& Biochemistry (hereinafter referred to as high-P_j disciplines). In contrast, significantly lower values were observed in Engineering, Computer Science, Neuroscience & Behavior (hereinafter referred to as low-P_j disciplines). For the Scopus data, the average and σ of P_j were 5.19 and 1.12, respectively, and the high-P_j discipline was absent, while the low-P_j disciplines were Engineering and Computer Science in common with the Web of Science. Regardless of which database we used, the relative deviation of the total number of papers was smaller for the normalized values than for the actual values in the high-P_j disciplines, and the opposite was true in the low-P_j disciplines. This suggests that the normalization worked as expected, according to the intention to make a relative increase in the number of papers in disciplines with low publication productivity.

To further investigate the similarity of P_j values between the two databases, we compared the P_j values of the top five well-populated disciplines in the Web of Science to those of the corresponding disciplines in Scopus (Figure 2). Despite slight differences in the magnitude, the observed variation in the P_j values was fairly similar, except for the pair of Physics in the Web of Science and Physics and Astronomy in Scopus. Analysis of the Kyutech's publication data shows that the results do not differ significantly between the two databases as far as the well-populated disciplines with similar field coverage are concerned. However, the robustness of the conclusion, should be further studied because our analysis was made mainly for disciplines included in a broad field, named Natural Sciences and Engineering; it has a relatively small difference in field coverage between the two databases in comparison to the other fields, i.e., Biomedical Research, Social Sciences and Arts and Humanities [18].

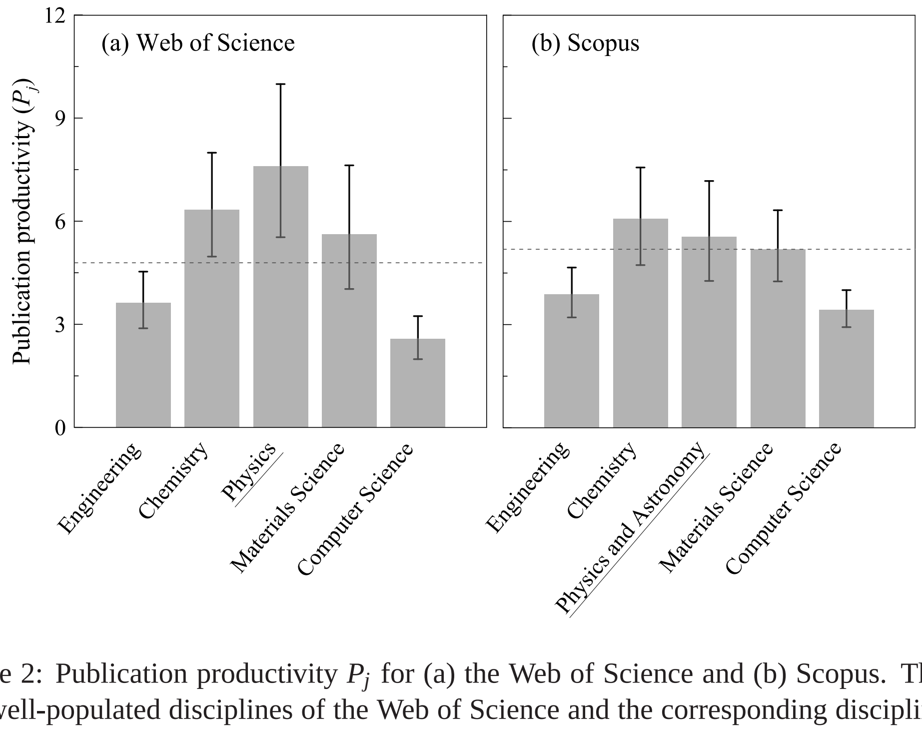

Figure 2: Publication productivity P_j for (a) the Web of Science and (b) Scopus. The top five well-populated disciplines of the Web of Science and the corresponding disciplines of Scopus are shown here. Underlined axis labels indicate the disciplines, field coverage of which is not in complete correspondence between the two databases. For thin dotted lines and error bars on the P_j bars, see the caption to Figure 1.

It is worth pointing out here that every researcher is equally assumed to have one manpower in the analysis (see Section 2.2). In fact, the number of hours that a given researcher can devote to a research field depends on his or her commitment to other tasks, e.g., activities for education and organizational operation. It would not be appropriate to compare the P_j values if the average conditions vary to a large extent among target groups.

3.2 Minimal sample size for P_j determination

It is crucial to determine the confidence interval of P_j with high accuracy for a comparison of the values among disciplines. As long as the minimal sample size required for accurate determination is unknown, it is difficult to guarantee that P_j values in Figure 1 are statistically significant. To estimate the minimal sample size, we take all the publication data as a parent population; the sample size is 542 for the Web of Science and 1340 for Scopus. The sample size substantially exceeds the number of authors (see Table 2) because many researchers are publishing papers in multiple disciplines. The difference is larger for the Scopus data because of an apparent increase in the amount of data caused by the allocation of paper count to multiple disciplines (see Section 2.2).

Figure 3 shows the 95% confidence intervals calculated by incrementing n in the bootstrap calculation procedure (1) (see Section 2.4) by 20; the n value starts from 20 to the extent that n does not exceed the sample size of the parent population. Because the results vary depending on which data are selected in the procedure (1), the calculation was made 100 times at each sample-size step with changing random seeds to obtain the average and σ of the upper and lower limits. As expected, a small sample size was accompanied by a large confidence interval, and the interval became narrower as n increased. The rate of change, defined as $(x_k - x_{k-1})/x_{k-1} \times 100$ (%), where x indicates arbitrary variable and k iteration step, was calculated for the average of the limits to detect convergence to a steady state; it fell below 1% when the sample size exceeded 200 for the Web of Science data, while the threshold was 1.5 times larger, around 300, for the Scopus data. The difference may be attributable to the apparent increase in the amount of the Scopus data, mentioned above.

3.3 Effects on researcher rankings

Statistical analysis in the previous subsection proved that various magnitudes of uncertainty existed in the P_j values because of insufficient data. Furthermore, the resulting P_j is the disciplinary average for the researchers of Kyutech; there would be some deviations in the normalized index DWPP calculated with the P_j from its correct values calculated with the disciplinary average for the total population. Nevertheless, it is worth presenting a concrete example of application of the index to researchers' performance evaluation in order to suggest predictable consequences.

The change of the measure from the number of papers to the DWPP resulted in a change in researcher rankings based on publication productivity (Figure 4). The sparsely populated disciplines, which had been excluded in Section 3.1, were included in this analysis to provide all those concerned their rankings, disregarding possible large statistical errors. In both cases of using the data from the Web of Science and Scopus, the higher or the lower the researcher ranking before the normalization was, the smaller the change in ranking was. The normalization caused a large change in ranking for the middle-ranking researcher (in the top 21%–80% before normalization); the maximal variation width was −80–60 based on the Web of Science data, while it was −60–50 based on the Scopus data. The slightly

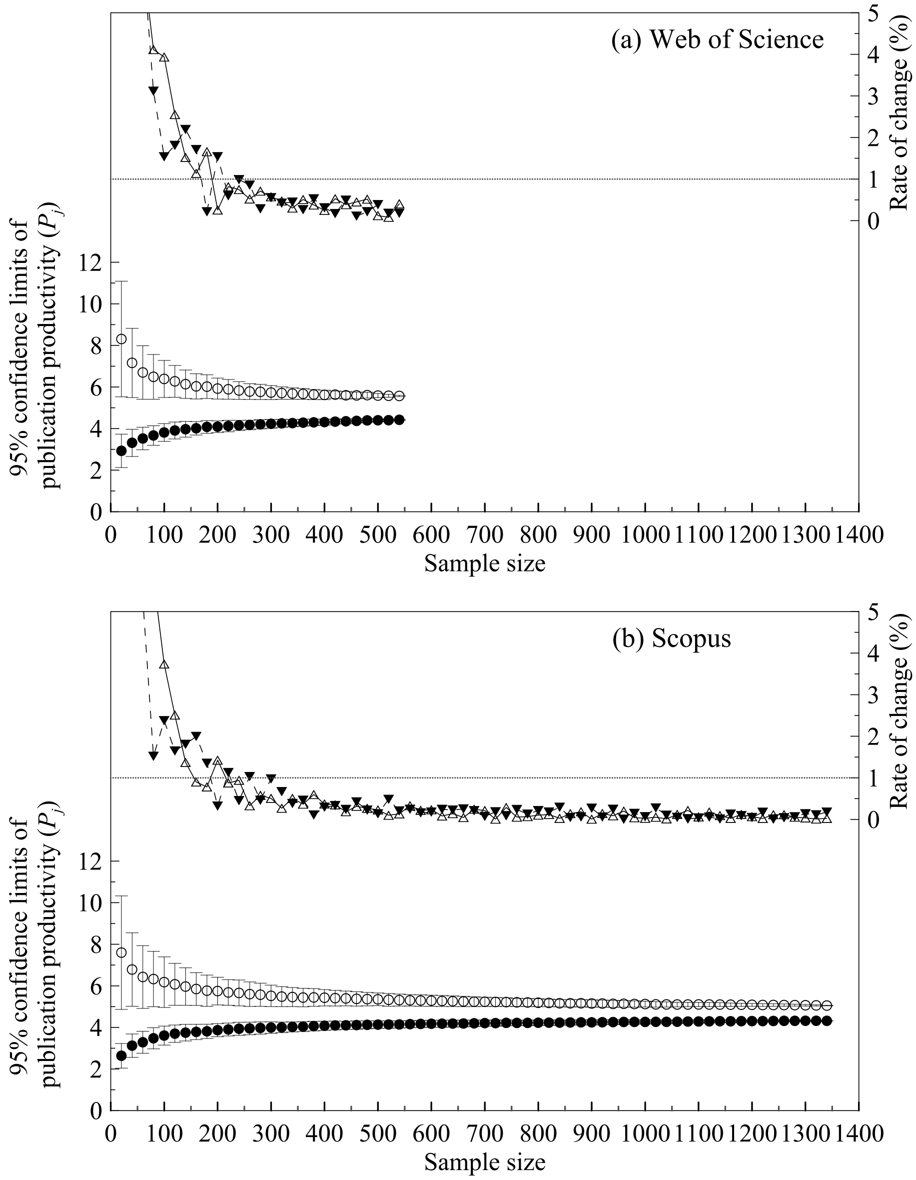

Figure 3: Dependence of 95% confidence limits of publication productivity P_j on the sample size. The average of upper limits (open circles) and lower limits (closed circles) were obtained from 100 times repetition of bootstrap calculations, and error bars indicate the standard deviation. Also plotted are the rates of change in the average of upper limits (open triangles) and lower limits (closed triangles). The values were calculated for the data of (a) the Web of Science and (b) Scopus.

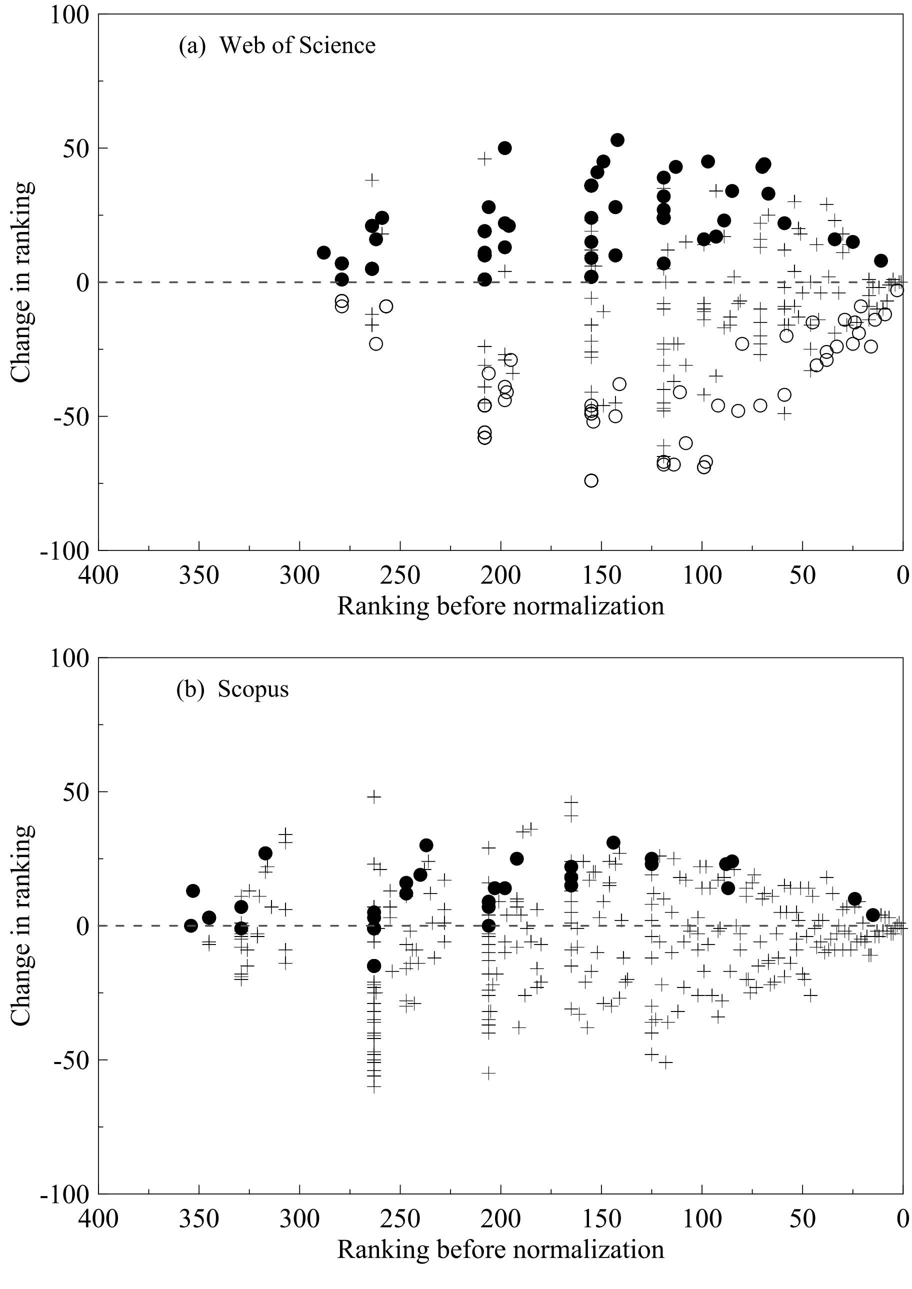

Figure 4: Change in researcher ranking calculated by subtracting the ranking before normalization from the ranking after normalization. Data are sorted by the ranking before normalization and plotted separately for researchers publishing papers mainly in the low-P_j disciplines (closed circles), researchers publishing papers mainly in the high-P_j disciplines (open circles), and otherwise (crosses). The analysis was done for the data of (a) the Web of Science and (b) Scopus.

small width for Scopus may be attributable to the larger number of disciplines of ASJC (see Table 2) or the larger number of disciplines per researcher caused by the allocation of paper count. In summary, the ranking of researchers publishing a moderate number of papers is changeable by normalization, and that of researchers who are prominently productive or unproductive is rather stable.

In every band of the ranking before normalization, most researchers who mainly published papers in the low-P_j disciplines, with the publication effort greater than 90%, enjoyed a marked rise in ranking, and the opposite can be said for researchers with the great effort devoted to the high-P_j disciplines, compared to the other researchers (Figure 4). Although they were few, some of the other researchers moved up in the ranking more than researchers publishing papers mainly in the low-P_j disciplines (see the upper envelope of the plots in Figure 4a and b). By examining the details of the publication data, we found such researchers publishing in disciplines with P_j well below its average, but the P_j was insignificant in light of a large confidence interval; in some cases, the discipline was sparsely populated and had been excluded in Section 3.1 (e.g., Geoscience in the Web of Science, and Neuroscience and Arts and Humanities in Scopus). A small minority of the other researchers moved down in the ranking more than researchers publishing papers mainly in the high-P_j disciplines (see the lower envelope of the plot in Figure 4a). They were publishing researchers in disciplines with statistically insignificant P_j well above its average (e.g., Microbiology in the Web of Science). These results suggest that, in determining individual rankings, it is crucial to determine P_j with accuracy in every discipline.

3.4 Challenges for the future

The most compelling work is to ensure a sufficient amount of publication data. In our analysis to determine the minimal sample size required for accurate determination of P_j, we found that the amount of data did not meet the requirements in every discipline. This has prevented us from making a firm conclusion about the disciplinary differences observed in P_j and using the normalized index DWPP for performance evaluation. We are now preparing to share information with other collaborative research institutes in Japan with the aim of increasing the amount of available data. Because the allocation of an individual's publication efforts to multiple disciplines is crucial for P_j determination, name disambiguation is an important task in the data collection. Furthermore, the data should include relevant information, such as position and affiliation, because recently researchers tend to work at more than one institution with different positions. There are a few more details to be studied: selecting a proper time period for the analysis and disciplinary classification, including document types other than article and review, and defining the authors to be included in the allocation of paper count.

There remains room for improvement in establishing a reliable index. First, we recognize that there is an intrinsic inaccuracy in the assumption on the allocation of an individual's publication efforts to multiple disciplines (see assumption (3) in Section 2.2); the proportional allocation is subject to an even difficulty of publication among disciplines at the individual level, which contradicts the disciplinary variation in P_j applied to normalize the number of papers. Although the inaccuracy was assumed to be ignorable in the results of this study, this is a future issue to be solved. Another important assumption we made regards the negligible effects of non-publishing researchers on the P_j calculation. According to [6], non-publishing researchers, the potential authors as they put it, are supposed to be considered in the quantification of publication productivity; their population in each discipline can be estimated with the Waring distribution, which is a probability model used for estimating zero-frequency from zero-truncated data. This could be a potential problem when applying this study's method to the publication data of other research institutions, which cover diverse disciplines of the arts and sciences. To increase the usefulness of the method, the applicability of such a probability model to our study should be examined in the future.

4 Conclusions

We proposed a method for the normalization of the number of papers to overcome the difficulties in fair comparison of researcher's publication productivity over diverse disciplines. In our analysis of the publication data of Kyutech indexed in the Web of Science and Scopus, similar variations were found in the calculated values of publication productivity P_j in the well-populated disciplines; for example, the P_j values of Engineering and Computer Science were statistically lower than the average in common. The normalization resulted in a relative increase in the total number of papers in disciplines with low publication productivity and a decrease in disciplines with high publication productivity. We further investigated the minimal sample size required for accurate determination of P_j. When all the Kyutech data were taken as the parent population and bootstrap calculations were made, changing the sample size little by little, it was suggested that the sample size should be greater than 200 for the Web of Science data and greater than 300 for the Scopus data for determining P_j with accuracy. Because the amount of data did not meet the requirements in every discipline, the confidence intervals of P_j are considered to have been overestimated in width in the present study. With the normalized index DWPP, compared to the naïve index of number of papers, researchers publishing papers in disciplines with low publication productivity enjoyed a marked rise in ranking. The change of a measure mostly affected the rankings of middle-ranking researchers; the ranking of researchers with extremely high or low productivity were less affected. In summary, the database selection did not matter for the results in the analysis of Kyutech's publication data. Note that the wider coverage of journals in Scopus enabled us to determine the rankings of more researchers.

The above-mentioned wide confidence intervals of P_j make it difficult to make a precise analysis for its disciplinary differences and creates uncertainty about the resulting normalized index. Therefore, we are planning to collaborate with other research institutions, to collect more publication data. For reliable analysis, name disambiguation of papers is crucial because the total publication effort ratio of researchers should be determined in each discipline. Our final aim is to conduct performance evaluation of research at various levels, such as the level of the whole institution, department, and individual, in a way that is as unbiased as possible, across disciplinary boundaries.

Acknowledgments

We are grateful to the staff of the General Affairs Division of Kyushu Institute of Technology for their contribution in conducting name disambiguation for papers in the Web of Science. We also thank the staff of the Research Cooperation Division of the institute for their support in collecting the publication data of Elsevier.

References

[1] H. Fukunari, ”Practical and applied bibliometric analysis for research strategy,” Journal of Information Processing and Management, 57(6), pp.376–386, 2014 (in Japanese).

[2] L. Butler, ”Assessing university research: a plea for a balanced approach,” Science and Public Policy, 34(8), pp.565–574, 2007.

[3] N. Furubayashi, ”Beyond bibliometrics. Research analytics and evaluation based on citation data,” Information Science and Technology Association, 64(12), pp.520–526, 2014 (in Japanese).

[4] Thomson Reuters, ”InCites Indicators Handbook,” pp.28, 2014.

[5] Elsevier, ”SciVal Metrics Guidebook,” pp.95, 2014.

[6] T. Koski, E. Sandström, and U. Sandström, ”Towards field-adjusted production: Estimating research productivity from a zero-truncated distribution,” Journal of Informetrics, 10(4), pp.1143–1152, 2016.

[7] L. Butler, ”Explaining Australia’s increased share of ISI publications – the effects of a funding formula based on publication counts,” Research Policy, 32(1), pp.143–155, 2003.

[8] J. W. Schneider, ”Publications or citations – does it matter? Beneficiaries in two different versions of a national bibliometric performance model, an existing publication-based and a suggested citation-based model,” Proceedings of ISSI 2015 Istanbul, pp.477–488, 2015.

[9] L. Colledge, ”Snowball metrics recipe book,” Amsterdam: Snowball Metrics Program Partners, pp.110, 2014.

[10] Research project for the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, ”Survey analysis report on performance evaluation of researchers,” 2015 (in Japanese).

[11] L. R. Vinluan, ”Research productivity in education and psychology in the Philippines and comparison with ASEAN countries,” Scientometrics, 91(1), pp.277–294, 2011.

[12] J. S. Barrot, ”Research impact and productivity of Southeast Asian countries in language and linguistics,” Scientometrics, pp.1–15, 2016; doi:10.1007/s11192-016-2163-3.

[13] D. Hicks, ”Performance-based university research funding systems,” Research policy, 41(2), pp.251–261, 2012.

[14] G. Sivertsen, ”Publication-based funding: the Norwegian model,” Research Assessment in the Humanities, Springer, pp.79–90, 2016.

[15] J. E. Hirsch, ”An index to quantify an individual’s scientific research output,” Proceedings of the National academy of Sciences of the United States of America, pp.16569–16572, 2005.

[16] G. Abramo and C. A. D'Angelo, "How do you define and measure research productivity?," Scientometrics, 101(2), pp.1129–1144, 2014.

[17] A. Schubert, W. Glänzel, and T. Braun, "Against absolute methods. Relative scientometric indicators and relational charts as evaluation tools," In: A. van Raan (Ed.), Handbook of Quantitative Studies of Science and Technology, North-Holland Publishing Company, pp.137–176, 1988.

[18] P. Mongeon and A. Paul-Hus, "The journal coverage of Web of Science and Scopus: a comparative analysis," Scientometrics, 106(1), pp.213–228, 2016.

[19] L. Egghe, R. Rousseau, and G. Van Hooydonk, "Methods for accrediting publications to authors or countries: Consequences for evaluation studies," Journal of the American Society for Information Science, 52(2), pp.145–157, 2000.

[20] J. Wang and Y. Sakurai, "An introduction to bootstrap methods," Kyoritsu Shuppan Co., LTD, pp.236, 2011 (in Japanese).

[21] R Development Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing," Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org, 2005.