Differential Topic Selections and Wording Behaviors among Funded Environmental Projects with Stakeholders

Invited Paper

Zhongyu Yao*, He Zhang*, Tianhang Chen*, Yuchen Wang*, Weidun Xie* and Ka-Chun Wong*
*Department of Computer Science
City University of Hong Kong
Kowloon Tong
Hong Kong SAR
Email: kc.w@cityu.edu.hk

2024 Conference on AI, Science, Engineering, and Technology (AIxSET) | 979-8-3503-9099-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIXSET62544.2024.00023

***Abstract*—Environmental research often attracts significant public attention for its relevance to society sustainability. Enormous resources (including research funding) have been allocated to the development of environmental research in the past years. To assess its effectiveness and relevance to stakeholders, we have conducted a data science study on the research projects funded by Environment and Conservation Fund (ECF) more than ten years. The results indicate that there are preferred project topics and researchers. In addition, sentiment analysis reveals the subtle ECF research project writing strategy trends which reflect the pragmatic aspects of topic selections and wording behaviors. Comparing to the stakeholders (i.e. government, mass media, and non-government organization), we observe interesting insights into the topic and wording differences among the relevant parties in a differential manner.**

I. Introduction

Environmental research is linked to different societal aspects such as sustainability, well-being, and economics [1]. Diverse stakeholder engagement has been recognized as the keys to successes in environmental research [2]. People have been exploring different approaches for solving diverse environmental issues such as climate change, ecological conservation, waste management, air/water quality, and urban pollution (e.g. light and noise). To address such diverse topics, trans-disciplinary collaboration has been discussed and conjectured as one of the promising solutions in the past [3]. Environmental research meta-science has been well-recognized as an important and highly desired area in ecology since 2015 [4], [5]. Efforts have been developed on environmental research meta-science to bridge the academic research with practitioners and policymakers; for instance, Sutherland et al. have developed an environmental evidence database tool for evidence synthesis and research decision-making [6]. Walsh et al. have also developed a classification taxonomy system to summarize the barriers and factors in environmental conservation practices, informing researchers and policy makers. Given the diverse environmental topics, challenges have been recognized and addressed to identify missing environmental research topics [7].

Failures in environmental research funding allocation can result in severe ecological consequences as evidenced by the recent study on the conservation studies on primates [8]. Indeed, the misallocation of environmental research funding can lead to impacts on ecological research directions; for instance, after analyzing the climate-related research grants from 333 donors around the world, Overland and Sovacool reported that the natural and technical science projects received 770% more funding than the social science counterparts [9]. Around the world, different governments have funded a multitude of environmental research projects for tackling those challenges. In particular, the Hong Kong government has invested billions of dollars in this aspect through the Environment and Conservation Fund (ECF). However, its effectiveness and relevance to stakeholders remains as an open topic.

II. Related Work

In the past, there is a myriad of works on environmental research meta-science; for instance, Burton et al. have identified that the US Environmental Protection Agency had the tendency not to fund the research on the impacts of releases of chemicals to the environment [10]. Ebi et al. have also identified that the current environmental funding policies in the U.S. and Europe were not well-aligned with the international agreements (the UN 2030 Agenda for Sustainable Development; the Sendai Framework for Disaster Risk Reduction; and the Paris Agreement under the United Nations Framework Convention on Climate Change) [11]. From the social science perspective, the environmental research funding is further complicated by different stakeholders such as NGO and funders [12]; for instance, a study on Florida has revealed that the environmental research spending can be influenced by the environmental pressure and budgetary politics [13]. Jolibert and Wesselink have also revealed the influence of stakeholders on biodiversity conservation progress [14]. In UK, Philipson et al. have also identified the importance of stakeholder engagement and knowledge exchange with diverse patterns [2]. Shahzad et al. even proposed to put stakeholders' pressures into the development of environmental practices

2024 Conference on AI, Science, Engineering, and Technology (AIxSET)

979-8-3503-9099-5/24/$31.00 ©2024 IEEE
DOI 10.1109/AIxSET62544.2024.00023

[15]. In Hong Kong, Xu et al. have identified the role of political institutions and players in environmental funding decisions [16]. Nonetheless, the study is limited to 19 non-recurrent environmental spending projects. On the other hand, there are declines in stakeholder trust on the government policy in environmental research [17]. Differential opinions on environmental policies (e.g. waste management) are also observed among different stakeholders [18]. Nonetheless, diverse environmental issues are still prevalent in Hong Kong such as air pollution [19], light pollution [20], noise pollution [21], landfill [22], and water quality [23]. Overall, there are ongoing debates on the interactions between stakeholders and environmental research projects. There is a need to delineate the commonalities and differences between them. Therefore, we propose to conduct a comprehensive data science study on hundreds of research projects for comprehensive insights into publicly funded environmental research in Hong Kong.

III. Methodology

A. ECF Research Project Data Collection

In August 2020, we have customized and implemented Python web-crawler scripts to collect the most updated "Environmental Research, Technology Demonstration and Conference Projects" information from the ECF web-page[1]. For the projects funded before 2014, the ECF archive web-page[2] has also been visited and crawled by customized Python web-crawler scripts. For each project, we have collected as much data as possible. Specifically, the project number, project title, principle investigator information, project funding amount, project duration, project status, project scope, and project outcomes have been collected, resulting in 382 project entries. However, we note that the project information before 2004 are highly heterogeneous without any regular pattern for fair and justified data formatting. Therefore, we have removed the project information before 2004, resulting in the final 292 project entries between 2004 and 2019.

B. EPD Milestone Text Data Collection

In August 2020, we have visited the Environmental Protection Department (EPD) website[3] to manually collect the milestone text data. Since the ECF research project data collected are ranged from 2004 to 2020, we also limit ourselves to collect the EPD milestone text data from 2004 to 2017. At the time of collection, the milestone text data after 2017 are still not available.

C. HKFP News Text Data Collection

In August 2020, we have adopted the ParseHub software to collect the "Environment & Health" category news data from the main website of Hong Kong Free Press (HKFP)[4]. In particular, we have collected the news title and date for each news entry as much as we can, resulting in 1635 news entries between 2015 and 2020.

[1]https://www.ecf.gov.hk/en/approved/artdp.html
[2]https://www.ecf.gov.hk/en/archives/archive_artdp.html
[3]https://www.epd.gov.hk/epd/english/resources_pub/history/history_hkep.html
[4]https://hongkongfp.com/category/topics/health-eco/

D. WWF Featured Story Text Data Collection

In August 2020, we have adopted the ParseHub Software to collect the "Feature stories" section from the main website of World Wide Fund for Nature (WWF) Hong Kong[5]. In particular, we have collected the news title, subtitle, and date for each story as much as we can, resulting in 278 featured story entries between 2010 and 2020.

E. Natural Language Processing for All Text Data

For all text data collected (e.g. project titles, milestone text, news text, story text), we have designed and followed the same protocol for natural language processing with the "tm" package (version 0.7-7) in R (version 3.6.2). Specifically, we have converted all text encoding into the "UTF-8" encoding for uniform processing. After that, we have converted all words into lowercase for semantic redundancy removal. Similarly, all non-semantic entities such as English stop-words, numbers, and punctuation marks have been removed. Lastly, all trailing white-spaces and tabs are trimmed out. For each text data type, we have also converted the text data into sentence-term matrices for data science analysis in the subsequent sections. If two matrices need to be compared, we have chosen to linearly normalize the term frequency to sentence-wide term frequency; for example, if we have ten words in one sentence, we divide each term frequency by ten.

IV. Results

In total, we have 292 ECF research project entries between 2004 and 2019. As the ECF scheme has been continuously developed in the past years, the number of funded projects should have been increased year-by-year. Basic funding statistics have been computed and described.

A. Basic Funding Statistics

A histogram is drawn to visualize the ECF research project counts across different years as shown in Figure 1. It is observed that there is a steady increase in the number of funded projects with some drops across years.

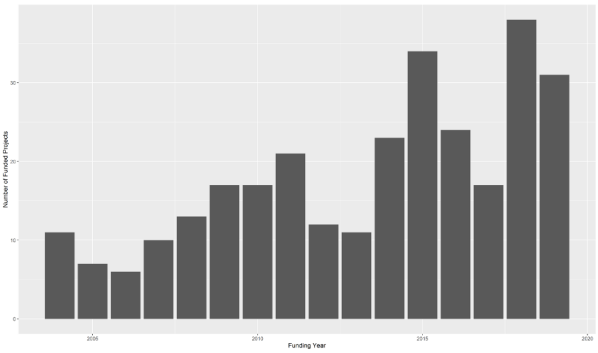

Fig. 1. Histogram for ECF Research Project Counts since 2004.

[5]https://www.wwf.org.hk/en/news/featuredstories/

On the other hand, we note that project funding can be varied across different projects; the number of projects may not be meaningful. Therefore, we have explored the relationships between project funding and project duration. In particular, we note that the funding situation could be varied across different years subject to the financial status of the government. Therefore, we have separated the projects according to the funding year in different colors as depicted in Figure 2. From the figure, we can observe different characteristics of the ECF projects across different years. Specifically, we can observe that the number of project months is positively correlated to the project funding amount. Secondly, we also observe that the projects are heavily centered around 24 months, Thirdly, we also observe that the very large funding amount can be seen around the year 2010. The early projects (in red) and recent projects (in purple) were not very well-funded.

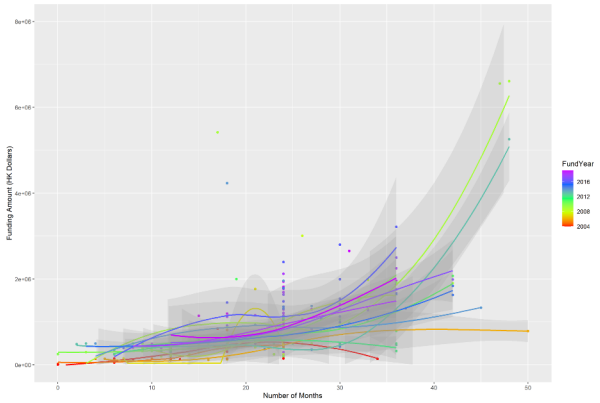

Fig. 2. Dot Plots on Project Funding versus Project Duration in Months. Different funding years are denoted in different colors. The fitted curves are drawn using the R command "geom_smooth()" where 95% intervals have also been drawn for each fitted line in grey colours.

B. Project Writing Sentiment Analysis

To evaluate the human factors in environmental research, we have conducted sentiment analysis on both the project titles and abstracts. Specifically, for each word, we have computed three sentiment scores: the sentiment score by the "sentimentr" package, the GI-dictionary-based sentiment score by the "SentimentAnalysis" package, and the QDAP-dictionary-based sentiment score by the "SentimentAnalysis" package in R. The sentiment results are visualized in Figures 3 and 4. Interestingly, we can observe that the sentiment scores of the project abstracts are always positive while those of the project titles fluctuate in recent years. It results in two key observations: (1) Such results reflect that the project abstract writing styles are more inclined towards positive words than the project title writings. It hints us that contrastive project writing could influence the project funding successes. (2) In the past, positive project titles were dominant. In recent years, it appears that critical or negative project titles may be needed, although the project abstract writing styles still remain positive.

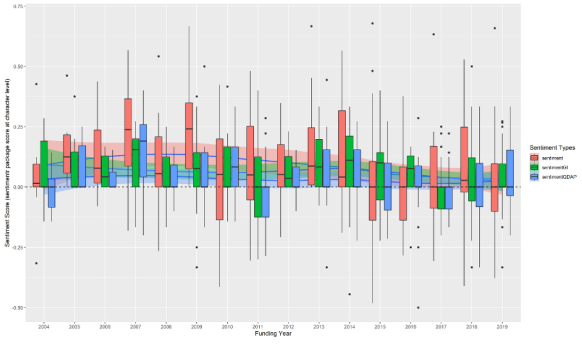

Fig. 3. Sentiment Analysis on Project Titles based on Three Sentiment Scoring Systems in Different Colors.

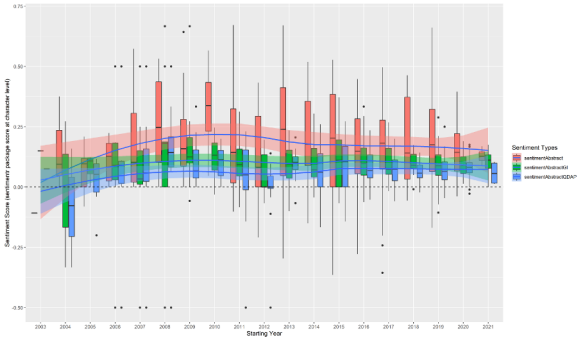

Fig. 4. Sentiment Analysis on Project Abstracts based on Three Sentiment Scoring Systems in Different Colors.

C. LDA Topic Modeling

In this section, we propose to apply Latent Dirichlet Allocation (LDA) topic modeling on the ECF project titles. To estimate its feasibility, we have plotted the word cloud on the project titles of the previously collected 292 "Environmental Research, Technology Demonstration and Conference Projects" between 2004 and 2019 from the ECF website based on the "wordcloud" software package in R as shown in Figure 5.

From the figure, we can observe that the ECF scheme is very related to the environmental research studies in Hong Kong since most of the project titles contain the keywords "hong" and "kong". In other aspects, we also observe that the keyword "conference" does not appear to be very large; it indicates that the conference projects may have been underutilized in the past. On the other hand, it seems that the project topics related to waste are favored by the ECF scheme in the past years. All those insights suggest that there are heavily weighted topics within the funded ECF projects. Therefore, extra data science techniques are needed for gaining insights into all available ECF projects.

In particular, we are also interested in the evolution of project topics because the project investigators can have different interests across different years. Therefore, we aim at developing the temporal data techniques in topic modeling

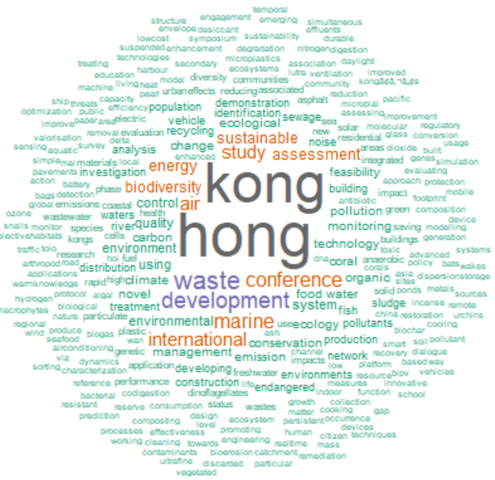

Fig. 5. Word Cloud on ECF Project Titles

across different years. It necessitates Latent Dirichlet Allocation (LDA) topic modeling.

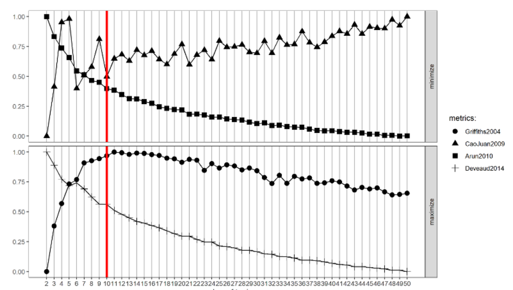

Fig. 6. Selection for the Optimal Number of LDA Topics" based on Bayesian model selection [24], density-based method for adaptive model selection [25], symmetric KL-Divergence of salient distributions [26], and latent concept modeling [27]. The red vertical line indicates that ten topics are selected.

To determine the number of latent topics for LDA, we have adopted and executed four topic number selection methods on all project titles, including Bayesian model selection [24], density-based method for adaptive model selection [25], symmetric KL-Divergence of salient distributions [26], and latent concept modeling [27]. Taking the minimum of its selection score summation, we choose the optimal number of LDA topics as shown in Figure 6. It is observed that the optimal number of LDA topics should be 10 which is fixed in the subsequent sections.

Given the 10 LDA topics, we proceed to stratify the project titles into the 10 LDA topics according to the gamma value of LDA modeling. In particular, we have counted the fraction number of projects as well as the total project funding amount under different LDA topics between 2004 and 2019 as depicted in Figures 7 and 8. Interestingly, we can observe that the research projects are mostly centered around specific themes such as ecological conservation, waste management, water pollution, green energy, and air quality. Comparing it to the EPD milestones collected, it is obvious that the research on noises (also known as sound pollution) and light pollution are under-represented. On the other hand, we also observe that the marine conservation topics are highly represented among the project titles. Such a view is further substantiated after we have broken down the project funding into individual project funding across the years as shown in Figure 8 where the green colors denote the marine conservation topics.

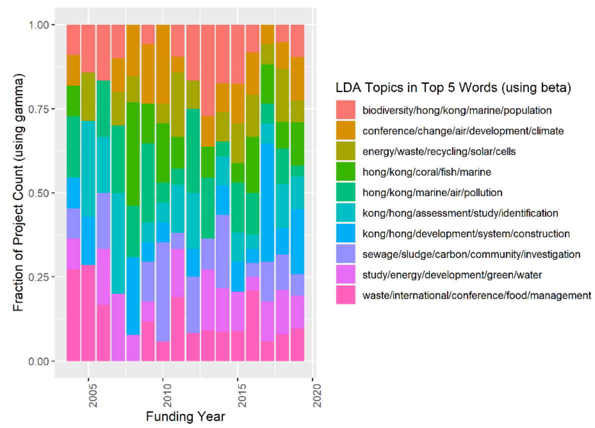

Fig. 7. Fraction of Project Count under Different LDA Topics

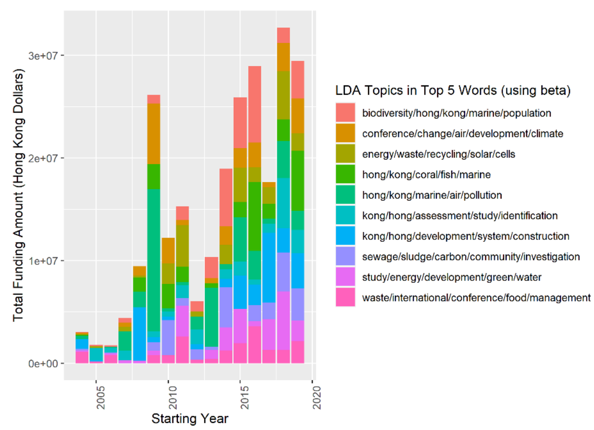

Fig. 8. Total Project Funding Amount under Different LDA Topics

Interestingly, the word enrichment analysis based on EPD milestones reveals that the existing research projects are highly enriched in the related topics of coral and other marine conservation. On the other hand, we can also observe that vehicles, power, landfill, and light pollution are highly depleted in the research projects. Such a different wording behavior observation provides insights on how to align future research projects with the EPD milestones in a complementary manner.

D. Differential Topic Comparison with Government

From the previous analysis, it is interesting that there are thematic focuses among the research projects. However, it

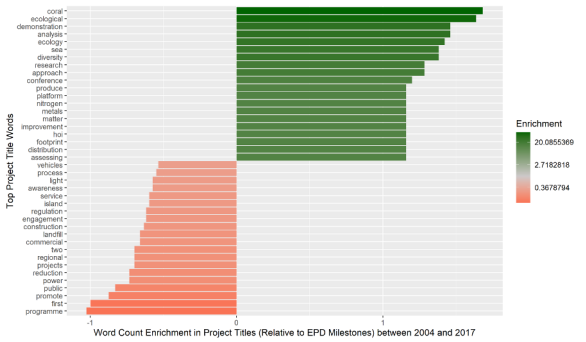

Fig. 9. Top 20 Enriched and Depleted Project Title Words Ranked by Word Count Enrichment Scores (Relative to EPD Milestones between 2004 and 2017). The enrichment score is computed as the term frequency ratio (in log10 scale) between project title and milestone text.

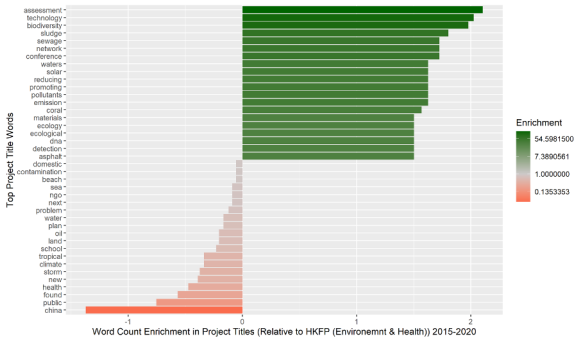

Fig. 10. Top 20 Enriched and Depleted Project Title Words Ranked by Word Count Enrichment Scores (Relative to HKFP media reports between 2015 and 2020). The enrichment score is computed as the term frequency ratio (in log10 scale) between project title and media report text.

actually makes sense since the proposed topics could be limited to the available researcher expertise in Hong Kong. Nonetheless, as funded by the government, it could still be interesting to see the textual divergence between the funded projects and the government policy. Therefore, we have conducted a differential topic comparison between the project titles and the EPD milestone words. After the sentence-level normalization (as described in the previous section), we have computed the ratios between the term frequencies of project titles and the term frequencies of EPD milestone text as the word count enrichment scores as depicted in Figure 9.

E. Differential Topic Comparison with Media

Arguably, media plays a significant role in monitoring environmental issues nowadays. Therefore, we are also interested in the topic differences between the media and the publicly funded research projects and conducted a differential text analysis between the project titles and the media report titles from Hong Kong Free Press (HKFP). After the sentence-level normalization (as described in the previous section), we have computed the ratios between the term frequencies of project titles and the term frequencies of HKFP report titles as the word count enrichment scores as depicted in Figure 10.

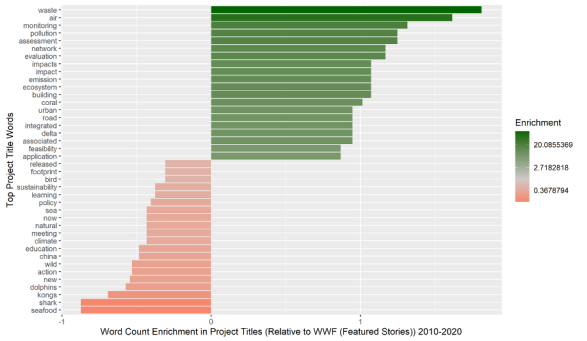

Fig. 11. Top 20 Enriched and Depleted Project Title Words Ranked by Word Count Enrichment Scores (Relative to WWF(HK) Featured Story Titles between 2010 and 2020). The enrichment score is computed as the term frequency ratio (in log10 scale) between project title and WWF story title.

The figure shows that the research project titles are more enriched in academic terms (e.g. assessment, technology, DNA, and biodiversity) than the media reports. In contrast, the media text usually contains layman terms (e.g. beach, sea, water, oil, land, and climate). It gives us insight into the observation that the research project titles are written in the technical and academic style while the media reports are written in the generally accessible style. Such a difference in writing styles imposes difficulties in directly comparing the semantic similarities, given the limited and sparse term frequencies available. On the other hand, it is also interesting to note that political words are also enriched in the media texts (e.g. NGO, Public, and China). Such findings could reflect another strategic aspect of media.

F. Differential Topic Comparison with NGO

To monitor and complement the government efforts in environmental protection, there are myriads of NGOs for environmental research and protection. As previously observed in the comparisons with media, the keyword “NGO” is well-enriched in the media. Therefore, it could be interesting if we can examine the differences between the “Feature stories” section from the main website of World Wide Fund for Nature (WWF) Hong Kong and the publicly funded research project titles. Therefore, we have also conducted a differential text analysis between them. After the sentence-level normalization (as described in the previous section), we have computed the ratios between the term frequencies of project titles and the term frequencies of WWF featured story titles and subtitles as the word count enrichment scores as depicted in Figure 11.

In general, we can observe that the title words “waste” and “air” are more prominent in the research project titles than WWF. It is actually expected since WWF focuses on the ecological conservation. However, it is surprising that the title term “coral” is still enriched in the research project titles compared to the WWF featured stories. It may imply that the research projects are heavily weighted towards coral studies. On the other hand, it is also interesting to observe that “shark”, “dolphins”, “seafood”, and “bird” are inclined

towards the WWF side; it not only reflects the thematic focuses of WWF in Hong Kong, but also shed lights on the research project topics which are still not well-proposed for publicly funded research under ECF. Another interesting aspect could lie in the advocating wordings in WWF such as "action", "education", "now", and "footprint". Such findings could also reflect another strategic aspect of the NGO.

V. Conclusions

In summary, we have conducted a data science study on the publicly funded research projects in the context of environmental research. In particular, we have performed different data collection and processing steps on the research project information from ECF. Based on the data statistics, we observed that there were specific trends and characteristics in the project topics and funding. The sentiment analysis revealed that positive sentiments were consistently observed in project abstracts across the years while diverse sentiments were observed in project titles, especially for recent years. The LDA modeling also revealed that there were specific topic preferences in those research projects; for instance, 5.4% (16/292) of those projects contain "coral" as one of the title words while only 0.39% (42/10793) of the NERC(UK) standard projects contain "coral" as one of the title words [6]. Such a thirteen-fold difference is interesting and reflects the thematic focuses of the environmental research projects in Hong Kong. Comparing the research project keywords with the stakeholders (government, mass media, and NGO), we also observed interesting insights into their commonalities and differences in environmental research. Coherent with the past study in Germany [28], we foresee that future research funding should be prioritized towards environmental technology.

References

[1] C. Campbell, E. Lefroy, S. Caddy-Retalic, N. Bax, P. Doherty, M. M. Douglas, D. Johnson, H. P. Possingham, A. Specht, D. Tarte *et al.*, "Designing environmental research for impact," *Science of the Total Environment*, vol. 534, pp. 4–13, 2015.

[2] J. Phillipson, P. Lowe, A. Proctor, and E. Ruto, "Stakeholder engagement and knowledge exchange in environmental research," *Journal of environmental management*, vol. 95, no. 1, pp. 56–65, 2012.

[3] C. Pohl, "Transdisciplinary collaboration in environmental research," *Futures*, vol. 37, no. 10, pp. 1159–1178, 2005.

[4] G. B. Stewart and C. H. Schmid, "Lessons from meta-analysis in ecology and evolution: the need for trans-disciplinary evidence synthesis methodologies." *Research synthesis methods*, vol. 6, no. 2, pp. 109–110, 2015.

[5] G. Stewart and J. Ward, "Meta-science urgently needed across the environmental nexus: a comment on berger-tal et al." *Behavioral Ecology*, vol. 30, no. 1, pp. 9–10, 2019.

[6] W. J. Sutherland, N. G. Taylor, D. MacFarlane, T. Amano, A. P. Christie, L. V. Dicks, A. J. Lemasson, N. A. Littlewood, P. A. Martin, N. Ockendon *et al.*, "Building a tool to overcome barriers in research-implementation spaces: The conservation evidence database," *Biological Conservation*, vol. 238, p. 108199, 2019.

[7] A. P. Christie, T. Amano, P. A. Martin, S. O. Petrovan, G. E. Shackelford, B. I. Simmons, R. K. Smith, D. R. Williams, C. F. Wordley, and W. J. Sutherland, "The challenge of heterogeneous evidence in conservation," *bioRxiv*, p. 797639, 2019.

[8] J. Junker, S. O. Petrovan, V. Arroyo-RodrÍguez, R. Boonratana, D. Byler, C. A. Chapman, D. Chetry, S. M. Cheyne, F. M. Cornejo, L. CortÉs-Ortiz *et al.*, "A severe lack of evidence limits effective conservation of the world's primates." *BioScience.*, 2020.

[9] I. Overland and B. K. Sovacool, "The misallocation of climate research funding," *Energy Research & Social Science*, vol. 62, p. 101349, 2020.

[10] G. A. Burton Jr, R. Di Giulio, D. Costello, and J. R. Rohr, "Slipping through the cracks: why is the us environmental protection agency not funding extramural research on chemicals in our environment?" 2017.

[11] K. L. Ebi, J. C. Semenza, and J. Rocklöv, "Current medical research funding and frameworks are insufficient to address the health risks of global environmental change," *Environmental Health*, vol. 15, no. 1, p. 108, 2016.

[12] M. Chewinski and C. Corrigall-Brown, "Channeling advocacy? assessing how funding source shapes the strategies of environmental organizations," *Social Movement Studies*, vol. 19, no. 2, pp. 222–240, 2020.

[13] X. Wang, "Exploring trends, sources, and causes of environmental funding: A study of florida counties," *Journal of Environmental Management*, vol. 92, no. 11, pp. 2930–2938, 2011.

[14] C. Jolibert and A. Wesselink, "Research impacts and impact on research in biodiversity conservation: The influence of stakeholder engagement," *Environmental Science & Policy*, vol. 22, pp. 100–111, 2012.

[15] M. Shahzad, Y. Qu, A. U. Zafar, X. Ding, and S. U. Rehman, "Translating stakeholders pressure into environmental practices: The mediating role of knowledge management," *Journal of Cleaner Production*, vol. 275, p. 124163, 2020.

[16] J. Xu, X. Wang, and H. Xiao, "How environmental bureaucrats influence funding legislation: an information processing perspective," *Environmental Politics*, pp. 1–22, 2020.

[17] R. M. Walker and P. Hills, "Changing dimensions of trust in government: An exploration in environmental policy in hong kong," *Public Administration and Development*, vol. 34, no. 2, pp. 123–136, 2014.

[18] C. Wan, G. Q. Shen, and S. Choi, "Differential public support for waste management policy: The case of hong kong," *Journal of Cleaner Production*, vol. 175, pp. 477–488, 2018.

[19] Z. Ai, C. Mak, and H. Lee, "Roadside air quality and implications for control measures: A case study of hong kong," *Atmospheric environment*, vol. 137, pp. 6–16, 2016.

[20] C. S. J. Pun, C. W. So, W. Y. Leung, and C. F. Wong, "Contributions of artificial lighting sources on light pollution in hong kong measured through a night sky brightness monitoring network," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 139, pp. 90–108, 2014.

[21] H. Li, C. Chau, and S. Tang, "Can surrounding greenery reduce noise annoyance at home?" *Science of the total environment*, vol. 408, no. 20, pp. 4376–4384, 2010.

[22] X.-W. Chen, J. T.-F. Wong, W.-Y. Mo, Y.-B. Man, C. W.-W. Ng, and M.-H. Wong, "Ecological performance of the restored south east new territories (sent) landfill in hong kong (2000–2012)," *Land Degradation & Development*, vol. 27, no. 6, pp. 1664–1676, 2016.

[23] X. Zhang, Q. Wang, Y. Liu, J. Wu, and M. Yu, "Application of multivariate statistical techniques in the assessment of water quality in the southwest new territories and kowloon, hong kong," *Environmental monitoring and assessment*, vol. 173, no. 1-4, pp. 17–27, 2011.

[24] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[25] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive lda model selection," *Neurocomputing*, vol. 72, no. 7-9, pp. 1775–1781, 2009.

[26] R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2010, pp. 391–402.

[27] R. Deveaud, E. SanJuan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval," *Document numérique*, vol. 17, no. 1, pp. 61–84, 2014.

[28] N. Weinberger, J. Jorissen, and J. Schippl, "Foresight on environmental technologies options for the prioritisation of future research funding lessons learned from the project roadmap environmental technologies 2020," *Journal of cleaner production*, vol. 27, pp. 32–41, 2012.

[6] http://gotw.nerc.ac.uk/ in Sep 2020