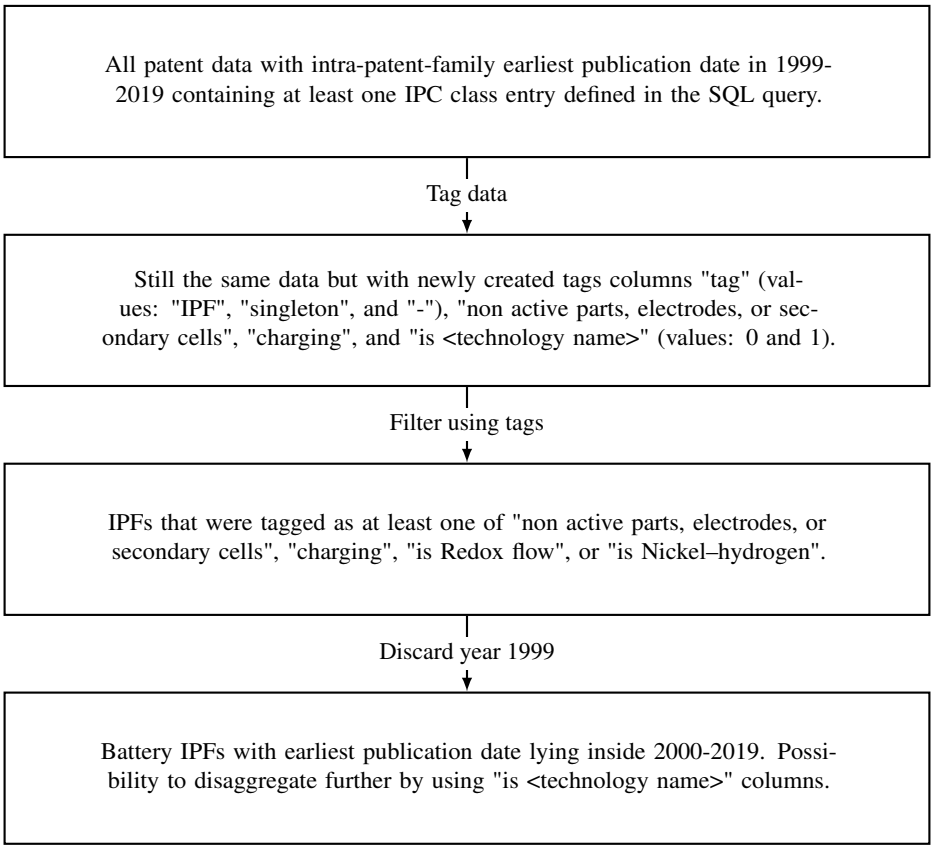

Figure 8. Flow chart depicting the data selection process for this study.

5.2 Counting patents

As already mentioned in the introduction of this work, the methodological setup of this study roughly follows the framework defined in the report by IEA and EPO[1]. This means that all dates in this study refer to the earliest publication date within the respective IPF and that the geographic distributions were calculated based on the geographic information assigned to the respective inventors in PATSTAT and where multiple inventors are indicated, each inventor was assigned an equal fraction of the respective count. In the eyes of the authors of this study, there is a limitation to this approach, which is described in the following: For identifying the inventors, their PATSTAT name attribute "psn_name" is used. The harmonization of this feature, which was carried out by PATSTAT, is not complete. For example, pairs of entries like "SHIH, I-FEN" and "Shih, I-Fen" exist, which in reality correspond to the same inventor, but consequently are treated as two different individuals. This shifts the fractions of countries of origin in these entries' patent family in favor of the country of the unharmonized name.

Figures 2 and 3 show the top countries in terms of their sum of the depicted values. The code used for counting patents by countries is contained in the Jupyter Notebook "counts_technologies_clustering.ipynb", which is part of the GitHub repository linked in the beginning of this chapter.

5.3 Methods: Battery technologies

Different to the report by EPO and IEA[1], in the current study fractional counting was also applied when breaking down counts by technological categories. Whenever an IPF was classified to belong to more than one category, each technology was assigned

an equal fraction of the respective count. As already mentioned above, this only happened in a very small minority of the cases since only 0.61% of all IPFs were assigned to more than one technology. The code used for counting patents by technologies is contained in the Jupyter Notebook "counts_technologies_clustering.ipynb", which is part of the GitHub repository linked in the beginning of this chapter.

5.4 Methods: Clustering

The metric R^2 applied for comparing the performance of several clustering algorithms using varying numbers of clusters can be characterized as follows (please note that said comparison was conducted on a non-varying dataset; a derivation of the relation below is provided for example by N. E. Helwig[25], page 4):

$$R^2 = \frac{SSB}{SST} = \frac{SST - SSW}{SST} = 1 - \frac{SSW}{SST} \in [0,1] \tag{1}$$

where

$$SSB = \sum_{i=1}^{p} n_i (\overline{X}_i - \overline{X})^2 = sum\ of\ squares\ between\ groups \tag{2}$$

and

$$SST = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2 = total\ sum\ of\ squares \tag{3}$$

and

$$SSW = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (X_{ij} - \overline{X_i})^2 = sum\ of\ squares\ within\ groups \tag{4}$$

with

$p = number\ of\ clusters,$

$n_i = number\ of\ elements\ in\ cluster\ i,$

$\overline{X_i} = centroid\ of\ cluster\ i,$

$\overline{X} = center\ of\ whole\ dataset,\ and$

$X_{ij} = jth\ element\ of\ cluster\ i.$

A higher R^2 value indicates a better clustering solution. Clustering algorithms that were compared are k-means and hierarchical agglomerative clustering using complete, average, single, and Ward linkage and the numbers of clusters that were tested ranged from two to nine.

The decision to use only the five dimensions "Lead-acid", "Redox flow", "Solid-state", "Sodium-ion", and "Lithium-Sulphur" resulted from extensive testing of other configurations, especially those that included "Lithium-ion", "Other Lithium" or a joint category of "Lithium-ion and other Lithium". These tests were not found to be satisfying, since it was observed that the Lithium-related IPFs were drowning out the other categories due to their sheer amount, resulting in clustering solutions that lacked the clear interpretability of the solution presented in this work. Lithium air batteries, another battery technology that has received increased attention in the recent years[4], was considered as a candidate feature for this analysis, but was discarded due to its still very low yearly IPF counts. The code used for clustering countries based on their technology distribution is contained in the Jupyter Notebook "counts_technologies_clustering.ipynb", which is part of the GitHub repository linked in the beginning of this chapter.

5.5 Methods: Title and abstract mining

Unigrams, bigrams, and trigrams were extracted from cleaned titles and abstracts from which meaningless words and phrases had been removed and in which certain synonymities and anomalies had been treated. The n-gram counts method simply counts occurrences and displays them as yearly sums whilst the n-gram intensities method does the same with the difference that its resulting values are scaled using each years' numbers of titles or abstracts, respectively. Three ways for presenting the identified n-grams were designed for this study:

- Method 1a:

  Sorted in descending order of increase over the given time frame of 2000-2019 with the measure used for sorting being $m_1 = count_{last} - count_{first}$.

- Method 1b:

  Sorted in ascending order of increase over the given time frame of 2000-2019 with the measure used for sorting being m_1. This method's purpose is to show n-grams that exhibit a negative increase, i.e. have decreased over the given time period.

- Method 2:

  Sorted in descending order with the measure used for sorting being $m_2 = \sum abs(year-to-year\ difference_{i,i+1})$. This method's purpose is to show n-grams whose count or intensity changed the most (in absolute terms) between all adjacent years.

The results displayed in Fig. 6 and Fig. 7 were obtained using method 1a, patent abstracts, and trigrams. The code for computing these results is contained in the Jupyter Notebook "title_and_abstract_mining.ipynb", which is part of the GitHub repository linked in the beginning of this chapter. The results obtained by using the methods and data combinations not presented in this paper can best be viewed by opening the HTML file "title_and_abstract_mining.html", which is also available in the same folder. The combinations for which results were computed can be characterized by the Cartesian product $c = \{n = 1,\ n = 2,\ n = 3\} \times \{n-gram\ counts,\ n-gram\ intensities\} \times \{method\ 1a,\ method\ 1b,\ method\ 2\} \times \{titles,\ abstracts\}$.

Data Availability

The raw data used for this study is available via subscription at PATSTAT Online. The data subset generated and used for this work can be reproduced using the queries and code made available at the GitHub repository associated with this work, which can be found by following this link: `https://github.com/ph1001/battery_patents.git`. On reasonable request said data subset is available from the corresponding author.

Acknowledgements (not compulsory)

... Acknowledgements ...

Author contributions statement

Might be subject to change.

Philipp Metzger designed the queries used for data selection, preprocessed and analyzed the data, prepared all figures, and wrote the manuscript text. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

1. IEA. Innovation in batteries and electricity storage, a global analysis based on patent data (2020).
2. De Rassenfosse, G., Dernis, H. & Boedt, G. An introduction to the patstat database with example queries. *Aust. Econ. Rev.* **47**, DOI: 10.1111/1467-8462.12073 (2014).
3. Griliches, Z. Patent statistics as economic indicators: A survey. *J. Econ. Lit.* **28**, 1661–1707 (1990).
4. Aaldering, L. J. & Song, C. H. Tracing the technological development trajectory in post-lithium-ion battery technologies: A patent-based approach. *J. Clean. Prod.* **241**, 118343, DOI: https://doi.org/10.1016/j.jclepro.2019.118343 (2019).

5. Malhotra, A., Zhang, H., Beuse, M. & Schmidt, T. How do new use environments influence a technology's knowledge trajectory? a patent citation network analysis of lithium-ion battery technology. *Res. Policy* **50**, 104318, DOI: https://doi.org/10.1016/j.respol.2021.104318 (2021).

6. Stephan, A., Bening, C. R., Schmidt, T. S., Schwarz, M. & Hoffmann, V. H. The role of inter-sectoral knowledge spillovers in technological innovations: The case of lithium-ion batteries. *Technol. Forecast. Soc. Chang.* **148**, 119718, DOI: https://doi.org/10.1016/j.techfore.2019.119718 (2019).

7. OECD. *OECD Patent Statistics Manual* (OECD PUBLICATIONS, 2, rue André-Pascal, 75775 PARIS CEDEX 16, 2009).

8. Dechezleprêtre, A., Ménière, Y. & Mohnen, M. International patent families: from application strategies to statistical indicators. *Scientometrics* **111**, 793–828, DOI: 10.1007/s11192-017-2311-4 (2017).

9. Schmoch, U. & Gehrke, B. China's technological performance as reflected in patents. *Scientometrics* **127**, 299–317, DOI: 10.1007/s11192-021-04193-6 (2022).

10. Schmoch, U. & Khan, M. Methodological challenges for creating accurate patent indicators. *Springer Handb. Sci. Technol. Indic.* 907–927, DOI: 10.1007/978-3-030-02511-3_37 (2019). Springer International Publishing.

11. Frietsch, R. & Kroll, H. China's foreign technology market perspectives. *Deutsch-chinesische Innov. Rahmenbedingungen, Chancen und Herausforderungen* pp. 365–375 (2020). Metropolis.

12. Vezzini, A. 15 - lithium-ion battery management. In Pistoia, G. (ed.) *Lithium-Ion Batteries*, 345–360, DOI: https://doi.org/10.1016/B978-0-444-59513-3.00015-7 (Elsevier, Amsterdam, 2014).

13. Neuhäusler, P., Rothengatter, O. & Frietsch, R. Patent applications - structures, trends and recent developments 2018. Studien zum deutschen Innovationssystem 4-2019, Leibniz Information Centre for Economics - ZBW, Berlin (2019).

14. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

15. Kluyver, T. *et al.* Jupyter notebooks – a publishing format for reproducible computational workflows. In Loizides, F. & Schmidt, B. (eds.) *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87 – 90 (IOS Press, 2016).

16. Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (eds.) *Proceedings of the 9th Python in Science Conference*, 56 – 61, DOI: 10.25080/Majora-92bf1922-00a (2010).

17. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362, DOI: 10.1038/s41586-020-2649-2 (2020).

18. Inc., P. T. Collaborative data science (2015).

19. Waskom, M. *et al.* mwaskom/seaborn: v0.11.2 (august 2021), DOI: 10.5281/zenodo.5205191 (2021).

20. Bird, S., Klein, E. & Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit* (" O'Reilly Media, Inc.", 2009).

21. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

22. Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).

23. The World Bank: Labor force, total. https://data.worldbank.org/indicator/SL.TLF.TOTL.IN (2022). Accessed: 10. Jan. 2022, 12:07.

24. National Statistics, Republic of China (Taiwan): Labor force, total. https://eng.stat.gov.tw/ct.asp?xItem=42761&ctNode=1609&mp=5 (2022). Accessed: 10. Jan. 2022, 12:49.

25. Helwig, N. E. One-way Analysis of Variance. http://users.stat.umn.edu/~helwig/notes/OneWayANOVA.pdf (2020). Accessed: 20. Jan. 2022, 11:18.