MDPI

Article

Revealing trends in geophysics using metadata analysis

Timofey Eltsov [1,†], Maxim Yutkin [2], Tadeusz W. Patzek [3]

[1] Ali I. Al-Naimi Petroleum Engineering Research Center, King Abdullah University of Science and Technology; timofey.eltsov@kaust.edu.sa
[2] Ali I. Al-Naimi Petroleum Engineering Research Center, King Abdullah University of Science and Technology; maxim.yutkin@kaust.edu.sa
[3] Ali I. Al-Naimi Petroleum Engineering Research Center, King Abdullah University of Science and Technology; tadeusz.patzek@kaust.edu.sa
* Correspondence: timofey.eltsov@kaust.edu.sa; Tel.: +966128087182
† Current address: 4700 KAUST, Thuwal, 23955-6900, Saudi Arabia

Version February 25, 2020 submitted to Geosciences

Abstract: Professional language evolution reveals the development of geophysics: researchers enthusiastically describe new methods of survey, data processing techniques, and objects of their study. Geophysicists publish their cutting-edge research at international conferences proceedings to share their achievements with the world. Tracking changes in the language allows one to identify trends and the current state of the science. Here, we describe the metadata analysis of the last 38 Annual Conferences organized b y the Society of Exploration Geophysicists, one of the biggest geophysical gatherings. The number of publications from oil companies reflects their financial situation; the number of papers from the academia of various countries indicates government financing of research. The USA academia has the most significant number of publications; in 2019, the number of papers from China was almost equal to those of the USA. An analysis of conference materials metadata allows one to identify trends in a specific field of knowledge and predict the development in the near future.

Keywords: geophysics; web data analysis; data mining; data analysis; metadata analysis

1. Introduction

The last four decades showed a tremendous change in geophysics. An increase in computing power and technological progress allowed geophysicists to solve more and more complicated tasks. At the same time, the field of application of geophysics is expanding; the market of geophysical services is changing. We assume that a change of geophysical tasks, applications, geography, and technology will inevitably lead to a shift in the professional language. If one can track changes in the frequency of terms used in recent years, one can shed light on the current state of the industry and possibly predict future changes. Authors apply language processing methods to analyze changes in the professional language in geophysics.

The biases of different origin complicate big data [1]. In machine learning, the difference between training data set and test data set can cause biases. Massive sample study can lead to bias associated with an error resulting from sampling or study design [2]. Supposedly, it is better to have a smaller and more representative data set rather than a much bigger but biased data. We want to understand what the modern geophysical language looks like and what the future of geophysics will be. In this paper, we analyze only scientific articles presented at the Society of Exploration Geophysicists (SEG) Annual Conference and Exhibition. The committee selects the papers for the conference each year; this

is the initial filtering. Also, it is worth noting that presenting at such a meeting is a demonstration of the technical capabilities of industrial companies and the scientific viability of academic institutions. Each annual conference proceedings is a cross-section of the state of geophysics, and we use it for analysis and predictions.

The SEG Annual Conference and Exhibition is one of the biggest gatherings of geophysicists in the world. Abstracts of the SEG Annual Conferences are a representation of the state of geophysical science, approximated mainly to the oil and gas industry. Articles in the electronic version for the 38 years are available for analysis [3]. The SEG conducted all their Annual conferences in the USA, and the last one was in San Antonio, TX. For analysis, the authors selected the proceedings of the SEG Annual Conference, as the most representative set, that reflects state-of-art-technologies in geophysics. Each conference proceedings is a reflection of the state of the industry in a particular year since, at this event, both academic institutions and the industry present their best achievements in the field.

Besides conference proceedings, one can use journal articles for data mining as the volume of the data for one year is comparable to the SEG Annual Conference and Exhibition Proceedings. The number of publications per year is smaller, but they consist of full-size papers. However, the release of articles in journals is carried out periodically, e.g., monthly or quarterly; at the conference, this happens once a year. The research materials are usually published in journals and reported at conferences; the proceedings include many of the results from full-sized articles. Moreover, the number of research teams presenting their work is several times larger in the case of analysis of conference materials compared to the study of one particular journal. SEG Annual Conference proceedings represent a collection of scientific research from a large number of scientific teams in one place for each of the 38 years. This approach allows one to conduct a unique study and trace the dynamics of changes in the industry.

2. Materials and Methods

We used the OnePetro online library [4] to get metadata of the reviewed papers. The OnePetro website offers ample opportunities for analyzing metadata. Along with OnePetro, CrossRef service can be used for metadata analysis. Different spelling versions and typos affected the study of affiliation. Besides, many organizations have since ceased to exist (acquired, bankrupt, split, *etc.*), which also complicated the analysis. We use open-source Python libraries: to transform, filter and process the text, and get metadata: TextBlob, NLTK (Natural Language Toolkit), argparse, Pandas, Scrapy, Requests-HTML, sqlite3, and NumPy. For printing the graphs, we use Matplotlib, Plotly, PIL (Python Imaging Library), and others.

In total, we analyzed 24,500 papers consisting of more than 57 million words or more than 383 million symbols. The number of non-unique authors for the entire time span chosen in this paper exceeds 75 thousand.

3. Results

More than 2400 industry companies and academic institutions from eighty-six countries heve presented their research at the SEG Annual Conference so far. The five companies with the most significant number of publications in the SEG Annual Conference are 1) Schlumberger, 2) WesternGeco, 3) CGG (Compagnie Générale de Géophysique), 4) BGP Inc. (BGP Inc., China National Petroleum Corporation), and 5) BP plc (formerly The British Petroleum Company plc and BP Amoco plc). These five companies accounted for about 30% of all affiliations in the past ten years. Schlumberger itself constitutes about 8% of all affiliations, with WesternGeco adding about 5% in the past ten years. The five most highly represented universities in SEG Annual Conferences over the 38 years are 1) Colorado School of Mines, 2) University of Houston, 3) China University of Petroleum, 4) Stanford University, and 5) Delft University of Technology.

Fig. 2 reveals the average number of papers for the academia by country. Each country is represented by a unique color; the size of the circles is the average number of publications. The inset

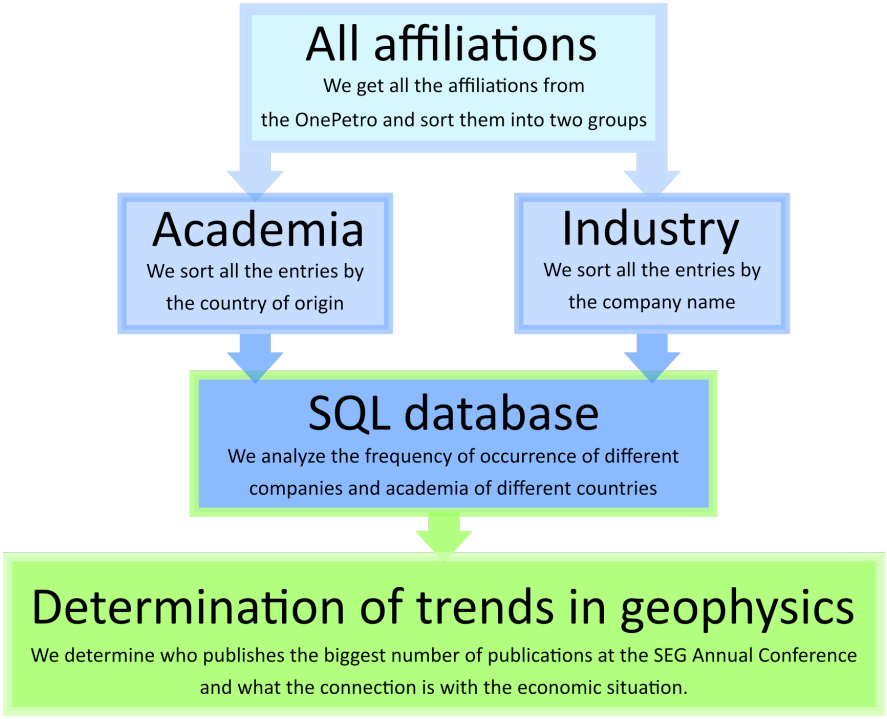

Figure 1. Data processing workflow.

in the lower-left corner shows a Europe zoom-in. The most considerable contribution is from the universities in the United States of America (the USA, see the five above) followed by universities in China (China University of Petroleum, Jilin University, and Tongji University), and Canada (University of British Columbia). The rest is shared among the Netherlands, France, Germany, and Russia.

China holds the leading position in Asia, followed by smaller but notable contributions from South Korea, Japan, and India. In South America, Brazil academia publishes the most abstracts. In the Middle East, the most represented country at the SEG Annual Meetings is the Kingdom of Saudi Arabia over the past ten years. In 2009 the average number of publications was about one, and in 2019, it is more than 23, which is indeed impressive. None of the academia of the other countries have shown such rapid relative growth in recent years. The total number of publications from the top-50 countries is presented in the appendix, and the full list can be accessed here [5].

The circle on Antarctica represents incomplete affiliations or affiliations with typos that were not correctly recognized; therefore, it was impossible to determine the location. It provides an estimate of the total error of the analysis.

Fig. 3 compares the annual number of publications from industry and academia. Both contributors show a steady growth over the years, which is associated with an increase in fossil fuel consumption, oil price, and constant-growth-economic paradigm. However, on the finer scale, there is a weak correlation with the oil price change. For example, the number of academic publications was hardly affected by the two recent crises in 2008 and 2014. On average, the number of industrial publications is only partially influenced by the oil price dynamics resulting in a slower growth rate in the last few years. It appears that, on average, the industry became more efficient in research expenditure optimization, which enabled them to maintain a high number of publications and even a slow but consistent growth during the shrinking market.

Curiously, the number of publications from the academia in the last year has fallen significantly. Fig. 3 indicates a decrease in the number of publications from the USA academia in 2019 compared to 2018. Perhaps this was due to a reduction of state funding of higher educational institutions [6]; see blue curve drop in 2018-2019 Fig. 5.

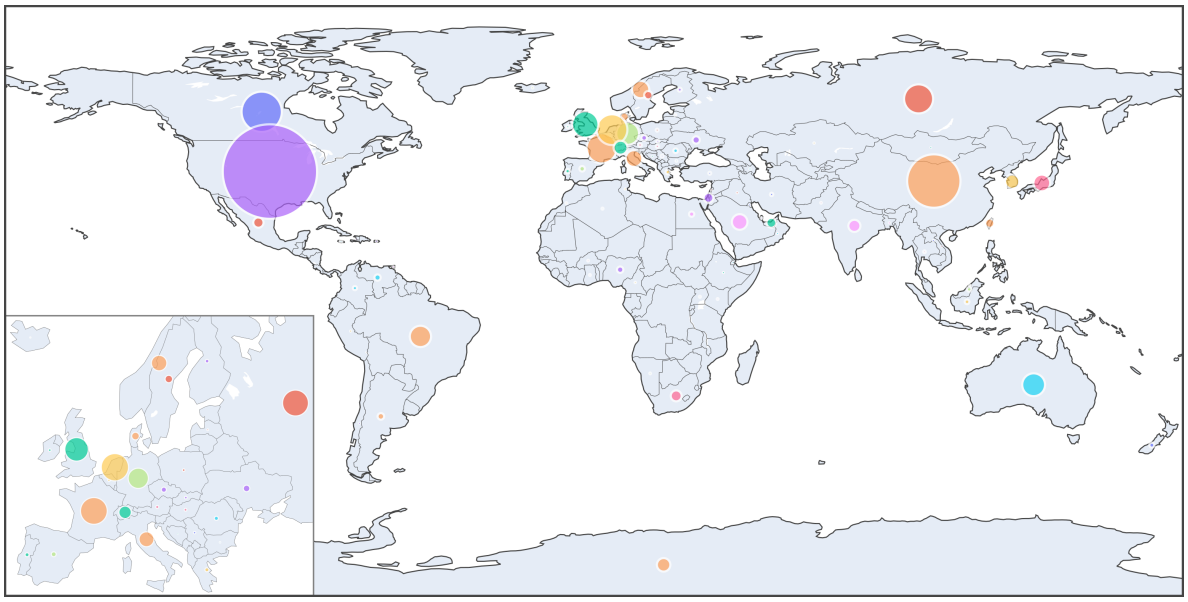

Figure 2. The total number of academic publications by country. Europe is in the zoomed inset on the lower left. The circle on Antarctica represents erroneous affiliations and serves as an error indicator.

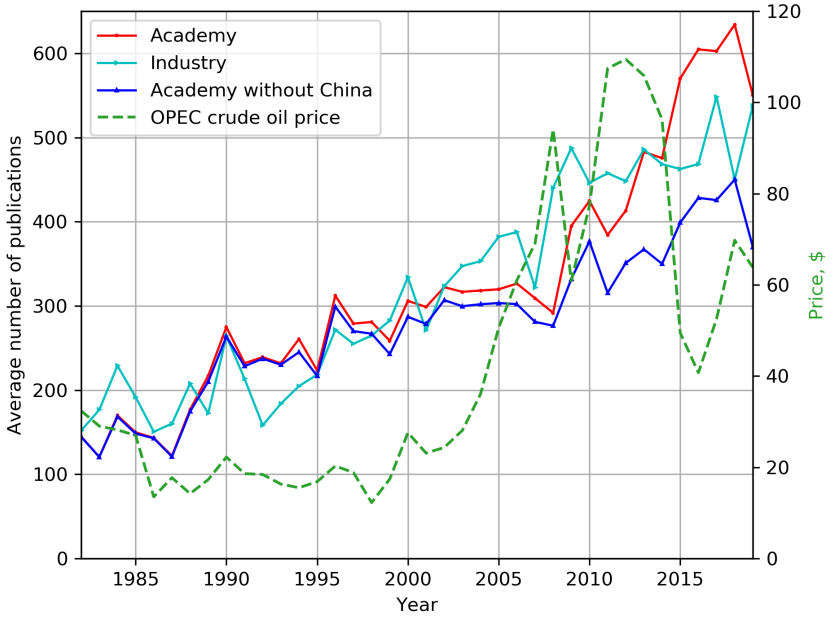

Figure 3. The annual breakdown of net industry and academia publications.

Fig. 4 shows that the number of co-authors per paper has increased and we observe a correlation with the world trend exemplified by a related field of the Earth and Planetary sciences. The increase in the number of authors per paper is a worldwide trend [7]; scientific research is becoming more interdisciplinary and thus more collaborative. The SEG average co-author number almost flattens out at 3.6 co-authors per paper, but in 2019 the number of authors per paper increased, reaching 3.9. With that, we see an increase in the number of organizations involved in the SEG Annual Conference.

Fig. 5 shows a breakdown of academic publications by country. The USA academia is ahead of everyone in the number of papers published annually, as well as total articles published. The Netherlands and Canada, have regular contributions, and their publication activity is constant over time. Other countries, like France and Germany, seem to follow the crude oil price trend. In contrast, China maintains a steady growth rate. In 2006 the Chinese government initiated a powerful program of

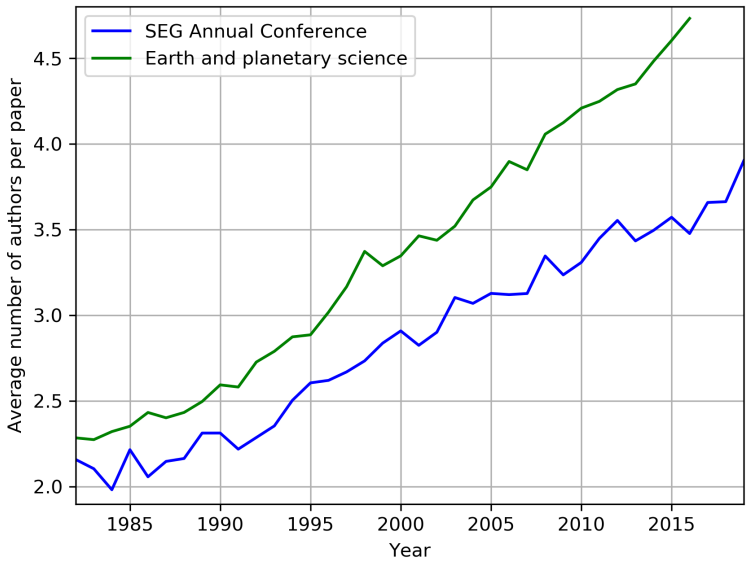

Figure 4. Number of co-authors growth rate for the SEG Annual Conference and for the Earth and Planetary science [7].

research development, "Medium and Long-term Plan for the Development of Science and Technology (2006–2020)" with a target of 2.5% GERD/GDP ratio by 2020 [8]. The strong support of the government for geoscience, allowed the Chinese academia to exhibit the fastest growth between 2008 and 2015. In 2013, we observed a 15% increase on R&D spending by China compared to 2012 [9]. The number of publications by the Chinese academia is now almost equal to those of the USA.

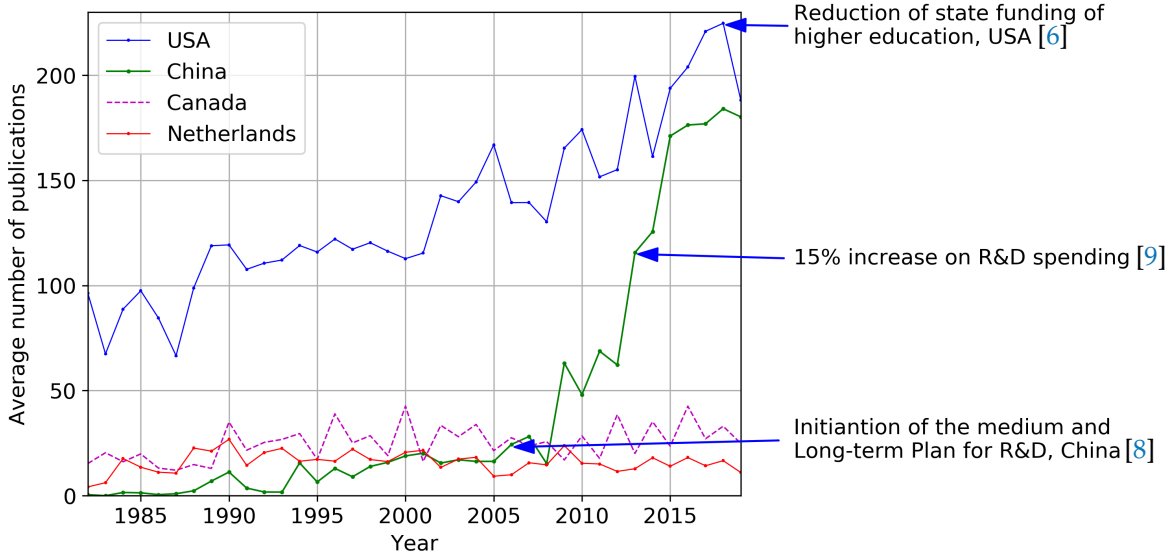

Figure 5. Average number of publications by academia of the TOP-four countries academia.

Next, we perform a similar analysis for industry publications. Fig. 7 shows the average number of papers by oilfield service companies. The most frequent guests at the SEG Annual Conferences are Schlumberger, WesternGeco, CGG, and BGP. Although WesternGeco is a part of Schlumberger, we show them separately according to the affiliation. Schlumberger dominates industrial geophysics research, followed by CGG and BGP. In general, the number of publications by the major oilfield

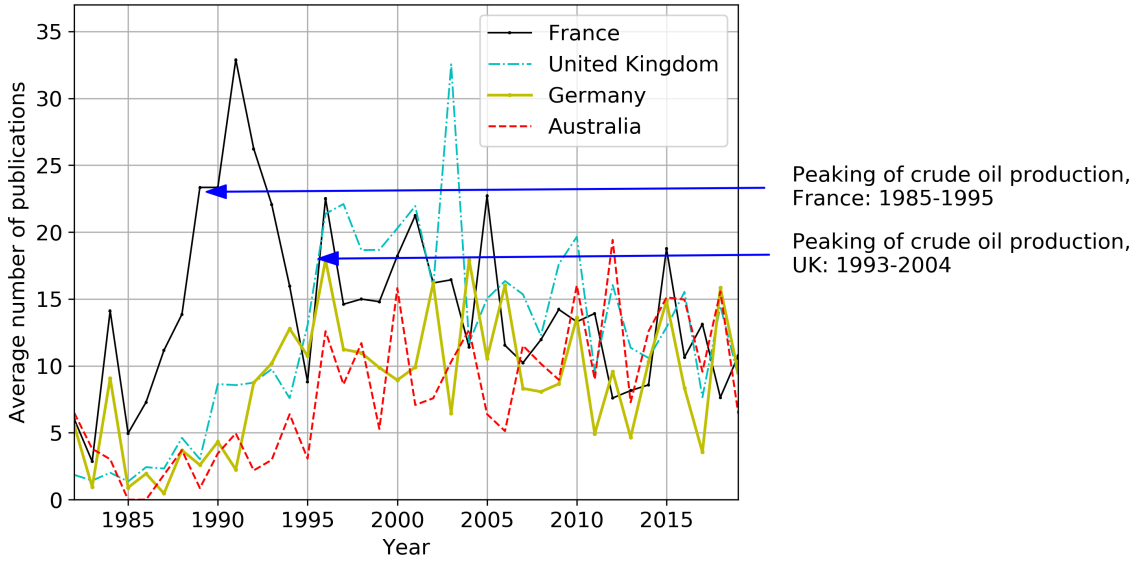

Figure 6. Average number of publications by France, UK, Germany, and Australia. Crude oil production data is presented online [10].

service companies grew steadily. Although oilfield service providers are dependent on oil prices, we surprisingly observe that after the 2014 crisis, the number of Schlumberger publications peaked for several consecutive years, followed by a decline in 2018. It should be noted that in 2014 Schlumberger reported an outstanding revenue of $48.6 billion. The dynamics of Schlumberger's papers reflect the dynamics of oil prices with a few years offset. The number of CGG publications number follows crude oil prices too, but since 2014, the number of publications from CGG has only declined. CGG Annual report states extremely difficult market environment [11], and cost reduction measures: reduction in the number of employees from 11060 to 7353, 55% general and administrative cost-cutting, 64% cost of marine monthly structure cost-cutting[1]. In January 2020 CGG reported its exit from marine acquisition, sale of ships and measuring equipment to Shearwater company [12]. This news suggests that we will see fewer publications from CGG in the coming years.

The change in the number of BGP publications shows a similar trend with the crude oil prices with one or two years of delay. In 2019 BGP became a leader by the number of publications among oil service providers.

Many oil and gas companies that no longer exist made significant contributions to the SEG Annual Conference in the 1980s and early 1990s. They are Arco Oil and Gas Co., Mobil E&P, OYO Corporation, Statoil, and others. These companies either merged with others, changed their names, or were acquired.

Fig. 8 displays five oil production companies with the most significant number of publications. The picture is conceptually different from oil service companies. For example, the number of publications by BP and ExxonMobil peaked in 2005 and nowadays, it is declining. 2005 was an outstanding year for ExxonMobil, with a net income of 36 billion and a 31% increase in the number of employees [15]. We observe steady growth in many economic indicators of the company since the beginning of 2000. At the same time, a decrease in the number of publications indicates a difficult period for the company. For instance, in 2014, we found only one paper from ExxonMobil, which has not happened over the past 15 years. The 2014 ExxonMobil Summary Annual Report [16] shows that compared to 2013, market valuation at the end of the year decreased by 12%, and we observe

[1] Comparing between 2013 and 2015

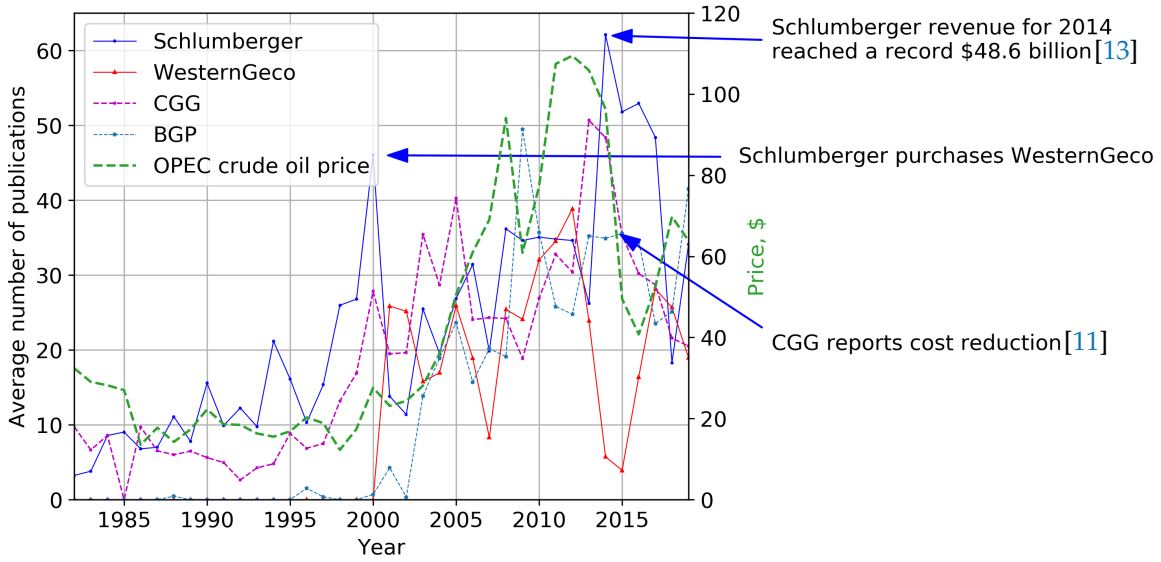

Figure 7. The average number of publications by oil-service companies and OPEC crude oil prices.

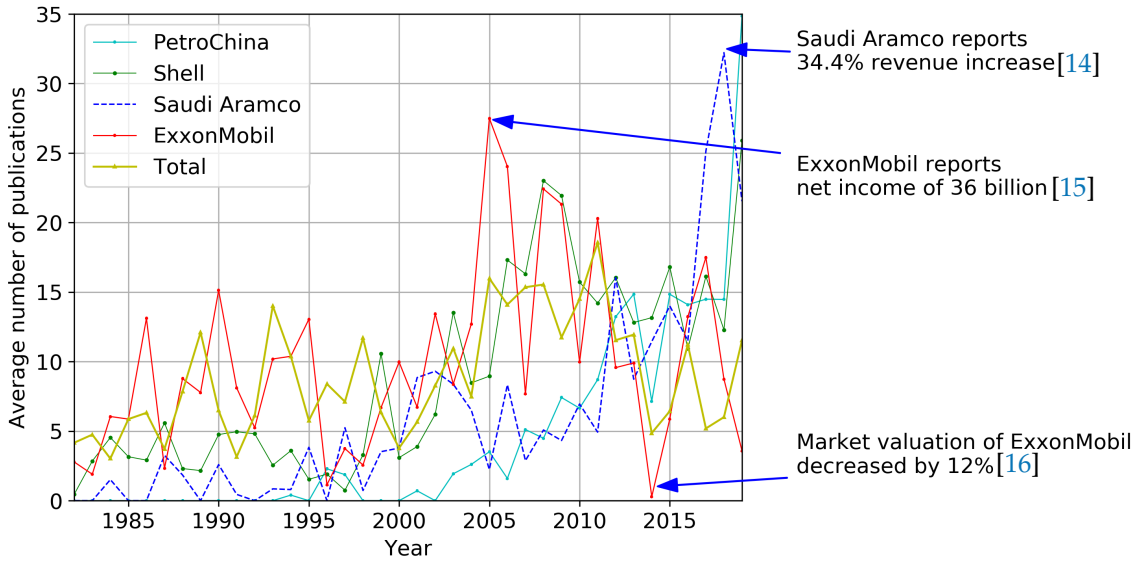

Figure 8. The average number of publications by oil production companies.

the decline in the stock market price of ExxonMobil in 2015. The decrease in profits immediately affects research financing. Saudi Aramco demonstrates steady growth; it had the biggest number of publications of all production companies in 2017 and 2018. Interestingly enough, the leadership was taken by PetroChina in 2019, followed by Shell and Saudi Aramco.

4. Discussion

The amount of hidden information is astounding. Such a study is unique because we have at our disposal a history of the development of geophysics.

5. Conclusions

We analyzed metadata of 24,500 papers prepared by more than 75,000 authors during 38 years. Academic institutions from 86 countries and more than 2400 industrial companies contributed to the SEG Annual Conference from 1982 to 2019. The USA academia has the most significant impact in

the proceedings of the SEG Annual Conference during the whole observation period. We observe that the number of papers from the Chinese academia growing, and it is almost equal to those of the USA. The activity of the companies at the SEG Annual Conference shows their economic condition, annual reports by CGG and ExxonMobil and other companies confirm this statement. Depending on the financial situation on the market, and the price of oil, the contribution of the academia and industry by publications changes in time. In 2018 we observed more abstracts from the academia, but in 2019 the number of publications from academia and industry were very close. In 2019 the most significant number of publication in the industry was made by BGP and PetroChina. The average number of authors per paper continues to grow over time in agreement with the global trend of Earth and Planetary science, but at a slower rate.

Author Contributions: Data mining and processing, software development, original draft preparation - Timofey Eltsov; software development and analysis, review and editing of the draft - Maxim Yutkin; supervision, project administration, historical analysis, review, and editing of the paper - Tadeusz W. Patzek.

Funding: Dr. Eltsov was supported by the KAUST Magnetic Sensor project, REP-2708.

Acknowledgments: Authors appreciate the responsiveness of the SEG team for permission to use digital data and especially SEG Digital Publications Manager, Jeno Mavzer, for the useful advice and help. The authors are grateful to their colleagues, and especially to Dr. Thomas Finkbeiner, for valuable and vital research recommendations. The authors are grateful to Ilya Kolganov for the useful advice on the design of the graphs. We also would like to acknowledge Dr. Charles Russell Severance for an informative Python course.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASCII	American standard code for information interchange
BP	BP plc., formerly The British Petroleum Company and BP Amoco
CGG	Compagnie Générale de Géophysique
CMP	Common Mid Point
EAGE	European Association of Geoscientists and Engineers
GDP	Gross Domestic Product
GERD	Gross domestic Expenditure on Research and Development
GPU	Graphics Processing Unit
HTML	HyperText Markup Language
NLTK	Natural Language Toolkit
NMO	Normal Moveout
PDF	Portable Document Format
PIL	Python Imaging Library
R&D	Research and Development
SEG	Society of Exploration Geophysicists
SPE	Society of Petroleum Engineers
SPWLA	Society of Petrophysicists and Well Log Analysts
USA	The United States of America

Appendix A. The total number of publication by country

#	Country name	The total number of publications
1	United States of America	5154.07
2	China	1649.95
3	Canada	952.47
4	Netherlands	608.58
5	France	546.53
6	United Kingdom of Great Britain and Northern Ireland	462.14
7	Germany	330.39
8	Australia	302.54
9	Brazil	264.0
10	Japan	205.62
11	Russian Federation	178.75
12	Norway	176.36
13	Italy	164.11
14	Saudi Arabia	145.46
15	Korea, Republic of	138.62
16	Switzerland	113.01
17	India	99.08
18	Mexico	62.15
19	Denmark	57.58
20	Israel	55.69
21	Taiwan	49.44
22	Sweden	41.45
23	Venezuela (Bolivarian Republic of)	28.01
24	Argentina	25.78
25	South Africa	24.53
26	Nigeria	23.96
27	United Arab Emirates	19.75
28	Czechia	19.34
29	Spain	17.86
30	Egypt	16.52
31	Singapore	14.98
32	New Zealand	14.77
33	Romania	13.7
34	Indonesia	13.42
35	Malaysia	12.66
36	Greece	12.06
37	Portugal	11.74
38	Ukraine	11.08
39	Colombia	10.25
40	Finland	9.8
41	Austria	6.92
42	Iran	6.78
43	Belgium	5.43
44	Slovakia	5.35
45	Ireland	5.26
46	Poland	5.22
47	Turkey	5.2
48	Serbia	4.5
49	Thailand	4.34
50	Jamaica	4.32

References

1. Glauner, P.; Valtchev, P.; State, R. Impact of Biases in Big Data. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 25-27 April 2018; pp 645–654.
2. Kaplan, R.; Chambers, D.A.; Glasgow, R.E. Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *CTS Journal* **2014**, *7*, 4, 342–346. doi:10.1111/cts.12178.
3. SEG Technical Program Expanded Abstracts. Available online: https://library.seg.org/series/segeab (Accessed on 2 December 2019).
4. OnePetro online library. Available online: https://www.onepetro.org (Accessed on 2 December 2019).
5. Eltsov, T. Data for SEG Annual Conferences analysis, 1982 - 2019. Available online: https://github.com/ANPERC-source/SEG_Annual (Accessed on 2 February 2020).
6. American Higher Education Hits a Dangerous Milestone. Available online: https://www.theatlantic.com/politics/archive/2018/05/american-higher-education-hits-a-dangerous-milestone/559457/ (Accessed on 2 December 2019)
7. Paper authorship goes hyper. Available online: https://www.natureindex.com/news-blog/paper-authorship-goes-hyper (Accessed on 17 October 2019).
8. UNESCO. *UNESCO SCIENCE REPORT, Towards 2030*; Report; United Nations Educational, Paris, France, 2015.
9. Ni, X. China's research & development spend. *Nature* **2015**, *520*, S8-S9. doi:10.1038/520S8a.
10. Global Economic Data, Indicators, Charts& Forecasts, CEIC. Available online: https://www.ceicdata.com (Accessed on 10 February 2020).
11. Compagnie Générale de Géophysique *ANNUAL REPORT 2015*; Report; CGG: Chicago, Ill., USA 2015.
12. Compagnie Générale de Géophysique CGG completes its exit from marine acquisition. Available online: https://www.cgg.com/en/Investors/Press-Releases/2020/01/CGG-Completes-its-Exit-from-Marine-Acquisition (Accessed on 2 February 2020).
13. Schlumberger. *2014 Annual Report*; Report; Schlumberger: Paris, France 2015.
14. Saudi Aramco. Saudi Arabian Oil Company, Consolidated financial statements for the year ended december 31, 2018. Available online: https://www.saudiaramco.com/-/media/publications/corporate-reports/saudi-aramco-results-2017-2018-full-financials.pdf (Accessed on 11 February 2020).
15. ExxonMobil. *Summary Annual Report 2005*; Report; ExxonMobil: Irving, TX, USA, 2005.
16. ExxonMobil. *Summary Annual Report 2014*; Report; ExxonMobil: Irving, TX, USA, 2014.