Accepted for publication in the
Collnet Journal of Scientometrics and Information Management

Identifying seminal works most important for research fields:

Software for the Reference Publication Year Spectroscopy (RPYS)

Lutz Bornmann *[a], Andreas Thor[b], Werner Marx[c], and Loet Leydesdorff[d]

* Corresponding author

[a] Division for Science and Innovation Studies
Administrative Headquarters of the Max Planck Society
Hofgartenstr. 8,
80539 Munich, Germany.
E-mail: bornmann@gv.mpg.de

[b] University of Applied Sciences for Telecommunications Leipzig
Gustav-Freytag-Str. 43-45,
04277 Leipzig, Germany.
Email: thor@hft-leipzig.de

[c] Max Planck Institute for Solid State Research
Information Service
Heisenbergstrasse 1,
70506 Stuttgart, Germany.
Email: w.marx@fkf.mpg.de

[d] Amsterdam School of Communication Research (ASCoR),
University of Amsterdam,
PO Box 15793, 1001 NG Amsterdam, The Netherlands.
Email: loet@leydesdorff.net

Abstract

Reference Publication Year Spectroscopy (RPYS) was proposed by Marx, Bornmann, Barth, and Leydesdorff (2014) to identify seminal publications in a research field which are most important in a historical context. We refined our RPYS toolbox by adding some features to the existing programs and we developed two new routines. First, a direct comparison of the results of different RPYSs is now possible, because the software transforms the results of the RPYS into percentiles for standardization. Second, we added routines that facilitate the user with retrieving the most-cited publications in specific years indicated by peaks in the spectrograms. Cited references can be aggregated across misspellings and variants. For this paper, two examples from the humanities and natural sciences are provided to demonstrate the functionalities and results of the programs. A more technical description of the usage of the programs can be found at http://www.leydesdorff.net/software/rpys/.

Key words

Cited references; Historical roots; Reference publication year spectroscopy; RPYS

1 Introduction

In the current atmosphere of research evaluation on all levels of scientific activities (single researchers, research groups, institutions, and countries), bibliometrics can be considered as an instrument for evaluation purposes (Pendlebury, 2008). For example, bibliometric data is one of the most important data sources for university rankings: The Leiden Ranking (Waltman et al., 2012) is solely based on bibliometric data and for the Academic Ranking of World Universities (http://www.shanghairanking.com) and the Times Higher Education World University Rankings (http://www.timeshighereducation.co.uk/world-university-rankings) this data plays an important role. The focus on evaluation purposes in the use of bibliometrics may close one's eyes to other possibilities of using this data. A good example for another use is the VOSViewer tool (http://www.vosviewer.com), which allows visualizations of scientific activities beyond evaluation (citation relations among scientific fields, university profiles and collaborations, co-citations of journals, etc.).

In this study, we would like to present significantly improved software for another tool which can be used to examine the seminal works or historical roots, respectively, of scientific fields. The method for examining historical roots has been developed on the base of a proposal of Bornmann and Marx (2013) to conduct cited reference analysis (Garfield, Pudovkin, & Istomin, 2003; Garfield, Sher, & Torpie, 1964; Leydesdorff, 2010). Bornmann and Marx (2013) argue for broadening the perspective in bibliometrics by complementing the (standard) times cited with cited reference analyses using field-specific impact measurements. For such a cited reference analysis, they propose to extract all cited references from a field-specific publication set and to analyze which papers (scientists or journals) have been cited most often and in which years.

A specific application of cited reference analysis is Reference Publication Year Spectroscopy (RPYS), which was introduced by Marx, et al. (2014): "RPYS is based on the

analysis of the frequency with which references are cited in the publications of a specific research field in terms of the publication years of these cited references. The origins show up in the form of more or less pronounced peaks mostly caused by individual publications that are cited particularly frequently" (p. 751).

Recently, RPYS has been used to examine the historical roots in some research fields: Marx, et al. (2014) used research on graphene and on solar cells to illustrate how RPYS functions. Leydesdorff, Bornmann, Marx, and Milojevic (2014) investigated the historical origins of iMetrics (information metrics, bibliometrics, and scientometrics) in scholarly literature. For example, they found that Lotka (1926) can be considered as the first source, but the intellectual program of iMetrics was especially shaped in the early 1960s. Whereas Barth, Marx, Bornmann, and Mutz (2014) examined the origins of the Higgs boson research and combined RPYS with a segmented regression analysis, Wray and Bornmann (2014) took a closer look at the roots of the philosophy of science. As the results of Marx and Bornmann (2014) show, RPYS can not only be applied to the identification of origins, but also to reveal scientific legends: "Charles Darwin, the originator of evolutionary theory, was given credit for finches he did not see and for observations and insights about the finches he never made" (p. 839). The analysis validated bibliometrically the known fact that a book from 1947 is the origin of the term "Darwin finches" (Lack, 1947).

Most RPYS papers are based on software which can be downloaded at http://www.leydesdorff.net/software/rpys. Recently, Comins and Hussey (2015) used RPYS to investigate the impact of Viterbi algorithm first published by Andrew Viterbi in 1967. They extended the method of RPYS with heat maps with the goal of comparing the results of different RPYS. A comparison is only possible when the results are standardized. In the present paper, we present new RPYS software that uses percentiles for this standardization. In our opinion, this transformation into percentiles improves on the rank transformation proposed by Comins and Hussey (2015). Furthermore, we add routines that facilitate the user

with retrieving the most-cited papers in specific years indicated by peaks in the spectrogram. Cited references can be aggregated across misspellings and variants.

2 Methods

2.1 Datasets used

As the first example in this study, we use the dataset of Wray and Bornmann (2014). They investigated the origins of the philosophy of science field. Their study is based on papers published in the journals *Philosophy of Science*, *British Journal for the Philosophy of Science*, *Studies in History and Philosophy of Science*, and *Erkenntnis* (n=8,757 records). Since the data comes from four journals, it can be used to compare the impact of important papers for the field published in these journals. For example, it can be revealed whether Thomas Kuhn's *Structure of Scientific Revolutions* (Kuhn, 1962) has the same or different impact in these journals in terms of citation. The relevant number of papers and cited references are provided in Table 1.

Table 1. Number of papers and number of cited references in four journals relevant for the philosophy of science (date of searching: March 2013).

Journal	Number of papers	Number of cited references
British Journal for the Philosophy of Science (since 1956)	2,814	35,985
Erkenntnis (since 2000)	714	15,973
Philosophy of Science (since 1956)	3,833	54,683
Studies in History and Philosophy of Science (since 1974)	1,396	49,919
Total	8,757	156,560

A second example deals with four research topics within the field of the natural sciences. The analysis is based on papers (n=18,451 records) investigating light scattering at small particles of four different materials: atmospheric aerosols, cosmic dust, colloids, and

nanoparticles. Here, the data can be used to compare the impact of publications shared among these research topics and thus indicated as most important at the level of the field. The corresponding numbers of papers and cited references are provided in Table 2.

Table 2. Number of papers and number of cited references for four research topics within the field of the natural sciences. The corresponding Web of Science search queries are given in parentheses (date of searching: March 2015).

Research topic	Number of papers	Number of cited references
Aerosols (TS=aerosol* AND (TS=atmospher* OR air) AND TS=scatter*)	5,315	198,473
Cosmic Dust (WC=astronomy AND TS=dust AND TS=scatter*)	3,109	144,199
Colloids (TS=colloid* AND TS="light scatter*"	5,817	240,090
Nanoparticles (TS=nanoparticle* AND TS="optical propert* AND TS=scatter*)	4,210	197,245
Total	18,451	780,007

The search was undertaken in the Web of Science (WoS), date of searching: October 2013 (first example) and March 2015 (second example).

2.2 Software

The programs rpys.exe, yearcr.exe, and RefMatchCluster.jar can be used to generate a RPYS of any document set downloaded from WoS. The procedure for how to use the routine is described in detail at http://www.leydesdorff.net/software/rpys/.

3 Results

3.1 First example

The starting point of a RPYS is the publication set representing a specific research field. In this study we use as a first example four journals in the philosophy of science. The

cited references of the publications in these sets are analyzed in terms of how often they have been cited. The results of the RPYS analyses for each journal are shown in Figure 1.

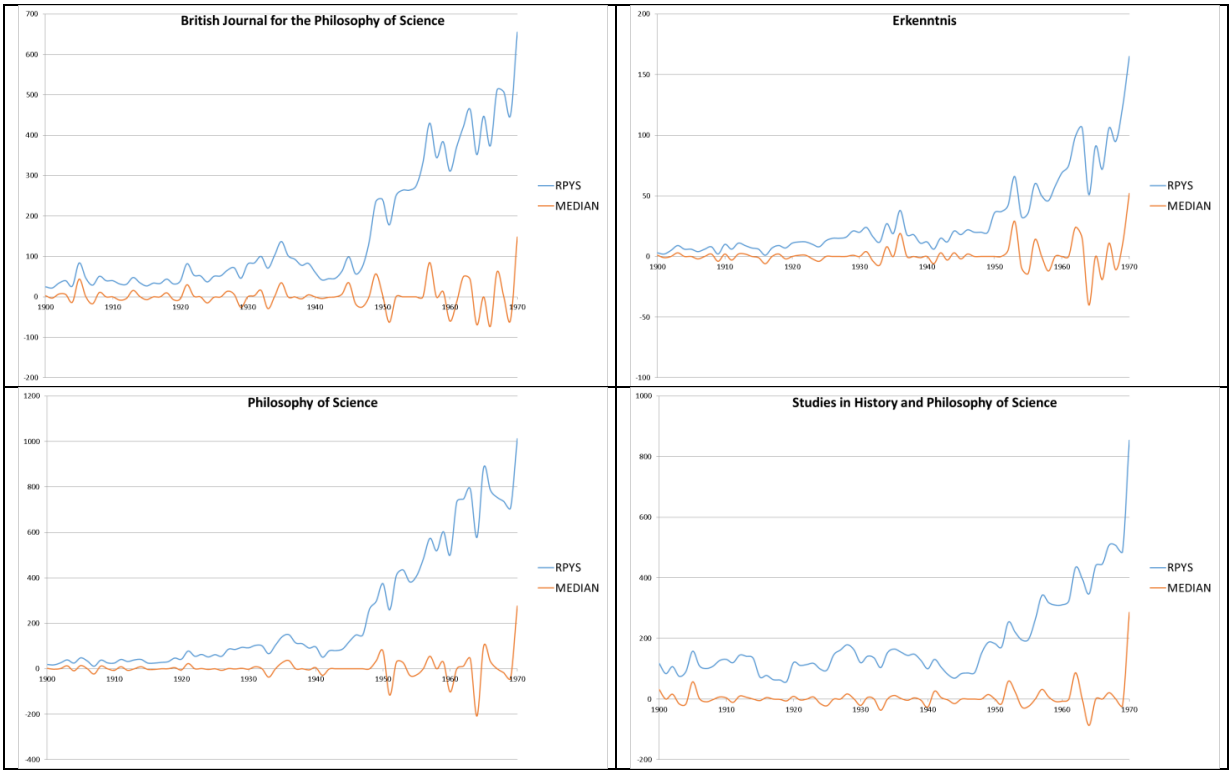
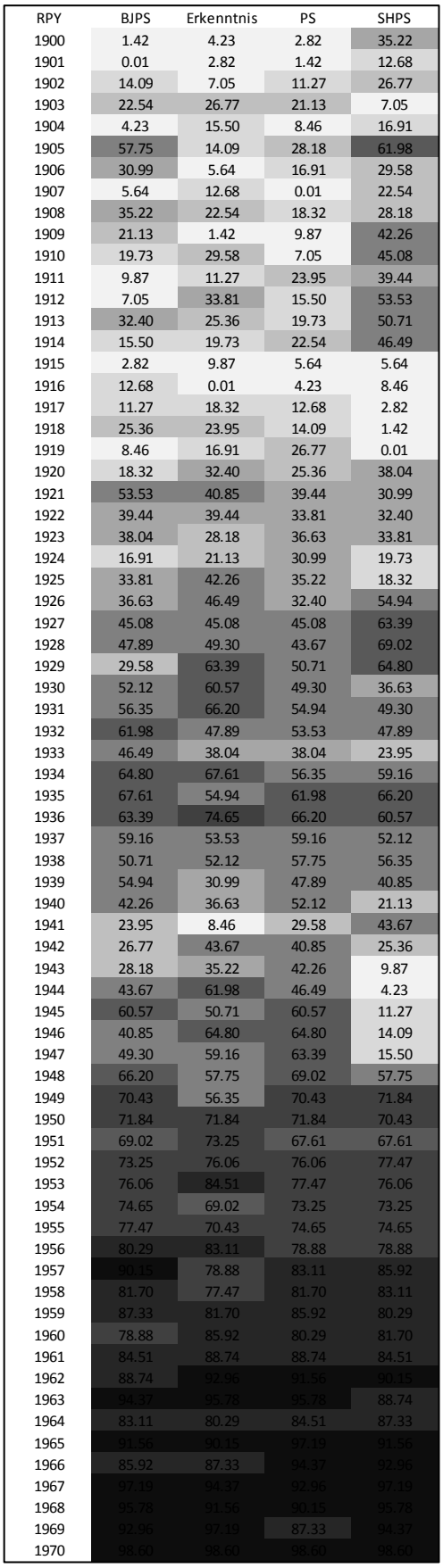


Figure 1. Results of the Reference Publication Year Spectroscopy for four journals in the philosophy of science field

Each graph in the figure visualizes the number of cited references per referenced publication year (blue line "RPYS") in a journal. In order to identify those publication years with significantly more cited references than other years, the deviation of the number of cited references in each year from the median of the number of cited references in the two previous, the current, and the two following years ($t - 2$; $t - 1$; t ; $t + 1$; $t + 2$) is visualized, too (brown line "Median"). This deviation from the five-year median provides a curve smoother than the one in terms of absolute numbers.

For the complete philosophy of science publication set, Wray and Bornmann (2014) identified the highest peaks in 1905, 1949, 1950, 1957,1962, 1965, and 1970. As the results in Figure 1 shows these peaks are not equally visible in the graphs of the different journals. For

example, the 1905 peak is visible in the *British Journal for the Philosophy of Science* and *Studies in History and Philosophy of Science*, but not in *Erkenntnis* and *Philosophy of Science*. In other words, the visual inspection is not very satisfying as a methodology for comparing the results for the four journals. For this reason, Comins and Hussey (2015) proposed a rank-transformation procedure: "To overcome the difficulty in making cross RPYS comparisons, we rank-transformed our results. To rank transform data, one takes the complete set of n values from a given RPYS analyses, say X_1, X_2, …, X_n. These observations are then sorted by the magnitude of the observed values. These values are then substituted in such a manner that the largest value takes the value *n*, the second largest take the value *n*-1, and so on." Using this rank-transformation procedure, however, a comparison would no longer be possible in the case of different numbers of years. For this reason, we recommend the transformation of the values into quantiles (percentiles) in order to make results with different publication years comparable.



RPY	BJPS	Erkenntnis	PS	SHPS
1900	1.42	4.23	2.82	35.22
1901	0.01	2.82	1.42	12.68
1902	14.09	7.05	11.27	26.77
1903	22.54	26.77	21.13	7.05
1904	4.23	15.50	8.46	16.91
1905	57.75	14.09	28.18	61.98
1906	30.99	5.64	16.91	29.58
1907	5.64	12.68	0.01	22.54
1908	35.22	22.54	18.32	28.18
1909	21.13	1.42	9.87	42.26
1910	19.73	29.58	7.05	45.08
1911	9.87	11.27	23.95	39.44
1912	7.05	33.81	15.50	53.53
1913	32.40	25.36	19.73	50.71
1914	15.50	19.73	22.54	46.49
1915	2.82	9.87	5.64	5.64
1916	12.68	0.01	4.23	8.46
1917	11.27	18.32	12.68	2.82
1918	25.36	23.95	14.09	1.42
1919	8.46	16.91	26.77	0.01
1920	18.32	32.40	25.36	38.04
1921	53.53	40.85	39.44	30.99
1922	39.44	39.44	33.81	32.40
1923	38.04	28.18	36.63	33.81
1924	16.91	21.13	30.99	19.73
1925	33.81	42.26	35.22	18.32
1926	36.63	46.49	32.40	54.94
1927	45.08	45.08	45.08	63.39
1928	47.89	49.30	43.67	69.02
1929	29.58	63.39	50.71	64.80
1930	52.12	60.57	49.30	36.63
1931	56.35	66.20	54.94	49.30
1932	61.98	47.89	53.53	47.89
1933	46.49	38.04	38.04	23.95
1934	64.80	67.61	56.35	59.16
1935	67.61	54.94	61.98	66.20
1936	63.39	74.65	66.20	60.57
1937	59.16	53.53	59.16	52.12
1938	50.71	52.12	57.75	56.35
1939	54.94	30.99	47.89	40.85
1940	42.26	36.63	52.12	21.13
1941	23.95	8.46	29.58	43.67
1942	26.77	43.67	40.85	25.36
1943	28.18	35.22	42.26	9.87
1944	43.67	61.98	46.49	4.23
1945	60.57	50.71	60.57	11.27
1946	40.85	64.80	64.80	14.09
1947	49.30	59.16	63.39	15.50
1948	66.20	57.75	69.02	57.75
1949	70.43	56.35	70.43	71.84
1950	71.84	71.84	71.84	70.43
1951	69.02	73.25	67.61	67.61
1952	73.25	76.06	76.06	77.47
1953	76.06	84.51	77.47	76.06
1954	74.65	69.02	73.25	73.25
1955	77.47	70.43	74.65	74.65
1956	80.29	83.11	78.88	78.88
1957	90.15	78.88	83.11	85.92
1958	81.70	77.47	81.70	83.11
1959	87.33	81.70	85.92	80.29
1960	78.88	85.92	80.29	81.70
1961	84.51	88.74	88.74	84.51
1962	88.74	92.96	91.56	90.15
1963	94.37	95.78	95.78	88.74
1964	83.11	80.29	84.51	87.33
1965	91.56	90.15	97.19	91.56
1966	85.92	87.33	94.37	92.96
1967	97.19	94.37	92.96	97.19
1968	95.78	91.56	90.15	95.78
1969	92.96	97.19	87.33	94.37
1970	98.60	98.60	98.60	98.60

Figure 2. Quantiles of the number of cited references for four philosophy-of-science journals: *British Journal for the Philosophy of Science* (BJPS), *Erkenntnis*, *Philosophy of Science* (PS), and *Studies in History and Philosophy of Science* (SHPS)

There are different methods available for the calculation of quantiles (Bornmann, Leydesdorff, & Mutz, 2013). We use the method proposed by Hazen (1914, p. 1550), because it accounts for the uncertainty in small sets (Leydesdorff, 2012) and is used very frequently nowadays. The formula for the transformation is $((i-0.5)/n * 100)$, whereby i is the rank (all years are ranked in decreasing order by their number of cited references) and n the total number of years. The resulting quantiles are comparable across results produced on the base of different numbers of referenced publication years. The quantile values are provided in an additional column of the file "median.dbf" generated by "rpys.exe".

Comins and Hussey (2015) proposed to visualize a comparison of RPYS using heatmaps. We followed their recommendation and produced Figure 2 based on quantiles. The higher the quantile for a certain journal and publication year, respectively, the darker is the corresponding cell. For example, the results shown in Figure 2 validate the results of visual inspection of Figure 1: the 1905 peak is pronounced for the *British Journal for the Philosophy of Science* and *Studies in History and Philosophy of Science*, but not for *Erkenntnis* and *Philosophy of Science.*

After the most important years for a research field have been identified, the most important publications in these years can be identified in a second step of analysis. These most important publications are often those that bring about the peaks. Especially in years of early science, single publications may produce pronounced peaks. Here, two programs can be used to identify the most important publications: (1) The program "yearcr.exe" produces "yearcr.dbf" in which the cited references are listed for specific years (column RPY) by their number of occurrences (column "N_CR"). In two further columns, the percent of a specific cited reference's number of occurrences as a percentage of the total number of all cited references' occurrences within that year (column "PERC_YR") and across all years (column "PERC_ALL") are provided. As a further information, the column "RPY" contains the publication years of the cited references. (2) Since many cited references appear with variants

in the list of cited references in “yearcr.dbf”, we wrote the program RefMatchCluster.jar in java. This program is able to identify, unify and aggregate cited references data.

Using these two programs, one can identify those publications in the *British Journal for the Philosophy of Science*, *Studies in History and Philosophy of Science*, *Erkenntnis*, and *Philosophy of Science*, which are the most frequently cited references. As a first step, “yearcr.exe” is used to produce “yearcr.dbf” with aggregated cited references information for the RPYs 1900 to 1970. As the alphabetical sorting of the cited references show many cited references appear with several variants in “yearcr.dbf.” For example, “kuhn ts, 1970, structure sci revolu, p102” and “kuhn ts, 1970, structure sci revolu, p115” are listed among the cited references in the dataset. Both references, however, refer to the same cited publication and should be jointly counted.

One can use RefMatchCluster.jar in a second step of analysis and detect the variants of the same cited reference, cluster them, and aggregate their occurrences (number of cited references). Note that the program does not merge occurrences across publication years. Thus, a reference to “kuhn ts, 1962”—the first edition of *The Structure of Scientific Revolutions*—or a reference to a later edition are not included automatically. However, the user can make these adjustments manually and then run the program again. The program needs some arguments. For example, the results which are presented in the following are based on the following command line (see http://www.leydesdorff.net/software/rpys/refmatchcluster.txt):

java -jar RefMatchCluster.jar -input=yearcr.dbf
-matcher=journal_short,Levenshtein,0.75 -matcher=lastname,Levenshtein,0.75
-match=yearcr_match.csv -cluster=yearcr_cluster.dbf -aggregate=cleaned.csv

In this case, the Levenshtein similarity function is used to determine the similarity value between character strings in a 0 to 1 range (here we used: 0.75 as a threshold) for the

journal/book title and author name fields in order to produce the unified and aggregated cited-references data. In our manifold experiences, a similarity value of 0.75 results in useful aggregated results. The results of the unification and aggregation process based on the philosophy of science journals' datasets are shown in Table 3: For each journal the most cited publications are shown – over all publication years, within a publication year, and in 1905.

For example, of all cited references in the papers published in the *British Journal for the Philosophy of Science* the reference "popper k., 1959, logic sci discovery" accounts for 0.8%. If percentages within single publication years are calculated, "einstein a, 1905, ann phys-berlin, v17, p891" has been most-frequently cited with 38%.

Table 3. Top-three most frequently cited publications for *British Journal for the Philosophy of Science* (BJPS), *Erkenntnis*, *Philosophy of Science* (PS), and *Studies in History and Philosophy of Science* (SHPS). The table shows the most frequently cited publications over all publication years (1900-1970), within a single publication year and in 1905. Subsequent to each cited reference the corresponding percentage is mentioned.

	Journal			
	BJPS	*Erkenntnis*	*PS*	*SHPS*
Most cited publication over all publication years	1. popper k., 1959, logic sci discovery (0.8%) 2. lakatos i., 1970, criticism growth kno (0.7%) 3. popper k. r., 1963, conjectures refutati (0.5%) hempel c., 1965, aspects sci explanat (0.5%)	1. quine w. v. o., 1960, word object (1%) 2. dretske fi, 1970, j philos, v67, p1007 (0.7%) 3. gettier e., 1963, analysis, v23, p121 (0.6%)	1. hempel c., 1965, aspects sci explanat (1%) 2. nagel e, 1961, structure sci (0.6%) 3. kuhn ts, 1970, structure sci revolu (0.4%)	1. kuhn ts, 1970, structure sci revolu (0.7%) 2. lakatos i., 1970, criticism growth kno, p91 (0.5%) 3. kuhn t, 1962, structure sci revolu (0.4%)
Most cited publication within the respective publication years	1. einstein a, 1905, ann phys-berlin, v17, p891 (38%) 2. keynes j. m., 1921, treatise probability (37%) 3. einstein a, 1915, sitzber k preuss aka, p778 (30%)	1. henning h, 1916, z psychologie, v73 (100%) 2. russell b., 1905, mind, v14, p479 (67%) 3. aristotle, 1941, basic works aristotl	1. einstein a, 1905, ann phys-berlin, v17, p891 (40%) 2. keynes j. m., 1921, treatise probability (35%) 3. reichenbach h., 1938, experience predictio (26%)	1. duhem p, 1954, aim structure physic (14%) 2. kuhn t, 1962, structure sci revolu (12%) 3. kuhn ts, 1970, structure sci revolu (11%)
Most cited publication from 1905	• einstein a, 1905, ann phys-berlin, v17, p891 (38%)	• russell b., 1905, mind, v14, p479 (67%)	• einstein a, 1905, ann phys-berlin, v17, p891 (40%)	• einstein a, 1905, ann phys-berlin, v17, p549 (7%)

In Table 3, it is interesting to note that the journals are different in terms of their most frequently cited publications: whereas – for example – publications of Karl Popper seem to be important for authors in the *British Journal for the Philosophy of Science*, publications of Thomas Kuhn seem to be especially relevant for authors in *Studies in History and Philosophy of Science*. However, there is one paper published by Albert Einstein which is prominently cited in three of the four journals (*British Journal for the Philosophy of Science*, *Philosophy of Science*, and *Studies in History and Philosophy of Science*). The title of this paper is as follows: "Zur Elektrodynamik bewegter Körper" [On the electrodynamics of moving bodies] (Einstein, 1905).

In the interpretation of the percentages for the journal *Erkenntnis* one should consider that the cited reference counts can be low. With a total of 15,973 cited references, this journal has the lowest number of cited references compared to the other journals (see Table 1). In case of low numbers of cited references, publications with one or only a few cited references may contribute high percentages within a single publication year. For example, with only one single citation the cited reference "henning h, 1916, z psychologie, v73" achieved a 100% share in 1916 (see Table 3).

3.2 Second example

As a second example in order to demonstrate the utility of our software we analyze four research topics within the field of the natural sciences: light scattering at small particles of different materials (atmospheric aerosols, cosmic dust, colloids, nanoparticles). Again, the cited references of the publications in the sets of these research topics have been analyzed in terms of how often they have been cited. The results of the RPYS analysis are shown in Figure 3.

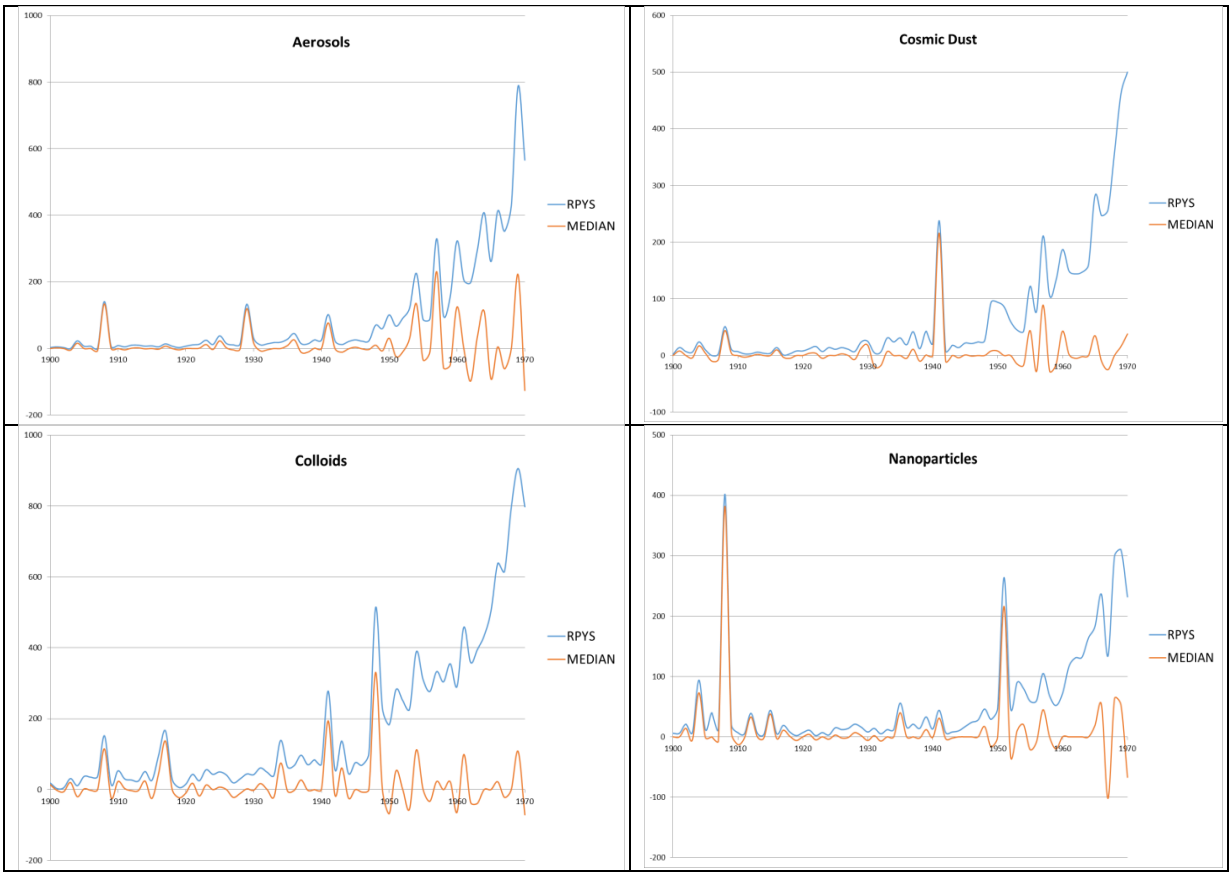


Figure 3. Results of the Reference Publication Year Spectroscopy for four research topics within the field of the natural sciences (atmospheric aerosols, cosmic dust, colloids, nanoparticles)

As in our first example, there are some common reference publication year peaks when comparing the four cases, in particular among earlier publication years. The quantiles of the numbers of cited references for the four natural sciences research topics are shown in Figure 4.

Figure 3 and Figure 4 reveal a distinct peak for the reference publication year 1908 which can be assigned to a paper published by the German physicist Gustav Mie (1908). This paper deals with the scattering of electromagnetic waves by small particles and was published in 1908 in *Annalen der Physik* under the title of "Contributions to the optics of turbid media, particularly of colloidal metal solutions". Mie (1908) applied Maxwell's electromagnetic theory to spherical gold colloids and provided a theoretical treatment of this phenomenon. The term "Mie scattering" is still an eponymy of his name: Mie scattering occurs when the

diameters of particles are similar to the wavelengths of the scattered light. For particles much larger or much smaller there are approximations. But for objects with sizes similar to the wavelength, e.g., aerosols or water droplets in the atmosphere, cosmic dust, colloids or nanoparticles, a more exact approach is needed. In contrast to Rayleigh scattering, Mie scattering is independent from the wavelength of the scattered light. This kind of light scattering occurs for example in the lower parts of the atmosphere and explains why clouds and the sky near the horizon appear white (and not blue like the upper part of the sky which underlies Rayleigh scattering). Mie's (1908) paper had hardly been cited before 1950 but meanwhile received altogether more than 6300 citations.

Other important early reference publication years for the four research topics analyzed here are 1904 and 1941: The peaks corresponding to the reference publication year 1904 in Figure 3 and Figure 4 can be attributed to a paper by Garnett (1904) entitled "Colours in metal glasses and in metallic films" which appeared in the *Philosophical Transactions of the Royal Society of London A*. The peaks corresponding to the reference publication year 1941 can be assigned to a paper by Henyey and Greenstein (1941) entitled "Diffuse radiation in the galaxy".

Similar to the first example of our analysis, the peaks corresponding to these most frequently referenced early publication years are not equally visible in the four research topics. For example, the paper by Garnett (1904) is less important for the colloids research topic and the paper by Henyey and Greenstein (1941) is particularly important for the astrophysics research topic, but this is not surprising considering the title.

RPY	Aerosols	Cosmic Dust	Colloids	Nanoparticles
1900	3,45	2,59	6,90	8,63
1901	6,04	26,73	0,00	7,76
1902	2,59	14,66	1,73	32,76
1903	0,00	11,21	13,80	12,94
1904	31,90	35,35	3,45	50,00
1905	11,21	18,11	16,38	20,69
1906	10,35	0,00	14,66	37,94
1907	0,87	6,04	15,52	22,42
1908	49,14	43,11	39,66	63,80
1909	6,90	20,69	4,31	30,18
1910	12,94	13,80	25,87	12,07
1911	5,18	5,18	12,94	6,04
1912	14,66	4,31	10,35	37,07
1913	13,80	10,35	7,76	6,90
1914	9,49	7,76	25,00	3,45
1915	12,07	6,90	9,49	39,66
1916	4,31	25,87	36,21	5,18
1917	24,14	1,73	40,52	29,31
1918	8,63	3,45	11,21	15,52
1919	1,73	17,25	2,59	1,73
1920	7,76	16,38	5,18	11,21
1921	17,25	22,42	20,69	17,25
1922	21,56	27,59	8,63	0,87
1923	33,63	12,94	27,59	10,35
1924	18,97	25,00	19,83	2,59
1925	37,94	19,83	24,14	25,87
1926	25,87	24,14	18,11	19,83
1927	16,38	18,97	6,04	25,00
1928	20,69	12,07	12,07	31,90
1929	48,28	34,49	21,56	27,59
1930	37,07	36,21	18,97	14,66
1931	15,52	8,63	28,45	24,14
1932	23,28	9,49	23,28	4,31
1933	26,73	38,80	17,25	18,97
1934	27,59	33,63	38,80	18,11
1935	36,21	37,94	29,31	43,97
1936	38,80	29,31	30,18	26,73
1937	25,00	39,66	35,35	31,04
1938	22,42	21,56	31,90	23,28
1939	35,35	41,38	34,49	36,21
1940	32,76	31,90	32,76	21,56
1941	46,56	56,04	45,69	38,80
1942	31,04	15,52	26,73	9,49
1943	18,11	28,45	37,94	13,80
1944	28,45	23,28	22,42	16,38
1945	34,49	31,04	33,63	28,45
1946	29,31	30,18	31,04	33,63
1947	30,18	32,76	37,07	34,49
1948	41,38	37,07	57,76	40,52
1949	39,66	47,42	43,11	35,35
1950	45,69	46,56	41,38	41,38
1951	40,52	45,69	46,56	59,49
1952	43,97	43,97	43,97	42,25
1953	47,42	42,25	42,25	49,14
1954	52,59	40,52	52,59	48,28
1955	42,25	49,14	49,14	44,83
1956	43,11	44,83	44,83	45,69
1957	56,04	55,18	50,00	50,87
1958	44,83	48,28	48,28	46,56
1959	50,00	50,00	50,87	43,11
1960	55,18	54,31	47,42	47,42
1961	51,73	52,59	56,04	51,73
1962	50,87	50,87	51,73	52,59
1963	54,31	51,73	53,45	53,45
1964	58,63	53,45	55,18	55,18
1965	53,45	58,63	56,90	56,04
1966	59,49	56,90	59,49	57,76
1967	57,76	57,76	58,63	54,31
1968	60,35	59,49	60,35	60,35
1969	62,94	62,07	62,94	61,21
1970	61,21	62,94	61,21	56,90

Figure 4. Quantiles of the number of cited references for four research topics within the field of the natural sciences (atmospheric aerosols, cosmic dust, colloids, nanoparticles)

Table 4. Top-three most cited publications of four research topics within the field of the natural sciences (light scattering at aerosols, cosmic dust, colloids, and nanoparticles). The table shows the most cited publications over all publication years (1900-1970), within a publication year and in 1908. Subsequent to each cited reference the corresponding percentage is mentioned.

	Research Topics			
	Aerosols	Cosmic Dust	Colloids	Nanoparticles
Most cited publication over all publication years	1. van de hulst h.c., 1957, light scattering sma (2.49 %) 2. chandrasekhar s., 1960, rad transfer (2.29 %) 3. mie g, 1908, ann phys-berlin, v25, p377 (1.91 %)	1. henyey lg, 1941, astrophys j, v93, p70 (3.98 %) 2. van de hulst h.c., 1957, light scattering sma (2.69 %) 3. chandrasekhar s., 1960, rad transfer (1.40 %)	1. verwey e.j.w, 1948, theory stability lyo (2.36 %) 2. kerker m., 1969, scattering light oth (1.59 %) 3. stober w, 1968, j colloid interf sci, v26, p62 (1.49)	1. mie g, 1908, ann phys-berlin, v25, p377 (9.03 %) 2. turkevich j, 1951, discuss faraday soc, p55 (3.73 %) 3. stober w, 1968, j colloid interf sci, v26, p62 (2.16 %)
Most cited publication within the publication years	1. harries c., 1907, ber dtsch chem ges, v40, p165 (100 %) 2. mie g, 1908, ann phys-berlin, v25, p377 (92.92 %) 3. koschmieder h., 1925, beitr phys atmos, v12, p33 (84.20)	1. plummer h.c., 1911, monthly notices of the royal astronomical society, v71 (100 %) 2. wolf m, 1917, astron nachr, v204, p41 (100 %) 3. mie g, 1908, ann phys-berlin, v25, p377 (86.27 %)	1. wilson ha, 1900, philos mag, v50, p238 (78.95 %) 2. pickering su, 1907, j chem soc, v91, p2001 (78.35 %) 3. mie g, 1908, ann phys-berlin, v25, p377 (76.98)	1. mie g, 1908, ann phys-berlin, v25, p377 (93.80 %) 2. gans r, 1912, ann phys-berlin, v37, p881 (92.30 %) 3. gans r, 1915, ann phys-berlin, v47, p270 (79.55 %)
Most cited publication in 1908	• mie g, 1908, ann phys-berlin, v25, p377 (92.92 %)	• mie g, 1908, ann phys-berlin, v25, p377 (92.92 %)	• mie g, 1908, ann phys-berlin, v25, p377 (92.92 %)	• mie g, 1908, ann phys-berlin, v25, p377 (92.92 %)

Again, the unified and aggregated cited references are based on the use of the Levenshtein similarity function (using the similarity value of 0.75 as a threshold) for the journal/book title and author names as follows:

java -jar RefMatchCluster.jar -input=yearcr.dbf -matcher=journal_short,Levenshtein,0.75 -matcher=lastname,Levenshtein,0.75 -match=yearcr_match.csv -cluster=yearcr_cluster.dbf -aggregate=cleaned.csv

The results of the unification and aggregation process are shown in Table 4: For each of the four research topics the three most frequently cited publications are shown – over all publication years (1900-1970), within a publication year, and in 1908.

Like the journals of the first example, the research topics of the second example show important differences (not surprising in consideration of the very different research topics) but also notable similarities. For example, four publications appear twice as the most cited publications over all publication years. Again, one publication is striking: The paper published by Mie (1908) is prominently cited by papers in all four research topics. It appears in all topics both in the category of the most cited publications during all publication years as well as in the category of the most cited publication in its publication year 1908.

Concerning the interpretation of the percentages of the most cited publication in Table 4, one should consider that – similar to our first example – the cited reference counts are frequently low. In case of low cited reference counts, publications with only one or a few cited reference counts may have very high percentile values within a single publication year. For example, with only one single citation the cited reference “harries c., 1907, ber dtsch chem ges, v40, p165“ is ranked in the 100^{th} percentile within 1907.

4 Discussion

Bibliometric data can be used not only for evaluation purposes, but also for scientifically related investigations in a historical context. The bibliometric cited-reference data in WoS can be traced back to the first ever published documents. RPYS was proposed by Marx, et al. (2014) to identify seminal publications in the knowledge base of a research field which are most important in a historical context. Based on ideas around the RPYS conducted by Comins and Hussey (2015), we refined our RPYS toolbox by adding some features to the existing programs and we developed two new routines. For this paper, two examples from the humanities and natural sciences have been used to demonstrate the functionalities and results of the programs. A more technical description of the usage of the programs can be found at http://www.leydesdorff.net/software/rpys/.

The toolbox can be used very flexibly in different contexts: (1) historical roots of research fields can be identified; (2) scientific myths can be uncovered (Marx & Bornmann, 2014); (3) most important publications for authors of specific journals can be identified (e.g. Leydesdorff, et al., 2014); (4) importance of single publications can be compared between different research fields, journals, and researchers. We would like to encourage scientists (bibliometricians) to follow the example of Comins and Hussey (2015) and to use the toolbox for one's own examples (i.e., for one's own research fields). We would be glad to receive hints for further improvements of the toolbox.

In our next project, we plan to develop a graphical user interface integrating the three programs. The user interface will be designed similarly to VOSViewer (van Eck & Waltman, 2010) and the CiteNetExplorer (van Eck & Waltman, 2014). For the user of the planned RPYS interface, it will be possible to show, interlink and adapt the results of a RPYS (graphs with RPY peaks and lists with cited references) in a flexible way.

References

Barth, A., Marx, W., Bornmann, L., & Mutz, R. (2014). On the origins and the historical roots of the Higgs boson research from a bibliometric perspective. *The European Physical Journal Plus, 129*(6), 1-13. doi: 10.1140/epjp/i2014-14111-6.

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: opportunities and limits. *Journal of Informetrics, 7*(1), 158-165.

Bornmann, L., & Marx, W. (2013). The proposal of a broadening of perspective in evaluative bibliometrics by complementing the times cited with a cited reference analysis. *Journal of Informetrics, 7*(1), 84-88. doi: 10.1016/j.joi.2012.09.003.

Comins, J., A., & Hussey, T. W. (2015). Compressing multiple scales of impact detection by Reference Publication Year Spectroscopy. *Journal of Informetrics, 9*(3), 449-454.

Einstein, A. (1905). Zur Elektrodynamik bewegter Körper [On the electrodynamics of moving bodies]. *Annalen der Physik, 17*, 891-921.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology, 54*(5), 400-412. doi: Doi 10.1002/Asi.10226.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia, PA, USA: Institute for Scientific Information.

Garnett, J. C. M. (1904). Colours in metal glasses and in metallic films. *Philosophical Transactions of the Royal Society of London Series a-Containing Papers of a Mathematical or Physical Character, 203*, 385-420. doi: DOI 10.1098/rsta.1904.0024.

Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of American Society of Civil Engineers, 77*, 1539-1640.

Henyey, L. G., & Greenstein, J. L. (1941). Diffuse radiation in the galaxy. *Astrophysical Journal, 93*(1), 70-83. doi: Doi 10.1086/144246.

Kuhn, T. S. (1962). *The structure of scientific revolutions* (2. ed.). Chicago, IL, USA: University of Chicago Press.

Lack, D. (1947). *Darwin's finches*. Cambridge, UK: Cambridge University Press.

Leydesdorff, L. (2010). What Can Heterogeneity Add to the Scientometric Map? Steps towards algorithmic historiography. In M. Akrich, Y. Barthe, F. Muniesa & P. Mustar (Eds.), *Débordements: Mélanges offerts à Michel Callon* (pp. 283-289). Paris: École Nationale Supérieure des Mines, Presses des Mines.

Leydesdorff, L. (2012). Accounting for the uncertainty in the evaluation of percentile ranks. *Journal of the American Society for Information Science and Technology, 63*(11), 2349-2350.

Leydesdorff, L., Bornmann, L., Marx, W., & Milojevic, S. (2014). Referenced Publication Years Spectroscopy applied to iMetrics: *Scientometrics, Journal of Informetrics,* and a relevant subset of JASIST. *Journal of Informetrics, 8*(1), 162-174. doi: DOI 10.1016/j.joi.2013.11.006.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 12*, 317 - 323.

Marx, W., & Bornmann, L. (2014). Tracing the origin of a scientific legend by reference publication year spectroscopy (RPYS): the legend of the Darwin finches. *Scientometrics, 99*(3), 839-844. doi: DOI 10.1007/s11192-013-1200-8.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology, 65*(4), 751-764. doi: 10.1002/asi.23089.

Mie, G. (1908). Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen (Contributions to the optics of turbid media, particularly of colloidal metal solutions). *Annalen der Physik, 25*(3), 377-445.

Pendlebury, D. A. (2008). *Using bibliometrics in evaluating research*. Philadelphia, PA, USA: Research Department, Thomson Scientific.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. doi: 10.1007/s11192-009-0146-3.

van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics, 8*(4), 802-823. doi: DOI 10.1016/j.joi.2014.07.006.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., . . . Wouters, P. (2012). The Leiden Ranking 2011/2012: data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419-2432.

Wray, K. B., & Bornmann, L. (2014). Philosophy of science viewed through the lense of "Referenced Publication Years Spectroscopy" (RPYS). *Scientometrics*, 1-10. doi: 10.1007/s11192-014-1465-6.