

Does data curation matter in citation and co-citation analysis? Evidence from a top service journal

Mehmet Ali Koseoglu[†]

Hasan Evrim Arici

Nagihan Cakmakoglu Arici

Mehmet Ali Koseoglu[†]
College of Management
Metropolitan State University
1501 Hennepin Ave.
Minneapolis
MN 55403-1897
USA

Hasan Evrim Arici^{*}
Faculty of Tourism
Kastamonu University
Kastamonu
Turkey

and

EU Business School
European University Munich
Theresienhöhe 28
Munich
Germany

Nagihan Cakmakoglu Arici[§]
Faculty of Tourism
Kastamonu University
Kastamonu
Turkey

E-mail

[†] koseoglu.mehmetali@gmail.com

^{*} hasan_evrilm.arici@euruni.edu

[§] heraas1983@yahoo.com

^{*} (Corresponding Author)

Bibliometric scholars have primarily evaluated massive data without refining any potential typing and/or spelling errors, resulting in two constraints: misinterpretation of findings and misleading future research in the knowledge domain. Thus, this study aims to introduce the data curation approach in order to reduce these restrictions. Utilizing a renowned service journal (Journal of Service Research) as the study sample, we first acquired all published papers and then constructed raw and clean datasets. We ran citation and co-citation analyses on these datasets separately. Our investigation reveals that clean data yielded more trustworthy and valid results than raw data with redundant references. This study provides an answer to how and why data in bibliometric analysis needs to be cleaned. It thus contributes to the literature by suggesting a new route for scholars to improve the accuracy and reliability of their bibliometric findings.

Keywords: Quantitative analysis, Cluster analysis, Longitudinal data analysis.

Introduction

To uncover the intellectual structure of a given domain, scholars adopt citation and co-citation analysis as a quantitative bibliometrics approach that reduces or eliminates the potential drawbacks of using a qualitative approach relying on mostly subjective norms and interpretations of independent scholars in scientific research (Köseoglu, 2020). In doing so, scholars improve the document citation and co-citation analysis

as complementary to the qualitative investigations as a part of the bibliometric analysis (Zupic & Čater, 2015). The citation and co-citation analyses clarify the intellectual connection of the knowledge domain by utilizing mathematical and statistical techniques to use citations or references of published documents as data (Hou, Yang, & Chen, 2018). Scholars have mostly adopted these approaches in their bibliometric studies within the last decade (Batistič, Černe, & Vogel, 2017; Hota, Subramanian, & Narayanamurthy, 2019; Köseoglu, Mehraliyev, Aladag, & King, 2021; Reale, Khelfaoui, Montiglio, & Gingras, 2020). These attempts have recently led to the launching of several software programs (such as VOSviewer) that significantly ease the analysis of a huge volume of data within a limited period. However, this procedure undoubtedly also brings about some significant disadvantages for the scholarship. Probably of utmost importance is to submit the raw data, including misspelled and/or duplicated publications or references, into the software without subjecting the data set to a cleaning procedure. The broad range of bibliometric studies, in the form of both a citation and co-citation analysis, rely on reviewing a large volume of data by ignoring the cleaning process, which can result in questioning of findings and aftereffects (e.g., Boyack & Klavans, 2010; Cunill, Salvá, Gonzalez, & Mulet-Forteza, 2019; Hota, Subramanian, & Narayanamurthy, 2019; Lazzeretti, Capone, & Innocenti, 2017; Mody, Hanks, & Cheng, 2021; Shin & Perdue, 2019). This can vitally affect the findings' reliability and validity, which results in questioning the findings of bibliometric studies. Due to the potential limitation of using the raw data for citation and co-citation analysis, the agenda generated from the uncleaned data may mislead future research about the study domain. This is one of the problems that bibliometric scholars have answered to warrant the reliability of findings. Thus, this current research aims to answer the research question "does data curation matter in bibliometric analysis?"

Overcoming this problem requires data cleaning to ascertain if there are some misuses and differences in the spelling of an author's name, the title of the study, or journal that need to be eliminated within the dataset before submitting it to the software for analysis (Köseoglu et. al., 2021). In other words, a bibliometric analysis first requires an explicit clean-up of all the inconsistencies in references used by previous publications that the dataset includes. Given the vital role of data cleaning in both citation and co-citation analyses, an application of data curation which refers to thoroughly preparing a dataset for further bibliometric analysis, is necessary. Such an implementation can elucidate how a bibliometric analysis can provide more clear, reliable, valid, and accurate findings that can lead to future research in the knowledge domain. Therefore, we have tested how data curation influences citation and co-citation analysis findings.

Given the paucity of articles using data cleaning in the field of bibliometric analysis including citation and co-citation (Köseoglu, Okumus, Dogan, & Law, 2019), the value of the data curation approach in the bibliometric analysis literature is manifold. Bibliometric studies review comprehensive literature, clarifying and revealing the most cited and influential studies within the knowledge domain (Ellegaard & Wallin, 2015; Glänzel, Schubert, & Czerwon, 1999; Scheffler & Brunzel, 2020). As understanding state-of-the-art study outcomes in the research field are significant, review studies with citation and co-

citation analysis could serve as a valuable basis for scholars. These review studies can also present promising directions for future investigation, generating a roadmap of extant knowledge and enabling innovative ideas to improve. However, to be able to achieve all the above-mentioned outcomes, researchers must first consider getting thoroughly clean data, which does not include any misspelled or duplicated publications or references. Thus, adopting a data curation approach can contribute to the discipline's progress.

Considering this background, this present work aims to introduce and display the utility of data curation as a complementary methodological approach to remove potential flaws and inconsistencies in the dataset. As an application example, we have shown the differences between before- and after-curation data to introduce how data curation could serve as a method to improve the bibliometric analysis field in disciplines and provide complementary viewpoints to citation and co-citation methods in the literature. To do so, we have employed a top service journal: the *Journal of Service Research*.

The study, furthermore, presents the history of citation and co-citation analysis. After summarizing the methodology employed herein, we demonstrate the utility of the new approach in the form of a step-by-step explanation. Last, we discuss the utility of the newly suggested complementary approach by comparing the findings obtained from our before and after data curation analyses of the articles chosen from the referred journal for the further implications of the new approach.

Literature review

History of citation analysis

Bibliometric analysis is a quantitative method utilized to identify the progress and intellectual structure of scientific domains of research (Liu, Yin, Liu, & Dunford, 2015). In general, bibliometric analyses have been comprehensively endorsed due to their inherent objective (Garfield, 1979). Bibliometric research could examine researchers' latent intellectual connections, study trends, and global connections (Benckendorff & Zehrer, 2013; Köseoglu et al., 2021). Unless indicated otherwise, the concept of citation has been utilized synonymously with the concept of bibliographic reference. Therefore, citation analysis can somewhat reflect the examination of bibliometric references that shape a part of the mechanism of scientific connection (Nicolaisen, 2007).

Citation analysis is considered a quantitative method, which examines the citation of scholarly journals, studies, and scholars to test their quantitative features and natural laws (Daud et al., 2022; Kalita, 2016). From its infancy stage, citation analysis has been the topic of perpetual claim regarding the coherence, validity, and information content it mirrors. However, it has been comprehensively utilized by scholars as the fundamental study approach for bibliometrics and scientometrics. At this point, it has constructed its conceptual mechanism, even though there have been several challenges in implementation (MacRoberts & MacRoberts, 2018).

The historical background of citation analysis provides a significant understanding of the evolution of this analysis method. Notably, in 1955, Garfield suggested seeking

scholarly literature through citations in his Citation Indexes for Science. In 1963, he further improved the extensive citation index database, the Science Citation Index (SCI) which was considered the groundwork of citation analysis and established a meaningful basis for citation analysis research (see Garfield, 1963). Nonetheless, as a domain including cross-over research, citation analysis has a comprehensive conceptual background; specifically, studies in and before the 1950s already made important breakthroughs in formulation, improvement, and following the progress of citation analysis.

Since the mid-1950s, the citation analysis method suggested by Garfield accomplished immediate improvement and expansion. In the 1960s, Garfield revealed SCI, presenting the necessary database for citation analysis, and Price's "little SCI Big SCI" argument provided a significant conceptual basis for the improvement of citation analysis.

Chronologically, Garfield (1979) defined citation analysis as an analytical method that utilizes reference citations of scholarly published studies. By Le Pair (1988), citation analysis is potentially a clear evaluation method for those scholarly subdomains whose journals are the academic platforms for connection. Lancaster (1991) drew attention to a vital class of bibliometric research that associates with what sources researchers cite. He indicated that citation analysis copes with several factors like the topics which have the connections among the citations (for example, who cites whom, which journal has been cited by which journal, and what knowledge domain is more cited in the literature of a given field of study, etc.). In a similar vein, Diodato in 1994 defined citation analysis as a dominant method of bibliometrics that regards citations (both to and from papers) (see Diodato & Gellatly, 2013). Citation analysis research could concentrate on papers, scholars, journals, and study locations as the builder of those papers. As for the 2000s, citation analysis has exposed a different improvement trend, specifically in study subjects. These subjects comprise of study evaluation index, knowledge visualization, and citation attitude, and have been shaping the highest citation publication periods, recently being a study hotspot.

Hou's (2017) study found three important historical stages in the development of citation analysis. The first stage: was the 1950s - 1960s, the formation time, shaped by Garfield's approach to the Science Citation Index in 1955, Price's argument of the exponential development of scholarly advancement in 1963, and Price's network analysis in 1965. The second period proceeded from the 1970s to the late 1990s, called the development period, formed by three breakthroughs, namely, Small's (1973) method of reference co-citation, author co-citation analysis (White & Griffith, 1981) and (McCain, 1990) and White and McCain's map of scholarly information based on author co-citation analysis in 1998. The third period started in 2000 when the quick diffusion time of citation analysis, described by distinct study topics, involved three main hot topics: scholarly study progress index, citation attitude, and knowledge visualization.

History of the co-citation analysis

The co-citation analysis method establishes the basis for improving the citation analysis study (Hou, 2017). Co-citation has been described by several scholars like Small

(1973), Cawkell (1976), and Bellardo (1980) as a topic similarity index. Co-citation analysis experienced its important development process between the 1970–the 1990s. We realized the progress of co-citation analysis when the first vital attempt appeared in 1973 with Small (1973), who suggested the document co-citation analysis method. It was a cornerstone in the background of co-citation analysis research. Small (1973) defined it as “the frequency with which two papers are cited together”. Moreover, the author indicated that cited papers are associated together via the procedure of co-citation, and this procedure is akin to the similarity tests of the co-occurrence of words. Cawkell (1976) provides a parallel description of co-citation as a topic similarity measure, and Bellardo (1980) described co-citation, yet in a very simple expression, as a procedure whereby a researcher cited two previous studies in a new study.

As for 1981, there happened a significant improvement in the background of co-citation analysis. White and Griffith’s (1981) seminal study first proposed the approach of author co-citation analysis. In the background of co-citation analysis, 1990 was another important year. McCain’s (1990) pioneering study systematically presented the analysis approach and procedures of total cited count by scholars and reviewed the procedure of relevant study since the analysis method of total cited count by scholars was suggested in 1981. In this study, she suggested an approach of cluster analysis drawing on authors’ co-citation, factor analysis, and information visualization. In the millennium, through a multi-perspective co-citation analysis, Chen, Ibekwe-SanJuan, and Hou (2010) adopted the knowledge visualization tool CiteSpace to reveal the progress and characteristics of the study domain from 1996 to 2008.

As of the 2000s, citation analysis has come in a time of fast development. The number of published studies, citation counts, and citation frequencies all increased considerably. Drawing on the existing theories of citation analysis, document co-citation, and author co-citation, citation analysis started to progress in different avenues, involving scholarly evaluation index, citation network, knowledge visualization, mapping, and citing attitudes. Additionally, the H index research for scholarly research progress is another current approach in citation analysis.

With the development of information technology, certain research in bibliometric analysis has mostly employed software programs and computer-aided automatic labelling, particularly the wide application of text-mining techniques and cluster algorithms in analyzing the structure of knowledge domains (Hou et. al., 2018). However, these computerized systems are not meant to replace biases derived from human intervention (Hou et. al., 2018). Rather, in the data collection process, human intervention could lead to some important biases. Among them, the most important one may be disregarding the data cleaning as many authors did (e.g., Cunill, Salvá, Gonzalez, & Mulet-Forteza, 2019; García-Lillo, Claver-Cortés, Úbeda-García, Marco-Lajara, & Zaragoza-Sáez, 2018; Gomezelj, 2016; Lazzeretti, Capone, & Innocenti, 2017; Mody, Hanks, & Cheng, 2021; Shin & Perdue, 2019). For this reason, we propose the approach, named “data curation”, in citation and co-citation analysis and demonstrate its important effect on scientific findings to avoid misleading future research in bibliometric studies. Simply put, we employed sample articles published

in three prestigious journals from service industries to show notable differences between before- and after-data curation approach in the results of citation and co-citation analyses.

Methodology

Study design

For understanding the intellectual structure of a body of knowledge, researchers choose one of two methods. They first depend on a qualitative technique, notably their experience and knowledge, judgments, and perceptions, compromising the reliability and validity of conclusions based on such qualitative methods (Zupic & Čater, 2015). Using quantitative approaches may overcome such limitations, and academics may conduct co-citation analysis to supplement qualitative aspects (Köseoglu, 2020). Co-citation analysis illuminates intellectual processes using statistical and mathematical techniques to analyze study citations or references as data (Arici et. al., 2021).

In the present research, we used two bibliometric methods: citation analysis and co-citation analysis using network analysis. These techniques improve impartiality by evaluating large datasets across time. Citation analysis entails counting references to highlight significant publications on a particular research topic during a specific time frame (Köseoglu et. al., 2021). Citation analysis may give insightful information into a research field's history, today, and future. Researchers can track development by studying the evolution of key references through time and finding the significance of valuable publications (Zupic & Čater, 2015). Co-citation analysis, on the other hand, investigates relationships between prominent references by visualizing intellectual linkages in the particular subject. Using statistics, it creates commonalities between cited article bibliographies, including the number of co-cited authors, articles, or journals (Arici et. al., 2021).

Data collection

The data utilized in the citation and co-citation analysis consist of references cited in service industries studies published by a scientific journal. These are considered to represent a confirmed type of knowledge. The studies were obtained from a top service journal, the Journal of Service Research (JSR). We have adopted the following procedure to select the journal. First, we sought Social Science Citation Index and Scopus, as commonly accepted international databases, and clarified the prominent service-focused journals. Second, we considered the top journal that has a long history and deep background so that we can better reflect the field of service industries. Last, we asked for three academicians' ideas regarding the suitability of the chosen journal. These procedures ended with the inclusion of the top journal in our sample.

Using the Scopus database, we downloaded studies as well as their references published in the service journal from the first issue (1998) to the latest one (December 2020). Later, these documents were transferred to an Excel spreadsheet. We also controlled if these

articles involved references and excluded any study without references. In the end, we deleted six documents, which yielded 5184 studies from JSR.

Before submitting the data for analysis, we followed two different procedures. First, we did not clean the data. As illustrated in Figure 1, we submitted the raw data for citation and co-citation analysis. Moreover, we adopted procedures illustrated in Figure 2 using the data curation approach. In particular, several references were inconsistent since the authors' names, the title of the study, and/or the name of the journal were illustrated differently. By following these procedures, we curated the data to fix the titles of the related journal article references and to enhance correctness. In particular, in the raw data, the types of references included journal articles, books, book chapters, theses, reports, and online addresses. However, drawing on previous bibliometric research (e.g., Köseoglu et al., 2021), in clean data we opted to confine their coverage to journal articles from these references by writing the reference type manually to increase reliability.

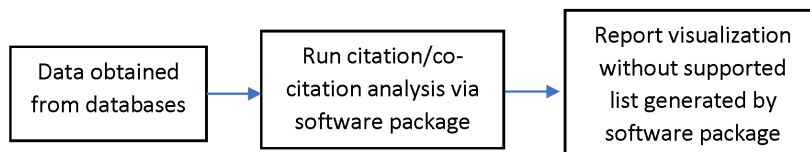


Figure 1

Current application of citation and co-citation without data cleaning/curation

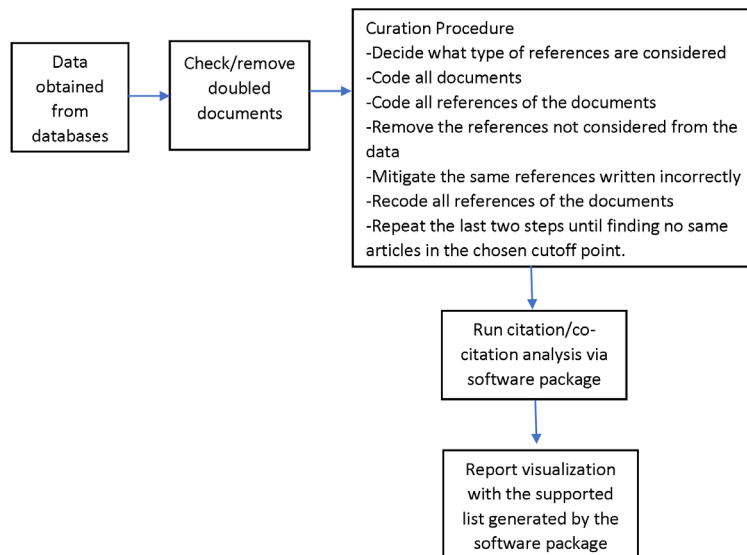


Figure 2

Suggested data cleaning/curation procedure in the citation and co-citation analysis

Data analysis

We performed citation and co-citation network visualization through VOSviewer software to analyse the raw and clean data. VOSviewer clarifies clusters utilizing modularity-based clustering (Köseoglu et. al., 2021). VOSviewer visualizes the most influential documents and co-cited articles in a knowledge field by counting the weight of these publications. This analysis relies on the numbers, i.e., citation and co-citation scores. This also allowed us to clarify the scholarly roots as clusters during the time frame. To normalize their study of the publications, the scholars used the correlation intensity technique (Arici et. al., 2021). Circles stand in for nodes in the network view, while bars depict the connections between nodes. The nodes' locational groupings are identified by their colors. The width of the nodes reveals the frequency of references to a certain publication. The data generated by this software has been provided in the Appendix, including raw and clean datasets. Raw data has been untouched, which means that we directly used it without refining, while the clean data was carefully checked, and we coded each reference separately to eliminate any duplication before submitting the dataset to the analysis.

Results

Citation analysis

We have considered the 200 articles mostly cited to reveal a bigger picture of the most influential articles within the data that can be described by counting citation documents available in the reference lists of the articles published in the journal set. The results have been displayed to explore potential differences among the selected databases. As can be seen in Appendix (online supplement), citation analysis results of the raw data obtained from the *Journal of Service Research* demonstrated that the three most cited ones of the first five articles were methodology-based studies, such as SEM (Anderson & Gerbing, 1988; Bagozzi, 1988; Fornell & Larcker, 1981a). Of the first five articles, other two studies were written about service quality (Parasuraman et. al., 1988) and relationship marketing (Morgan & Hunt, 1994). However, the most cited article (Fornell & Larcker, 1981a) was listed three times like different studies, as the 1st, 75th, and 166th ranks in raw data. Similarly, Vargo and Lusch's (2004) study was listed separately (in the 2nd and 207th ranks). We can illustrate many articles listed as different studies in the raw data (Anderson & Gerbing, 1988; Bagozzi & Yi, 1988; Oliver, 1997; Parasuraman et. al., 1988, 1985; Zeithaml et. al., 1996).

On the other hand, clean data results showed that the most cited article was likewise about SEM (e.g., Fornell & Larcker, 1981), but the second most cited article changed from Parasuraman et. al. (1988) (see Appendix A- online supplement). As we saw in the previous results, the data curation approach has changed the order of the most influential articles. For example, Bitner, Booms, and Tetreault's (1990) study was listed as the 10th most cited article in the raw data, but this study rose to the fourth rank in the clean data. Another surprising result shows that the article of (Zeithaml et. al., 1996) was listed as

the sixth most cited article in the clean data, whereas this article could not be seen within the top ten cited documents in the raw data. The reason behind such a dramatic change was that raw data illustrated the same article as two different documents in the 23rd rank and the 33rd rank, respectively (see Appendix B- online supplement). We want to close this paragraph by providing one more interesting example from both datasets. An article written by (Morgan & Hunt, 1994) was listed differently in the raw data in the rank of 31 (with weights of 12) and 51 (with weights of 10) since the first name of the first author was differently typed (i.e., Morgan, R.M. vs Morgan, Robert, M.). This was a good example for justifying our supposition that such typing or spelling mistakes need to be eliminated by adopting a “data curation approach” to reveal more reliable results.

There is a distinct difference between raw and clean data. Since citation and co-citation analysis influentially directed future research in the knowledge domain, analyzing such raw data including many intentional or unintentionally duplicated entries, would dramatically mislead future researchers. Overall, the results of such a citation analysis relying on the raw data might inaccurately dominate the study field.

Bibliometric scholars have mostly considered the first 20 or 30 most cited articles in citation analysis (e.g., Köseoglu et. al., 2019, 2021). However, utilizing larger data would increase the possibility of finding additional repeated works in the dataset. As seen, our results confirmed the supposition that some studies were illustrated several times in the ranks above the first 100 documents. That is, our study shows that as long as the sample size increases, the duplicating possibility of a document also increases. Thus, we used the 200 most cited documents in our citation analysis so that we could better reflect how many times a document is repeated in the raw data.

Co-citation analysis

Drawing on the co-citation analysis, we identified and visualized main clusters, each portraying a knowledge domain. Appendices featured relevant data concerning the different clusters and their components. As we did in citation analysis, co-citation analysis results of articles obtained from the top journal were illustrated for raw and clean data separately to reveal the main differences.

Co-citation analysis has still suffered from the lack of commonly accepted cut-off points (Köseoglu, 2020). Therefore, scholars have tried various thresholds to achieve the best interpretable dataset (Batistič et. al., 2017). According to Köseoglu (2020), it is reasonable and applicable. Hence, in our study, we used a cut-off level of 10 for revealing mostly cited 100 articles.

We also employed the co-citation analysis for the studies taken from the *Journal of Service Research* to be able to reveal the main differences between raw and clean data in service literature. Co-citation analysis of raw data produced seven different clusters (see Figure 3-A). Interestingly, the cluster map seemed in a heap, which means that six clusters were gathered on the left side of the map, while only one cluster in orange, including two documents, was placed on the right side. More importantly, the clusters involve many duplicate documents. For example, Anderson and Gerbing’s (1988) seminal methodology-

focused work was illustrated within different clusters (see Appendix C- online supplement, clusters 1 and 7). Morgan and Hunt's (1994) study was also listed in clusters 3 (darker blue colored) and 7 (orange colored) with weights of 12 and 10. That is, two studies constituting cluster 7 were illustrated within previous clusters and thus made cluster 7 redundant. As another duplicate study, Zeithaml et. al., (1996) work could be demonstrated (see Appendix

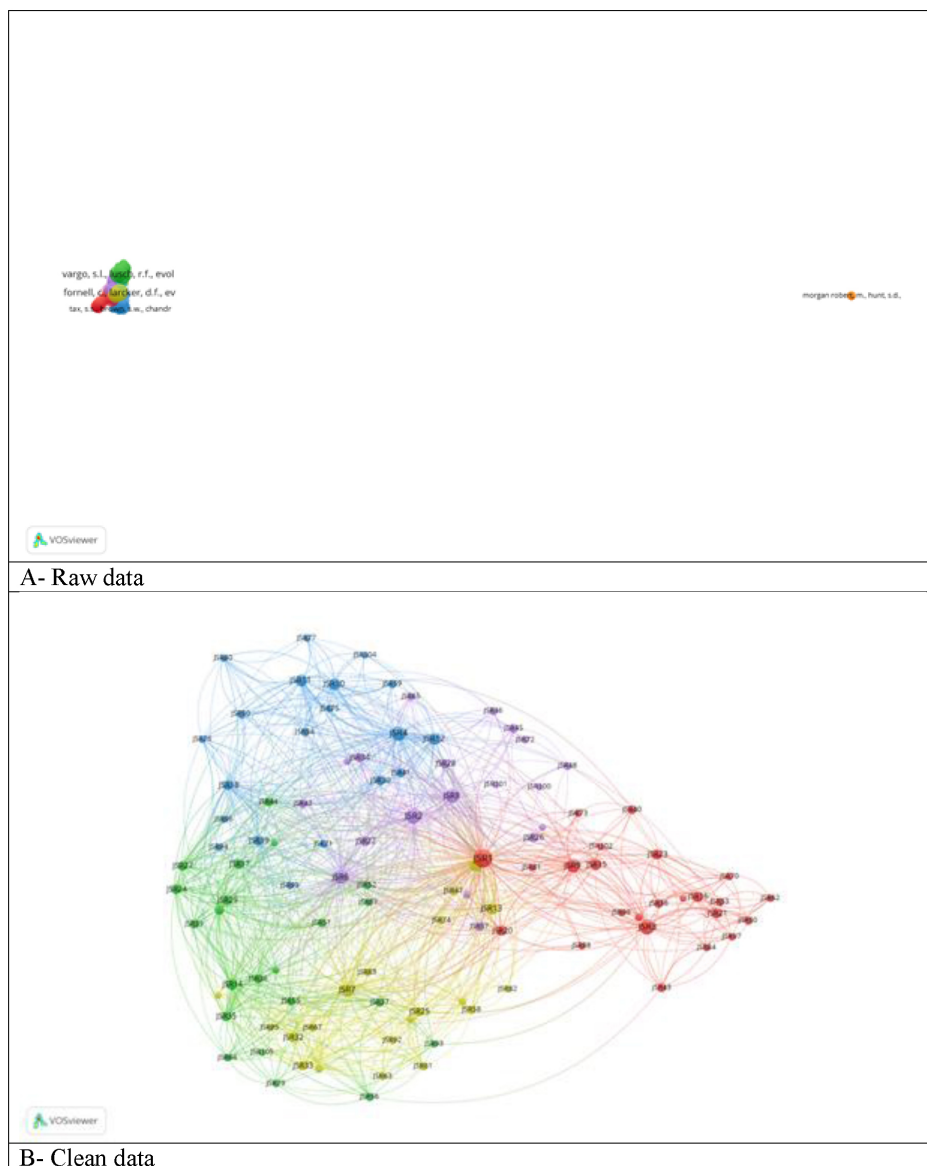


Figure 3
Network analysis of the most cited 100 documents in JSR

C, clusters red colored-1 and light violet colored-5). Nunnally's (1978) pioneering work focusing on psychometric theory was also duplicated across clusters with different weights like distinct studies. Many repeated studies could be found within the cluster map produced by co-citation analysis of the raw data (e.g., Bagozzi & Yi, 1988; Bolton, 1998; Fornell & Larcker, 1981; Parasuraman, Zeithaml, & Berry, 1985).

On the other hand, the clean data results generated five unique clusters in co-citation analysis (see Figure 3-B). Unlike the raw data results, there is no repeated study within and across clusters (see Appendix D- online supplement). The focus of the top ten documents with the highest weights in the first cluster (red colored) is on methodology (Fornell & Larcker, 1981a; Podsakoff et. al., 2003) and service quality (e.g., Bitner, Booms, & Tetreault, 1990; Parasuraman, Zeithaml, & Berry, 1988, 1985; Zeithaml, Berry, & Parasuraman, 1996). The second cluster is green-colored and has mostly concentrated on customer satisfaction as well as its antecedents and consequences (Anderson & Sullivan, 1993; Bolton, 1998; Fornell, 1992; Mittal & Kamakura, 2001; Rust & Zahorik, 1993). In darker blue colored cluster 3, service encounters and customer experiences are dominant topics (e.g., Bitner, 1990; Bitner, Booms, & Tetreault, 1990; Smith, Bolton, & Wagner, 1999; Tax, Brown, & Chandrashekar, 1998). Moreover, customer relation was the main foci of the subsequent cluster colored in yellow (e.g., Berry, 1995; Crosby, Evans, & Cowles, 1990; Dwyer, Schurr, & Oh, 1987; Garbarino & Johnson, 1999; Gremler & Gwinner, 2000; Gwinner, Gremler, & Bitner, 1998; Morgan & Hunt, 1994). Last, documents constituting cluster 5 (light violet colored) more focused on studying service quality as the main research topic (Cronin & Taylor, 1992; Parasuraman et. al., 1988, 1985; Schneider et. al., 1998; Zeithaml, 1988; Zeithaml et. al., 1996).

To check if additional documents potentially affect results in co-citation analysis, drawing on Köseoglu's (2020) suggestion, we expanded our sample by reducing the threshold value to cover more documents. As seen in Appendix (online supplement), this complementary analysis using the threshold value of 6 covered 211 documents out of 34096 and revealed six clusters different from the previous analysis, including 133 documents with a threshold of 7 (Figure 4-A). Another important difference between these analyses was regarding the main themes of clusters. For example, in the former analysis of raw data, three of the five most cited documents of the first cluster were methodology-focused studies (Anderson & Gerbing, 1988; Baron & Kenny, 1986; Bollen, 1989), while the latter's first cluster (red colored) concentrated more on studies examining service context and logic (Ostrom et. al., 2010, 2015; Vargo & Lusch, 2004, 2008).

Furthermore, the complementary analysis also involved many repeated documents as its counterpart. As can be seen in Appendix E (online supplement), when we excluded the duplicate documents, the last cluster (light blue colored cluster 6) included only one study (i.e., Cooil, Keiningham, Aksoy, & Hsu, 2007), because other four studies were already listed in previous clusters (i.e., Anderson & Gerbing, 1988; Bolton, 1998; Morgan & Hunt, 1994; Oliver, 1997). The most cited document by (Vargo & Lusch, 2004) was also demonstrated as a distinct study in cluster 4 (yellow colored), with a weight of 6. In addition, Bitner's (1990, 1992) two studies were repeated in the complementary analysis of the raw data.

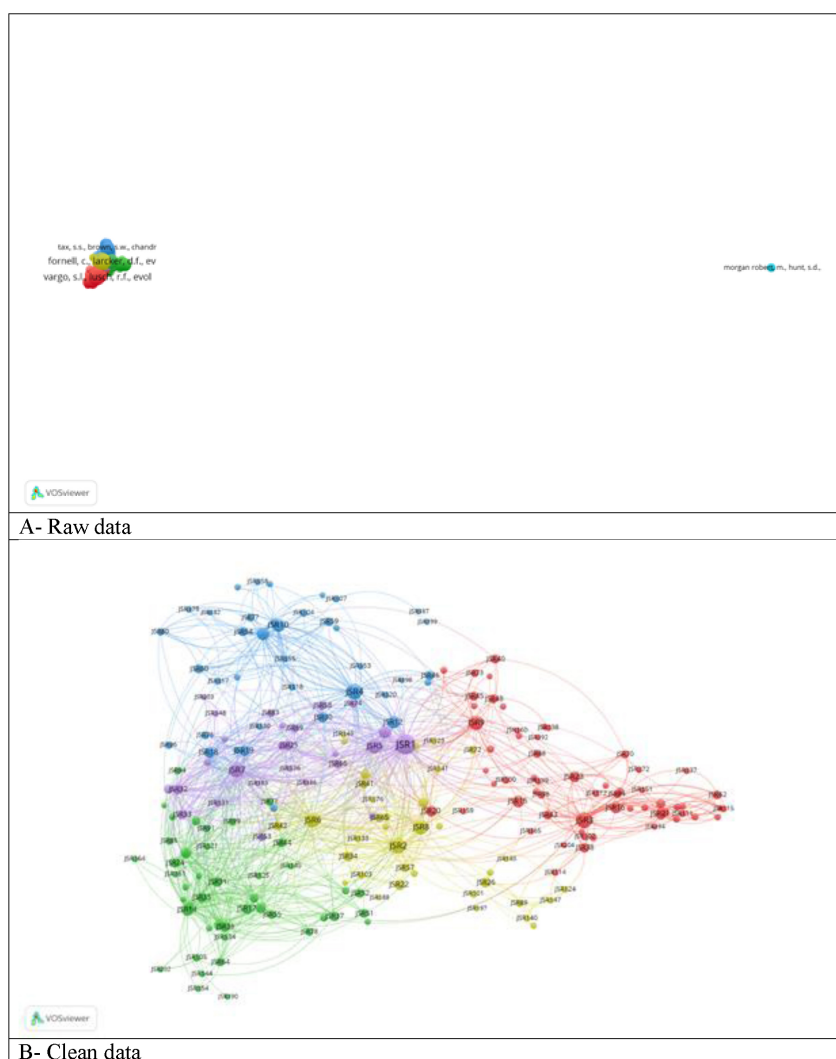


Figure 4

Network analysis of the most cited 200 documents in JSR

We used the same procedure for the clean data. Through expanding the sample, we analyzed 204 documents at least 13 times cited. This complementary analysis provided additional support for our clean data. As shown in Figure 4-B, this complementary analysis revealed five clusters as the former analysis of clean data (with 100 documents). More importantly, there was no significant change between the two analyses of clean data. In particular, within the first 100 documents, the majority of studies illustrated in both analyses were the same, except for Fornell and Larcker's (1981) methodological work. Even though this study had the same weighting (i.e., 132 citations), it was illustrated

in cluster 5 (light violet colored) in the complementary analysis (see Appendix F). The results confirmed that even if the threshold was altered and the number of documents was enhanced, the results showed internal consistency, referring to no significant changes which were detected in the co-citation analysis of both datasets applied data curation. However, in the raw data analysis, the results were significantly changed when the number of documents was increased. Therefore, data curation provides an important advantage to researchers by ensuring the reliability of the findings in co-citation analysis.

As a closing remark, we reveal that data curation plays a significant role in achieving more reliability in citation and co-citation analysis findings. Therefore, future research should not ignore the importance of this approach in bibliometric analyses.

Discussion and Conclusion

“What makes the data curation worthwhile in citation and co-citation analyses?” Even though bibliometric analysis literature has suffered from this ambiguity to our best knowledge no study has attempted to elucidate this knowledge void to date. Therefore, the main focus of our research is to clarify if data cleaning plays a significant role in citation and co-citation analyses. Only even the possibility that a data cleaning procedure possesses to provide more reliable findings deserves an academic effort. In other words, we need to choose our way: *“faster but problematic results from raw data or strong and reliable results from clean data, which one?”* To find the truth, we set out to compare the findings obtained from the samples of raw and clean data. Thus, we considered a highly prestigious journal from the service industries and performed citation and co-citation analyses using both raw and clean data. By doing so, we aimed to reveal the weaknesses of raw data, while demonstrating the strengths of data cleaning. Undoubtedly, our comparative results confirm that data cleaning named the “data curation” approach has a considerable potential to produce more reliable findings in citation and co-citation analysis.

Specifically, the data curation approach eliminated three flaws of raw data analysis in clarifying the intellectual connection of scientific disciplines. As observed in clean data analysis, there was no repeated document within the clusters. That is, each cluster had a unique document. On the contrary, raw data analysis results conducted for the journal presented clusters including several duplicate studies. Second, in the raw data analysis, a cited article presented in a cluster also be also seen in another cluster. However, as for the data curation approach, there is no doubt if a document is illustrated in various clusters in co-citation analysis because any repeated one is eliminated by carefully checking all documents before analysis. Third, visualization of raw data analysis is another problem. As seen, the cluster map generated from the raw data analysis has a skewed distribution of clusters. Most of the clusters were gathered at a point, whereas one or two other clusters were placed in a separate location on the map. In the clean data, all cluster maps provided a good visualization, which means that a reader can easily understand and interpret the distribution of clusters and their contents. In addition, the data curation approach leads to obtaining findings that illustrate internal consistency even when the sample size is

expanded. In the application, we first examined 100 data and then included 200 documents to test whether the change in sample size influences the findings. In clean data, for both samples, the results show consistency which means that there is no significant change in clusters and the documents constituting these clusters. However, raw data analysis could not provide such consistency since the number of clusters and their weightings, and the studies involved in clusters also considerably changed in both datasets. For example, Fornell and Larcker's (1981a) study presented among the most cited articles in the knowledge domain, whereas the same article was repeated many times both within the same cluster and across clusters. Such duplicate studies result in an unexpected change in the number of clusters and their weightings when the sample size is raised. Additionally, the majority of bibliometric studies do not report the most cited document lists as an appendix within the manuscript. If these documents are attached to the studies, we then could better realize the duplicate studies in the lists of most cited studies reported. Hence, we strongly encourage future researchers to report these additional lists (i.e., most cited 30, 100, or 200 studies) to their manuscript to be able to evidence if there is a duplicate document.

We cannot ignore the potential importance of previous bibliometric studies adopting raw data analysis to provide a research agenda in the given field. Still, it seems like trying to find an exit from the dark corridor, since raw data involves many duplicate documents, including the authors' name, the title of papers and journals, and publishers. As illustrated in our findings, analyzing such raw data will probably produce inaccurate cluster contents, consequently, a cluster label. More importantly, a scientific effort, which conducts citation and co-citation analyses relying on raw data has a great potential to manipulate the findings, intentionally and/or unintentionally, which could dramatically mislead future research in the knowledge domain. From this viewpoint, the vitality of the "data curation" approach needs scientometrics/bibliometrics researchers. In short, as a response to the above-mentioned question, our study reveals that data curation is worthwhile in citation and co-citation analyses for gaining reliable findings and providing a solid agenda for future research in the scientific field.

Despite its potential to unlock Pandora's box, this study has inherently suffered from some limitations. First, even if we tried to cover the main global database, considering only a top journal from the service industries could jeopardize the generalizability of our findings. Future research, thus, could perform additional efforts to scrutinize the usage of the data curation approach against raw data analysis so that we could better understand the generalizability of our proposed approach. Second, our approach, proposed for cleaning the data used to analyze the intellectual structure of the scientific field, requires several rigorous stages, including checking/removing doubled documents, coding all documents and their references separately, mitigating the same references written incorrectly, and re-coding all references of documents. Unfortunately, to our best knowledge, to clean the data for citation and co-citation analyses, the current software programs cannot check, code, remove (double documents), and re-code the remaining references. Thus, in some circumstances, particularly for analyzing big data, it can be a time-consuming process. Moreover, we used VOSviewer software to analyze both the raw and clean data, and thus,

adopting other software programs in the future study would pay dividends. Lastly, like its counterparts, our research has also relied on interpreting findings that can consist of biases.

References

- [1] Anderson, E. W., & Sullivan, M. W. The antecedents and consequences of customer satisfaction for firms. *Marketing Science*, 12(2), 125–143 (1993).
- [2] Anderson, J. C., & Gerbing, D. W. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3), 411 (1988).
- [3] Arici, H. E., Arici, N. C., Köseoglu, M. A., & King, B. E. M. Leadership research in the root of hospitality scholarship: 1960–2020. *International Journal of Hospitality Management*, 99, 103063 (2021).
- [4] Bagozzi, R. P., & Yi, Y. On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 16(1), 74–94 (1988).
- [5] Baron, R. M., & Kenny, D. A. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173 (1986).
- [6] Batistič, S., Černe, M., & Vogel, B. Just how multi-level is leadership research? A document co-citation analysis 1980–2013 on leadership constructs and outcomes. *The Leadership Quarterly*, 28(1), 86–103 (2017). <https://doi.org/10.1016/j.leaqua.2016.10.007>
- [7] Bellardo, T. The use of co-citations to study science. *Library Research*, 2(3), 231–237 (1980).
- [8] Benckendorff, P., & Zehrer, A. A network analysis of tourism research. *Annals of Tourism Research*, 43, 121–149 (2013).
- [9] Berry, L. L. Relationship marketing of services—Growing interest, emerging perspectives. *Journal of the Academy of Marketing Science*, 23(4), 236–245 (1995).
- [10] Bitner, M. J. Evaluating service encounters: The effects of physical surroundings and employee responses. *Journal of Marketing*, 54(2), 69–82 (1990).
- [11] Bitner, M. J. Servicescapes: The impact of physical surroundings on customers and employees. *Journal of Marketing*, 56(2), 57–71 (1992).
- [12] Bitner, M. J., Booms, B. H., & Tetreault, M. S. The service encounter: Diagnosing favorable and unfavorable incidents. *Journal of Marketing*, 54(1), 71–84 (1990).
- [13] Bollen, K. A. *Structural equations with latent variables*. New York: John Wiley (1989).
- [14] Bolton, R. N. A dynamic model of the duration of the customer’s relationship with a continuous service provider: The role of satisfaction. *Marketing Science*, 17(1), 45–65 (1998).
- [15] Boyack, K. W., & Klavans, R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389–2404 (2010).

- [16] Cawkell, A. E. Citations, obsolescence, enduring articles, and multiple authorships. *Journal of Documentation*, 32(1), 53–58 (1976). <https://doi.org/10.1108/eb026615>.
- [17] Chen, C., Ibekwe-SanJuan, F., & Hou, J. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, 61(7), 1386–1409 (2010).
- [18] Cooil, B., Keiningham, T. L., Aksoy, L., & Hsu, M. A longitudinal analysis of customer satisfaction and share of wallet: Investigating the moderating effect of customer characteristics. *Journal of Marketing*, 71(1), 67–83 (2007).
- [19] Cronin Jr, J. J., & Taylor, S. A. Measuring service quality: A reexamination and extension. *Journal of Marketing*, 56(3), 55–68 (1992).
- [20] Crosby, L. A., Evans, K. R., & Cowles, D. Relationship quality in services selling: An interpersonal influence perspective. *Journal of Marketing*, 54(3), 68–81 (1990).
- [21] Cunill, O. M., Salvá, A. S., Gonzalez, L. O., & Mulet-Forteza, C. Thirty-fifth anniversary of the International Journal of Hospitality Management: A bibliometric overview. *International Journal of Hospitality Management*, 78, 89–101 (2019).
- [22] Diodato, V. P., & Gellatly, P. *Dictionary of Bibliometrics*. New York: Routledge (2013).
- [23] Daud, A., Hayat, M. K., Alshdadi, A. A., Banjar, A., & Alharbi, W. M. Measuring the impact of co-author count on citation count of research publications. *COLLNET Journal of Scientometrics and Information Management*, 16(1), 35–48 (2022).
- [24] Dwyer, F. R., Schurr, P. H., & Oh, S. Developing buyer-seller relationships. *Journal of Marketing*, 51(2), 11–27 (1987).
- [25] Ellegaard, O., & Wallin, J. A. The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, 105(3), 1809–1831 (2015).
- [26] Fornell, C. A national customer satisfaction barometer: The Swedish experience. *Journal of Marketing*, 56(1), 6–21 (1992).
- [27] Fornell, C., & Larcker, D. F. Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50 (1981a).
- [28] Fornell, C., & Larcker, D. F. Structural Equation Models with Unobservable Variables and Measurement Error: Algebra and Statistics. *Journal of Marketing Research*, 18(3), 382–388 (1981b). <https://doi.org/10.1177/002224378101800313>
- [29] Garbarino, E., & Johnson, M. S. The different roles of satisfaction, trust, and commitment in customer relationships. *Journal of Marketing*, 63(2), 70–87 (1999).
- [30] García-Lillo, F., Claver-Cortés, E., Úbeda-García, M., Marco-Lajara, B., & Zaragoza-Sáez, P. C. Mapping the “intellectual structure” of research on human resources in the “tourism and hospitality management scientific domain”: Reviewing the field and shedding light on future directions. *International Journal of Contemporary Hospitality Management*, 30(3), 1741–1768 (2018). <https://doi.org/10.1108/IJCHM-04-2017-0187>
- [31] Garfield, E. Citation indexes for science. *Science*, 122(3159), 108–111 (1955).

- [32] Garfield, E. Citation indexes in sociological and historical research. *American Documentation*, 14(4), 289–291 (1963).
- [33] Garfield, E. Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4), 359–375 (1979).
- [34] Glänzel, W., Schubert, A., & Czerwon, H.-J. A bibliometric analysis of international scientific cooperation of the European Union (1985–1995). *Scientometrics*, 45(2), 185–202 (1999).
- [35] Gomezelj, D. O. A systematic review of research on innovation in hospitality and tourism. *International Journal of Contemporary Hospitality Management*, 28(3), 516–558 (2016). <https://doi.org/10.1108/IJCHM-10-2014-0510>
- [36] Gremler, D. D., & Gwinner, K. P. Customer-employee rapport in service relationships. *Journal of Service Research*, 3(1), 82–104 (2000).
- [37] Gwinner, K. P., Gremler, D. D., & Bitner, M. J. Relational benefits in services industries: The customer's perspective. *Journal of the Academy of Marketing Science*, 26(2), 101–114 (1998).
- [38] Hota, P. K., Subramanian, B., & Narayanamurthy, G. Mapping the intellectual structure of social entrepreneurship research: A citation/co-citation analysis. *Journal of Business Ethics*, 1–26 (2019).
- [39] Hou, J. Exploration into the evolution and historical roots of citation analysis by referenced publication year spectroscopy. *Scientometrics*, 110(3), 1437–1452 (2017).
- [40] Hou, J., Yang, X., & Chen, C. Emerging trends and new developments in information science: A document co-citation analysis (2009–2016). *Scientometrics*, 115(2), 869–892 (2018).
- [41] Köseoglu, M. A. Identifying the intellectual structure of fields: Introduction of the MAK approach. *Scientometrics*, 125(3), 2169–2197 (2020). <https://doi.org/10.1007/s11192-020-03719-8>
- [42] Köseoglu, M. A., Mehraliyev, F., Aladag, O. F., & King, B. Origins, evolution and themes of scholarly hospitality sources: 1960–2019. *International Journal of Hospitality Management*, 94, 102817 (2021). <https://doi.org/10.1016/j.ijhm.2020.102817>
- [43] Köseoglu, M. A., Okumus, F., Dogan, I. C., & Law, R. Intellectual structure of strategic management research in the hospitality management field: A co-citation analysis. *International Journal of Hospitality Management*, 78, 234–250 (2019).
- [44] Lancaster, F. W. *Indexing and Abstracting in Theory and Practice*. London: Library Association (1991).
- [45] Lazzeretti, L., Capone, F., & Innocenti, N. Exploring the intellectual structure of creative economy research and local economic development: A co-citation analysis. *European Planning Studies*, 25(10), 1693–1713 (2017).
- [46] Le Pair, C. The citation gap of applicable science. In *Handbook of quantitative studies of science and technology* (pp. 537–553) (1988). Elsevier.

- [47] Liu, Z., Yin, Y., Liu, W., & Dunford, M. Visualizing the intellectual structure and evolution of innovation systems research: A bibliometric analysis. *Scientometrics*, 103(1), 135–158 (2015).
- [48] MacRoberts, M. H., & MacRoberts, B. R. The mismeasure of science: Citation analysis. *Journal of the Association for Information Science and Technology*, 69(3), 474–482 (2018).
- [49] McCain, K. W. Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433–443 (1990).
- [50] Mittal, V., & Kamakura, W. A. Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics. *Journal of Marketing Research*, 38(1), 131–142 (2001).
- [51] Mody, M. A., Hanks, L., & Cheng, M. Sharing economy research in hospitality and tourism: A critical review using bibliometric analysis, content analysis and a quantitative systematic literature review. *International Journal of Contemporary Hospitality Management*, ahead-of-print(ahead-of-print). (2021). <https://doi.org/10.1108/IJCHM-12-2020-1457>
- [52] Morgan, R. M., & Hunt, S. D. The commitment-trust theory of relationship marketing. *Journal of Marketing*, 58(3), 20–38 (1994).
- [53] Nicolaisen, J. Citation analysis. *Annual Review of Information Science and Technology*, 41(1), 609–641 (2007). <https://doi.org/10.1002/aris.2007.1440410120>
- [54] Nunnally, J. C. *Psychometric Theory*. New York: McGraw-Hill (1978).
- [55] Oliver, R. L. *Satisfaction: A behavioral perspective on the consumer: A behavioral perspective on the consumer*. New York: McGraw-Hill (1997).
- [56] Ostrom, A. L., Bitner, M. J., Brown, S. W., Burkhard, K. A., Goul, M., Smith-Daniels, V., ... Rabinovich, E. Moving forward and making a difference: Research priorities for the science of service. *Journal of Service Research*, 13(1), 4–36 (2010).
- [57] Parasuraman, A., Zeithaml, V. A., & Berry, L. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. 1988, 64(1), 12–40 (1988).
- [58] Parasuraman, A., Zeithaml, V. A., & Berry, L. L. A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, 49(4), 41–50 (1985). <https://doi.org/10.1177/002224298504900403>.
- [59] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879 (2003).
- [60] Reale, D., Khelifaoui, M., Montiglio, P.-O., & Gingras, Y. Mapping the dynamics of research networks in ecology and evolution using co-citation analysis (1975–2014). *Scientometrics*, 122(3), 1361–1385 (2020).
- [61] Rust, R. T., & Zahorik, A. J. Customer satisfaction, customer retention, and market share. *Journal of Retailing*, 69(2), 193–215 (1993).

- [62] Scheffler, M., & Brunzel, J. Destructive leadership in organizational research: A bibliometric approach. *Scientometrics*, 125(1), 755–775 (2020). <https://doi.org/10.1007/s11192-020-03621-3>
- [63] Schneider, B., White, S. S., & Paul, M. C. Linking service climate and customer perceptions of service quality: Tests of a causal model. *Journal of Applied Psychology*, 83(2), 150 (1998).
- [64] Shin, H., & Perdue, R. R. Self-Service Technology Research: A bibliometric co-citation visualization analysis. *International Journal of Hospitality Management*, 80, 101–112 (2019). <https://doi.org/10.1016/j.ijhm.2019.01.012>
- [65] Small, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269 (1973).
- [66] Smith, A. K., Bolton, R. N., & Wagner, J. A model of customer satisfaction with service encounters involving failure and recovery. *Journal of Marketing Research*, 36(3), 356–372 (1999).
- [67] Tax, S. S., Brown, S. W., & Chandrashekar, M. Customer evaluations of service complaint experiences: Implications for relationship marketing. *Journal of Marketing*, 62(2), 60–76 (1998).
- [68] Vargo, S. L., & Lusch, R. F. Evolving to a new dominant logic for marketing. *Journal of Marketing*, 68(1), 1–17 (2004).
- [69] Vargo, S. L., & Lusch, R. F. Service-dominant logic: Continuing the evolution. *Journal of the Academy of Marketing Science*, 36(1), 1–10 (2008).
- [70] White, H. D., & Griffith, B. C. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163–171 (1981).
- [71] White, H. D., & McCain, K. W. Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, 49(4), 327–355 (1998).
- [72] Zeithaml, V. A. Consumer Perceptions of Price, Quality, and Value: A Means-End Model and Synthesis of Evidence. *Journal of Marketing*, 52(3), 2–22 (1988). <https://doi.org/10.1177/002224298805200302>.
- [73] Zeithaml, V. A., Berry, L. L., & Parasuraman, A. The behavioral consequences of service quality. *Journal of Marketing*, 60(2), 31–46 (1996).
- [74] Zupic, I., & Čater, T. Bibliometric methods in management and organization. *Organizational Research Methods*, 18(3), 429–472 (2015).
- [75] Deepjyoti Kalita. Citation Analysis of Science, *COLLNET Journal of Scientometrics and Information Management*, 10:2, 237–254 (2016), DOI: 10.1080/09737766.2016.1213967.
- [76] Ali Daud, Malik Khizar Hayat, Abdulrahman A. Alshdadi, Ameen Banjar & Wael Mansour Alharbi. Measuring the impact of co-author count on citation count of research publications, *COLLNET Journal of Scientometrics and Information Management*, 16:1, 35–48 (2022), DOI: 10.1080/09737766.2021.2016356.
- [77] <https://github.com/bibliometricresearcher/datacuration.git>.