A REFLECTION OF SOUTHERN FORESTS: A JOURNAL OF FOREST SCIENCE USING BIBLIOMETRICS

S Grobbelaar[1], R. Oosthuizen[1]

[1]Department of Engineering and Technology Management, Faculty of Engineering, Built Environment & IT, University of Pretoria, Private bag X20, Hatfield, 0028, South Africa.

Corresponding author: schalk.grobbelaar@up.ac.za

Abstract

Bibliometrics is used to determine patterns in published research. The aim of this paper is to illustrate the observable bibliometric patterns in the journal Southern Forests: A Journal of Forest Science. Frequency analysis and co-occurrence network analysis were performed to identify patterns. Natural Language Processing and Supervised Machine Learning were used to perform text classification. The objective of the text classification was to classify articles into fifteen themes. Each article was categorised in terms of the two main themes associated with the article. The analysis included 1 574 publications from 1941 to 2020 and confirmed that the journal's change in name and aims were successful in increasing the number of international researchers publishing in the journal. The research institute co-occurrence network diagram illustrates that there are two main research collaboration clusters. The one surrounds Stellenbosch University, and the other encompasses several South African universities and research institutes. Mondi and Sappi were found to be the companies that collaborated the most with independent research institutes. The keywords and theme analysis confirmed that the journal's aim and scope were supported in the publications. The theme analysis also identified areas with very few publications. The methods illustrated in this paper can be used to identify research strengths and weaknesses and may assist in strategic planning for future research prioritisation.

KEYWORDS: Bibliometrics, Supervised Machine Learning, Sustainability, Network Analysis

1. INTRODUCTION

Southern Forests: A Journal of Forest Science is published in association with the Southern African Institute of Forestry and is a leading forestry journal in the Southern Hemisphere (Informa, 2021). It is registered in the United Kingdom, and research from South America, Africa, Australasia, and Asia in forest science and the management of fast-growing, planted, or natural forests are especially encouraged. The journal started publishing in 1938 as the Journal of the South African Forestry Association. In 1962 the journal's name changed to South African Forestry Journal, in 1997 to Southern African Forestry Journal, in 2007 to Southern Hemisphere Forestry Journal, and from 2008 it was known as Southern Forests: a Journal of Forest Science.

The initial objectives of the journal were to publish findings related to tree planting, promote forest conservation, establish recreational use of forests, improve the application of scientific forestry, and stimulate open discussion related to forestry (Owen, 2002). The journal became available on the internet in 2003 (Owen, 2003). The journal witnessed substantial forestry productivity increases in Southern Africa, South America, and Australasia, even though the competition from other land uses and legislation led to reduced afforested land (Owen (2002); Owen (2003); Owen (2006); Owen (2005a); Owen (2005b); Owen (2006); Marais (2007); Owen (2008) and Du Toit (2011)).

Scopus has indexed 2 424 articles published by the journal (Elsevier, 2021).

This paper aims to illustrate a bibliometric analysis of this journal. Bibliometrics includes any quantitative analysis of research publications (Broadus, 1987) and entails a systematic study of a scientific field to identify research patterns (Oosthuizen and Pretorius, 2020). It is used to illustrate the development of a discipline by analysing

historical movements (Pritchard, 1969) and in this case it was used to evaluate the primary genus, keyword frequency, author affiliation frequency, author affiliation network, author country, author frequency, co-author network, theme frequency and theme network.

This paper attempted to answer six research questions. The first objective was to determine which genera were most frequently researched. This would confirm whether most of the publications were related to fast-growing species. In addition, the frequency of keywords could indicate the popularity of specific research themes, indicating whether the journal achieved its aims and scope. Reviewing the research institutions' collaboration and diversity would indicate whether the journal provides diverse findings while stimulating open discussion. The journal initially focused on South African research but changed its scope to include research relevant to the Southern Hemisphere. By evaluating the frequency of publications by countries outside of South Africa, it could be determined whether the journal was successful in achieving this objective. Lastly, an analysis of the frequency of themes and their relatedness could indicate whether the journal includes substantive scientific research on topics relevant to the aims and scope of the journal.

However, the main objective is to indicate how Southern Forests: A Journal of Forest Science has contributed to the development of forest research and to identify research gaps. Themes that have low frequency but are very relevant to the opportunities or risks of the industry could be considered as research gaps.

2. MATERIALS AND METHODS

Bibliometric data for 2 424 articles were downloaded on 17 September 2020 using Scopus. This included the article title, keywords, abstract, authors, author affiliations,

country of research organisation, and publication date. Duplicate articles, obituaries, prefaces, comments to the editor, notes or comments from the editors, editorials, announcements, advertisements, corrections, letters, correspondence, reviews, speeches, reports and articles without an abstract were removed from the data source. This left 1 574 articles to be analysed. Frequency tables and graphs were developed for the genus and author country using Microsoft Excel. Frequency and co-occurrence network diagrams were developed for the keywords, author, and author affiliation, using Gephi (Bastian et al., 2009).

One of the main aspects of this paper entailed the categorisation of papers into different themes through text classification (Kowsari et al., 2019) using Natural Language Processing (NLP) (Nadkarni et al., 2011) and Supervised Machine Learning (SML) (Gianey and Choudhary, 2018), a subset of Artificial Intelligence (AI) (Zhang, 2020). The benefit of using this methodology is that it can automatically categorise a large number of papers. A combination of NLP and SML (Fisher et al., 2016) was used to classify (Singh and Hasan 2015) 1 574 articles using a training dataset of 543 articles.

Each article in the training dataset for the SML model was classified in terms of the two primary themes of the article. The themes included were: breeding, diseases, fire, forecasting, forest economics, harvesting, inventory management, non-industrial forests and natural areas, pests, processing, silviculture of plantations, site classification, sustainability, transport, and wood properties. The themes were developed using the main sections in the South African Forestry Handbook (Bredenkamp and Upfold, 2012). Regarding sustainability, the broad sense of sustainability that includes social, economic, and environmental aspects were considered (Smith, 2012).

Microsoft Azure Machine Learning Studio (classic), a public cloud-based application that supports a wide range of statistical machine learning algorithms (Notz et al., 2019), was utilised to perform the classification. It uses standard blocks, but customised modules can also be created using Python or R (Pliuhin et al., 2019). A process similar to the Microsoft Azure Machine Learning Studio (classic) Text Classification Template (Microsoft, 2015) was used.

The process started with data preprocessing. This entailed removing irrelevant data, removing duplicates, removing empty entries, structuring the columns and finalising the Comma-Separated Values (CSV) file. Once the file was ready, it was uploaded to Microsoft Azure Machine Learning Studio (classic) to start with text preprocessing using R (R Core Team, 2014). This entailed removing special characters, removal of duplicate characters, replacement of numbers, conversion to lower case, removing stop words, and stemming of the words. Stemming was used to reduce words to their root form to improve the accuracy and speed of the model (Rajput and Khare, 2015). Thereafter, the text was divided into N-grams.

In some cases, word order is imperative to understand the context of words. For this reason, N-grams (a contiguous sequence of terms) were included in the Bag-of-Words (Wang et al., 2007). For this paper, tri-grams were used. A frequency of occurrence extraction was performed to calculate the number of times a tri-gram occurred in the corpus. This was followed by feature selection to reduce dimensionality (a hashing bitsize of 15 was used), improve the classifier's performance, reduce over-filtering, and improve the model's interpretability (Sarkar and Goswami, 2013). The Chi-squared method was used with 5 000 features.

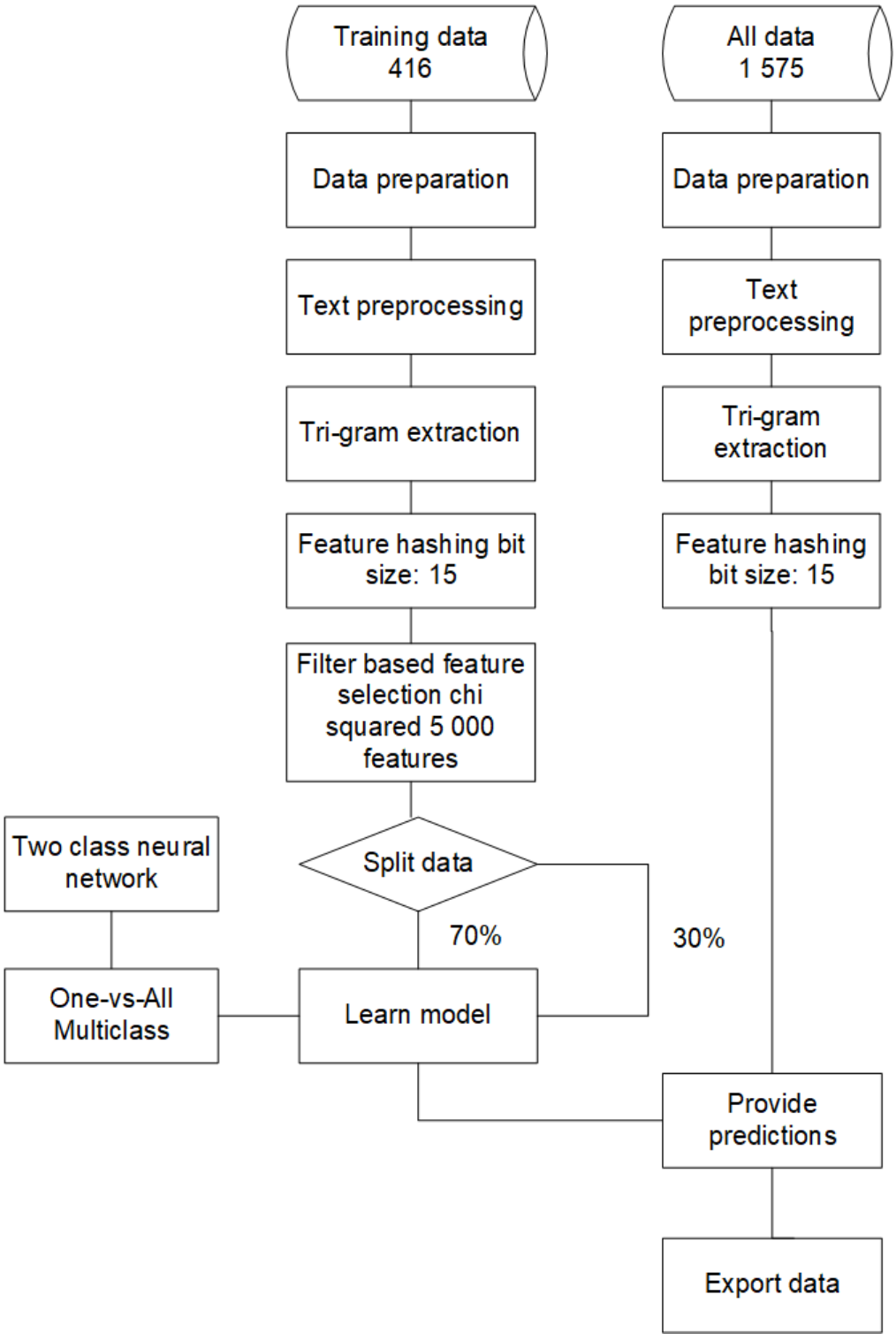

Figure 1: SML data pipeline for text classification

The data was split into training and evaluation sets. 70% of the training dataset was used to train the model, and 30% was used to evaluate the model. Cross-validation was not used for this study to limit the computational capacity required to run the model. However, future research may include this to optimise the model. A two-class neural network model was used to train the model with the training dataset. The parameters used were a single parameter trainer mode, a fully-connected case, 100 hidden nodes, a learning rate of 0,1, 100 learning iterations, an initial learning weight parameter of 0,1, no momentum term and a Min-Max normalizer was used. The two-class neural network model's performance (90% average accuracy) was compared with that of a support vector machine (SVM) model (89% average accuracy) and a multi-class neural network model (88% average accuracy). Since the two-class neural network model performed better than the other two, it was used to score the complete dataset. Once the model was trained, it was used to predict primary and secondary themes for the entire dataset. Finally, the results were exported to a CSV to perform further analysis using Excel and Gephi (Bastian et al., 2009). Network diagrams were compiled using the OpenOrd toolbox (Martin et al., 2011).

3. RESULTS

For the articles that indicated which genus was studied, the majority focused on hardwoods (*Eucalyptus* and *Acacia*). However, most papers investigated the genus *Pinus*, and the second most popular was *Eucalyptus* (Table 1).

Table 1: Genus prevalence

Hardwood/Softwood	Genus	Proportion	Total
Hardwood	*Eucalyptus*	33%	60%
Hardwood	*Acacia*	11%	
Hardwood	Other	16%	
Softwood	*Pinus*	37%	40%
Softwood	Other	3%	

Most of the authors were affiliated with South African research institutes. The second most came from the United States of America, and this was followed by Australia and Brazil (Figure 2).

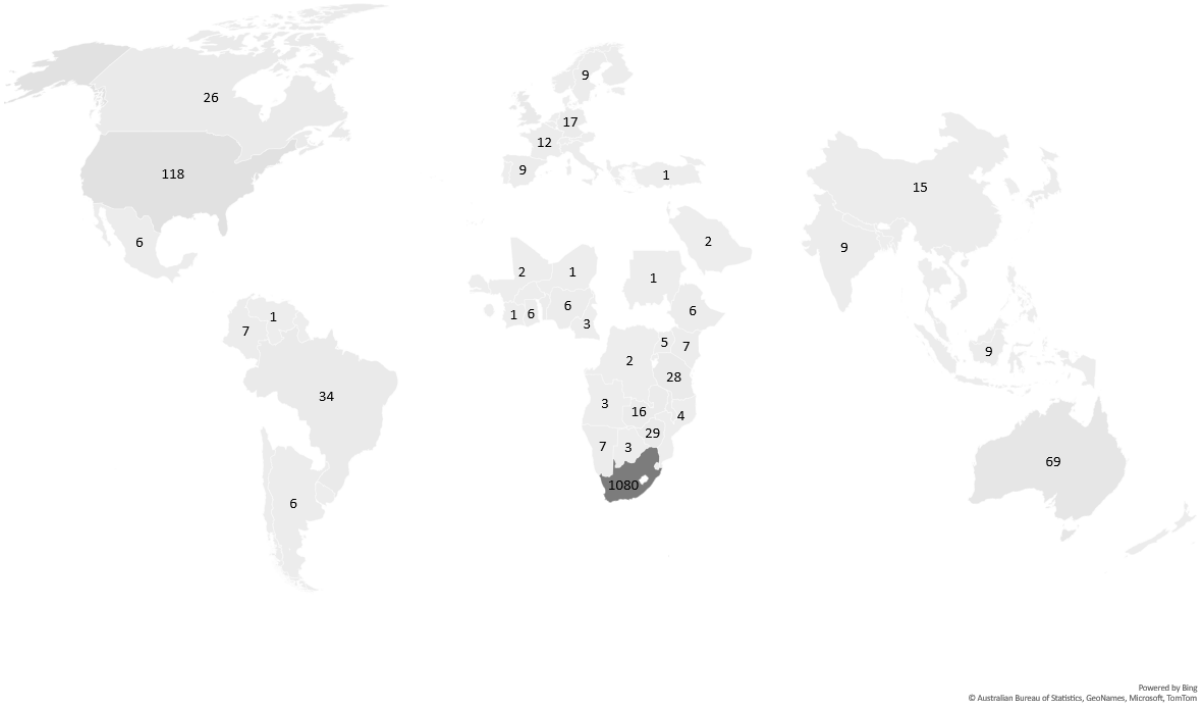

Figure 2: Research institute country

However, in recent decades the portion of authors outside Sub-Saharan Africa has increased, and from 2010 to 2020, only 51% of the authors were affiliated with Sub-Saharan African research institutes (Figure 3).

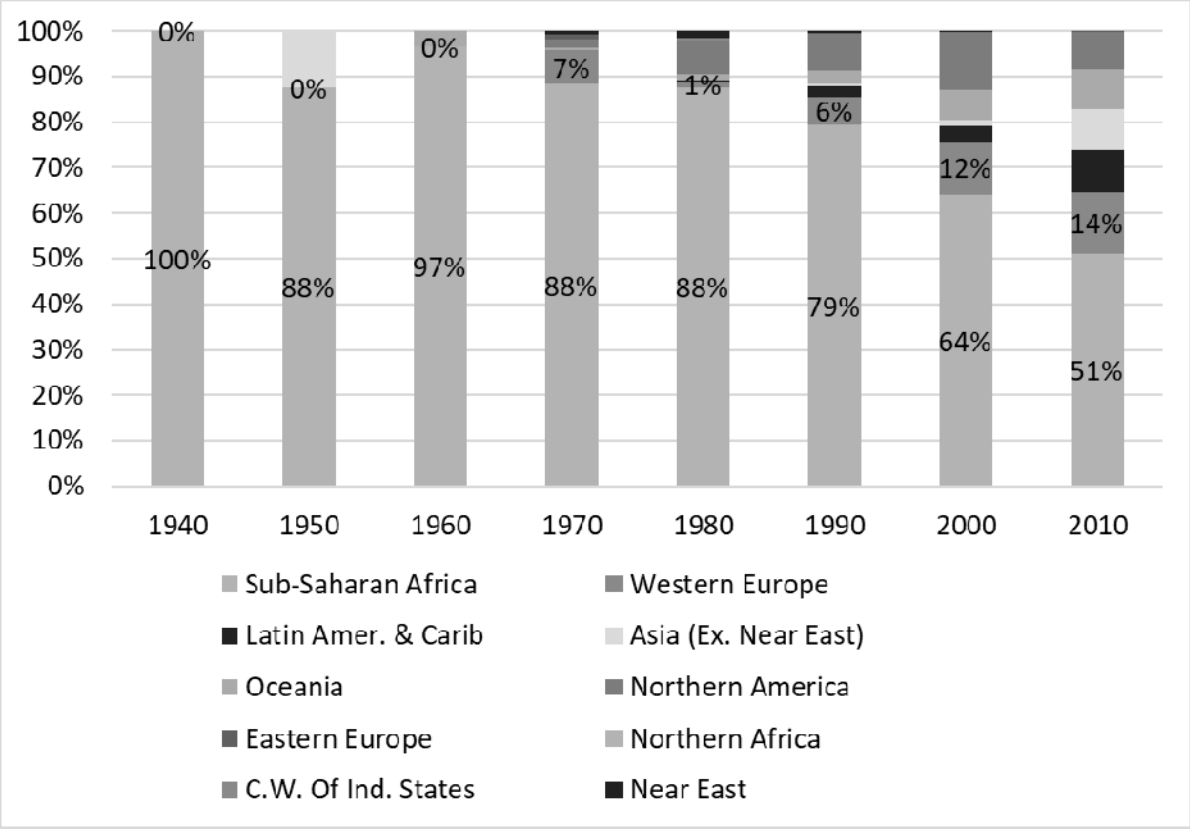

Figure 3: Research institute region

A keyword frequency and co-occurrence analysis indicated that the most popular keywords include *P. patula*, *E. grandis*, *Eucalyptus*, Growth, Survival, South Africa, Productivity, Biomass, Heritability, Wood Density, *P. elliottii*, *Fusarium circinatum*, Harvesting, Provenances, Height, *Acacia mearnsii*, *P. radiata*, Site Preparation, Remote Sensing, Forestry, Plantations, Sustainability, Genetic Parameters, Modulus of Elasticity, Climate Change, Cuttings, and Non-Timber Forestry Products. The proximity of a word to another word indicates the co-occurrence strength (Figure 4).

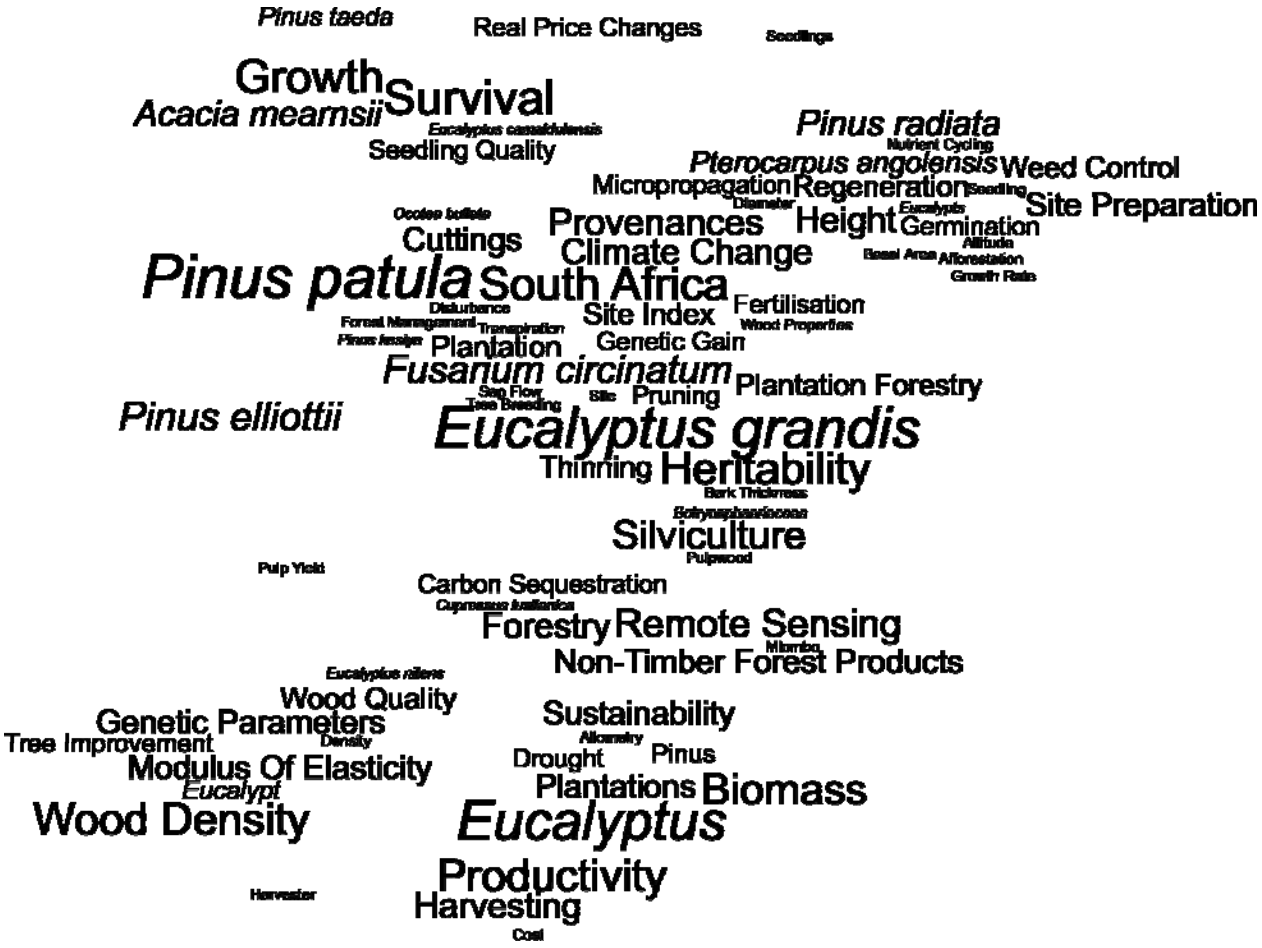

Figure 4: Keyword frequency and co-occurrence

The frequency and co-occurrence network diagram for author affiliations is illustrated in Figure 5. The node's size represents the frequency that a research institute occurred. The thickness of the edges and the proximity of two nodes to each other illustrate how frequently two research institutes collaborated on a publication. The research institutes with the most publications include Stellenbosch University, Nelson Mandela Metropolitan University, The Council for Scientific and Industrial Research, University of KwaZulu-Natal, University of Pretoria, and the Institute for Commercial Forestry Research. The DR De Wet Forestry Research Station, Mondi, Sappi, South African Forestry Research Institute, World Agroforestry, and the Plant Protection Research Institute, Pretoria were among the institutes that frequently published in the journal. The network diagram also

illustrates that Stellenbosch University has a high number of collaborations with numerous research institutes and subsequently forms a collaboration cluster separated from the other research institutes that are highly represented.

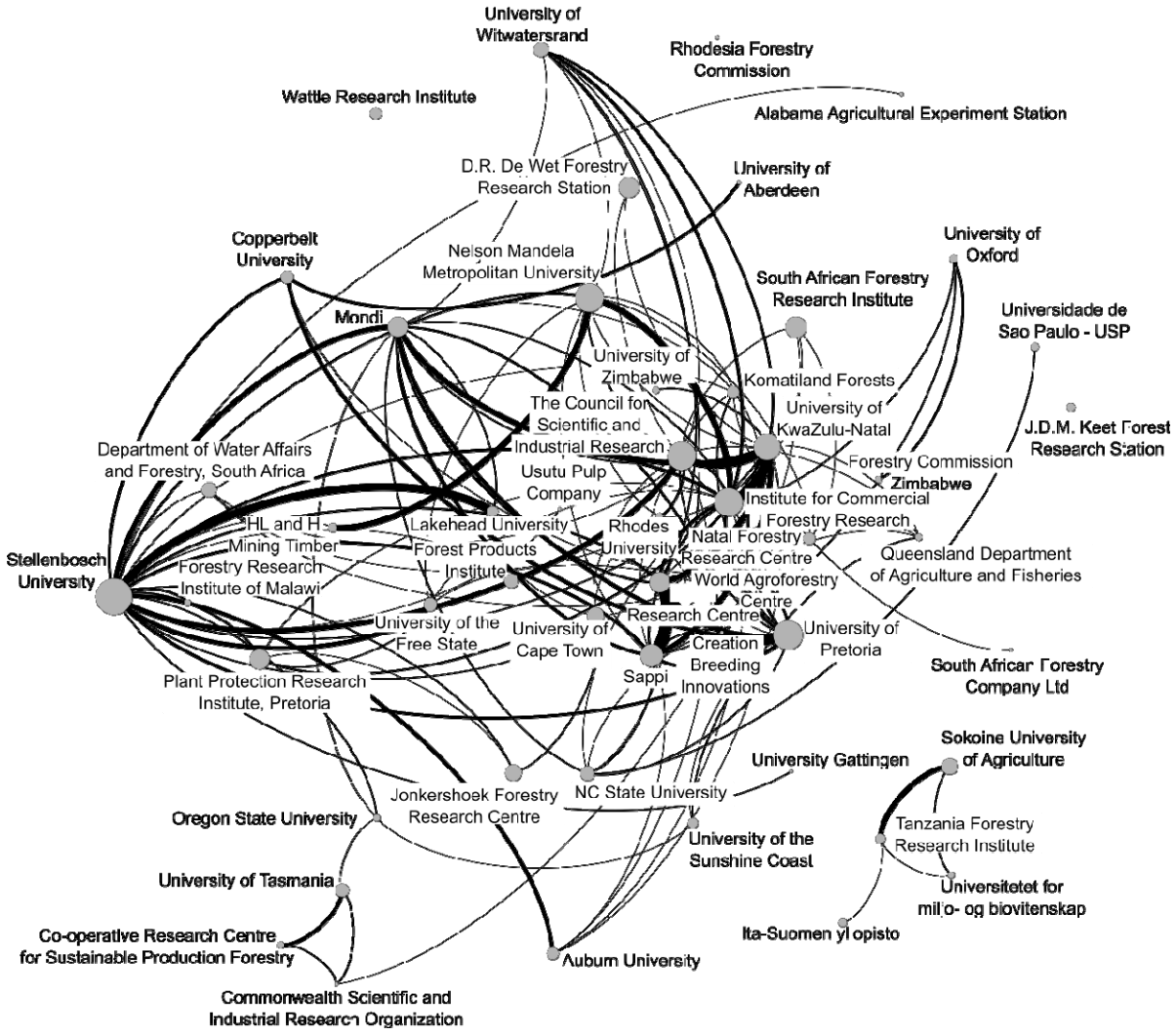

Figure 5: Researcher affiliation frequency and co-occurrence network diagram

The data pipeline's primary theme prediction accuracy for the training data was 90%, and for the secondary theme, it was 82%. The prediction accuracy is close to the expected accuracy of 85% (Scharkow, 2013). The prediction accuracy of a two-class support vector machine model and a multi-class neural network model was slightly weaker than the two-

class neural network model's accuracy. Figure 6 and Table 2 illustrate the frequency and co-occurrence network diagram for the research themes. The node's size demonstrates the frequency of a specific theme, and the thickness of the edges and the proximity of themes to each other illustrate the frequency of co-occurrence. The most popular themes include sustainability, silviculture of plantations, and inventory management. This was followed by breeding, site classification, forecasting, wood properties, harvesting, and forest economics. It is also noteworthy that there were very few publications concerning transport, processing, pests, diseases and fire. From 2010 to 2020, there was only one publication concerning fire management (Little and Nadel, 2014).

Table 2: Frequency per theme

Theme	Frequency	Proportion
Sustainability	1 171	37,2%
Silviculture of plantations	474	15,1%
Inventory management	318	10,1%
Site classification	230	7,3%
Breeding	224	7,1%
Wood properties	146	4,6%
Forecasting	145	4,6%
Harvesting	104	3,3%
Forest economics	99	3,1%
Non-Industrial Forests and Natural Areas	65	2,1%
Diseases	63	2,0%
Fire	59	1,9%
Pests	30	1,0%
Processing	15	0,5%
Transport	5	0,2%
Total	3 148	

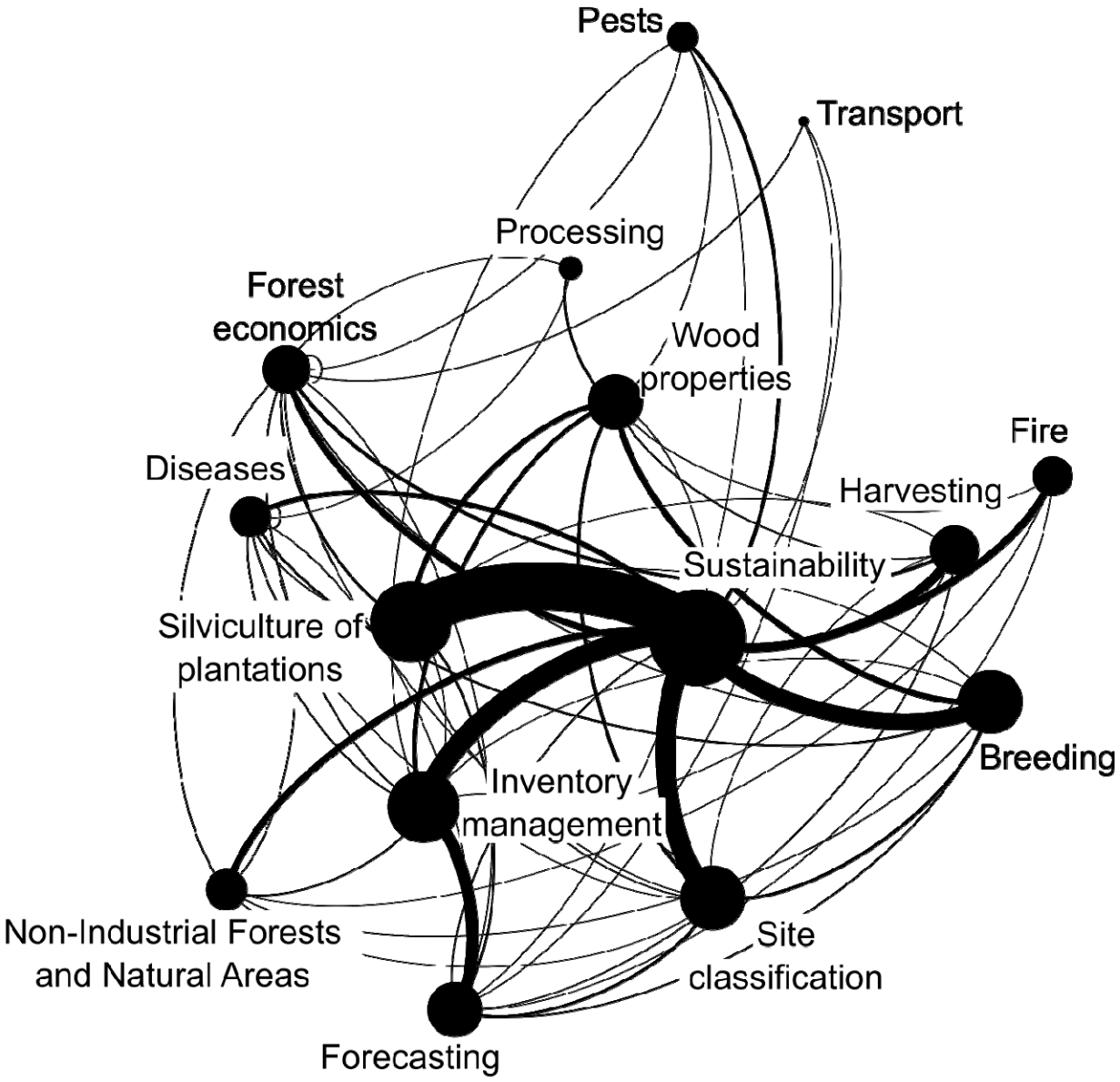

Figure 6: Thematical frequency and co-occurrence network diagram

4. DISCUSSION

This research paper aims to identify research patterns within Southern Forests: A Journal of Forest Science through the application of bibliometrics. The journal started as primarily a South African journal. The research institute's country distribution in Figure 3 supports the presumption that the journal's initial aim was to focus on South Africa. During the late 2000s, the journal's name and focus changed to include research from the Southern

Hemisphere. During the last decade, South African research institutes represented only 51% of the institutes that published in the journal. This metric illustrates that the name and focus change was successful in broadening the journal's international exposure.

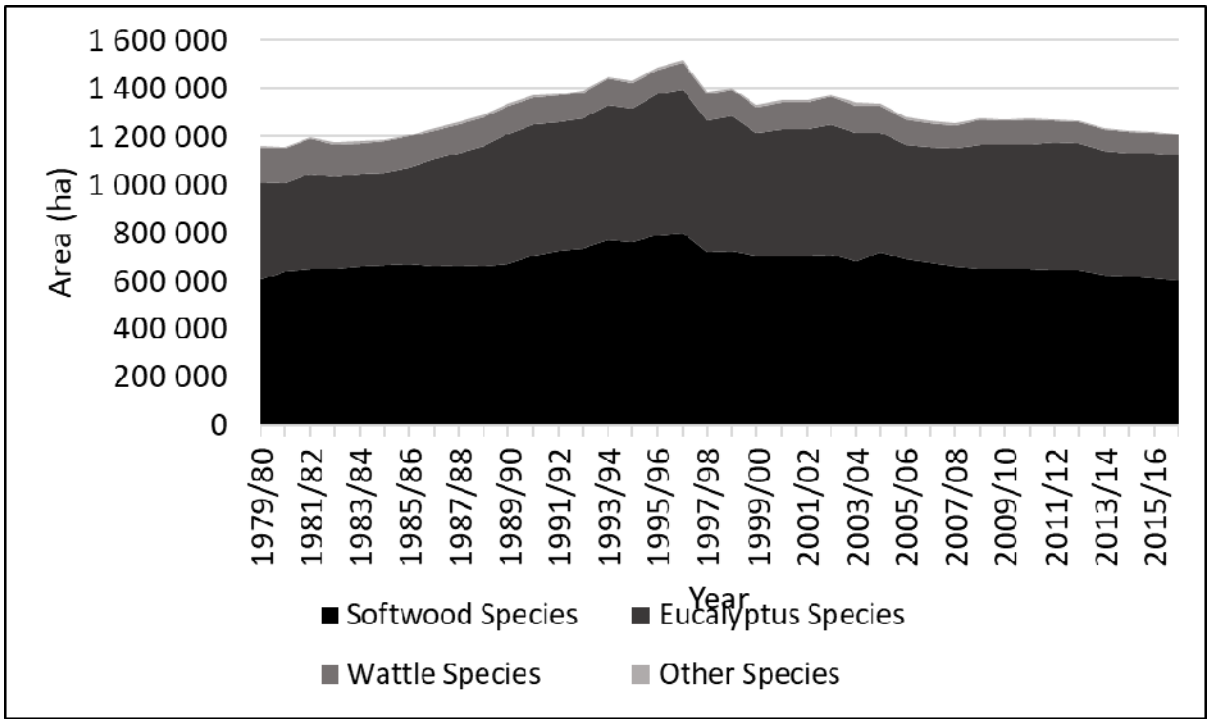

Figure 7: Planted area per genus in South Africa

The primary genera planted in South Africa are *Pinus*, *Eucalyptus* and *Acacia* (Wattle) (See Figure 7). Thus, one would expect that the majority of research would also focus on these genera. The results confirm that most of the research considered *Pinus* and *Eucalyptus.* A smaller portion considered *Acacia* (Wattle), and very little research was performed on other species. Subsequently, the genus distribution and research per genus follow the same pattern. This supported the journal's aim to publish on fast-growing species, especially planted species.

The keyword frequency analysis supported the popularity of *Pinus*- and *Eucalyptus*-related research. The dominant species included *P. patula* and *E. grandis*. The analysis

also indicated that research focused on productivity, genetic improvement, *Fusarium circinatum*, silviculture, and, to some degree, wood properties. This supports the journal's original objectives to promote scientific practices to improve productivity regardless of various threats, including pests, disease, and climate change, as Du Toit (2011) argued.

The frequency and co-occurrence network diagram for author affiliations indicated mainly two clusters of collaborations between research institutes. Stellenbosch University has provided higher education in forestry since 1932 (Wicht, 1958) and, to date, is the only University in South Africa that provides graduate and postgraduate B.Sc degrees in forestry and wood science (Stellenbosch University, 2022). Considering Stelllenbosch University's established role in South African forestry, it is also no surprise that it achieved the most publications and is also connected to a high number of research institutes. Mondi and Sappi represent the largest amount of contributions from a corporate point of view. Both of these companies collaborate extensively with various research institutes. The results indicate that the journal has remained a forum that stimulates open discussion from diverse institutions as originally intended.

SML is approximately 15% less reliable than manual coding, is not suitable for low-frequency categories, and has trouble classifying text that would have been difficult for a coder. However, even with very little text preprocessing, SML has been found to be robust for classifying large quantities of data (Scharkow, 2013). The text classification analysis indicated that sustainability was a popular topic. This is aligned with the journal's original objective to ensure the prosperity of the forestry industry. Considering that most research aims to achieve at least some type of benefit, either economic, social or environmental, it makes sense that this aspect was found to be included in many articles.

It is interesting to note that silviculture was the second most popular topic. The journal's aim and scope include fast-growing and planted forests, and for this reason, it can be expected that this theme is prevalent. Inventory management, breeding and site classification is also primarily associated with plantation forests and support the journal's main aims and scope.

Considering the significant financial and environmental impact that transport could have on a forestry company (Ackerman et al., 2017), it is surprising to notice that it is the theme with the least amount of publications. During the period considered, there were only five publications about transport. The relatively small amount of forestry specific researchers in South Africa could contribute to the low amount of publications in specialised fields like forestry transport. The occurrence of only one article relevant to fire management in the last decade is also an interesting finding.

This paper primarily used bibliometric methods to analyse the journal. Future studies may include qualitative analyses to enhance the narrative to journal and forestry perspectives. Detailed qualitative analysis may also shed light on the reasons for the low numbers of publications related to transport and fire management. The adaptation to climate change, the policies to promote the use of sustainable materials (to counter climate change), and the implementation of value-adding strategies, like mass timber construction (Grobbelaar & Visser, 2021), are research areas that could be expected to increase in popularity.

5. CONCLUSIONS

This paper evaluated publications in Southern Forests: A Journal of Forest Science using bibliometrics. The author's country analysis confirmed that the journal's change in name

and focus was effective in luring researchers outside of South Africa to take part in the journal. The journal's focus on fast-growing plantations is clearly illustrated through the high number of publications related to the *Pinus* and *Eucalyptus* genus, as determined through the genus search and keyword analysis. The pattern is associated with the plantation area per genus in South Africa. Considering that the journal originally focused on South African plantations, the association is to be expected. The researcher and researcher affiliations analysis indicated that there are mainly two research clusters that published in the journal. The one cluster is around Stellenbosch University (the research institute with the most contributions and the most expansive network of collaborators). The other includes several South African universities and research institutes. The theme frequency analysis indicated that the prevalence of silviculture, breeding, inventory management, and site classifications is associated with the journal's main aim and scope. However, it also indicated that very little research concerning transport had been published, and more recently, there has also been very little research regarding fire management. This paper didn't intend to determine the causes for these patterns. Future research may consider determining this. However, both aspects play a critical role in forestry management for short-rotation plantations, and thus the lack of research into these aspects may be worrying.

The results of analyses like these can be used to identify strengths and weaknesses in the available literature to prioritise research effort. The methods used in this research can be automated, and a web service can be developed. The implementation of such a service may assist researchers, industry, the public sector and other stakeholders to continuously reprioritise research and research funding to assist with achieving the objectives of forestry stakeholders in the Southern hemisphere. A tool of this nature can

also be used by the editors of the journal to confirm how changes in the aims and scope influence the reach and impact of the journal. Future research may include other journals to expand the scope of the analysis or compare the results of various models and configurations.

6. REFERENCES

Ackerman, P.A., van der Merwe, E.A. & Pulkki, R.E. 2017. A South African softwood sawtimber supply chain case study. *Southern Forests*. 79(4):329–338.

Bastian, M., Heyman, S. & Jacomy, M. 2009. Gephi: an open source software for exploring and manipulating networks. in San Jose: International AAAI Conference on Weblogs and Social Media.

Bredenkamp, B.V. & Upfold, S.V. 2012. *South African Forestry Handbook*. 5th ed. B.V. Bredenkamp & S.J. Upfold (eds.). Menlo Park, South Africa: The Southern African Institute of Forestry (SAIF).

Broadus, R.N. 1987. Toward a definition of “bibliometrics”. *Scientometrics*. 12(5–6):373–379.

Elsevier, B.V. 2021. *Scopus*. [Online], Available: https://www.scopus.com/.

Grobbelaar, S. & Visser, J.K. 2021. An analysis of South African sawmilling competitiveness. *Southern Forests*. 83(1):28–37.

Informa, U.K. 2021. *Journal Southern Forests: a Journal of Forest Science*. [Online], Available: https://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=tsfs20 [2021, March 18].

Little, K.M. & Nadel, R. 2014. Testing pelargonic acid and pyraflufen-ethyl with glyphosate

as alternatives to paraquat dichloride for the preparation of fire-break tracer lines at Underberg, South Africa. *Southern Forests: a Journal of Forest Science*. 76(2):67–73.

Marais, G. 2007. President's message on the first issue of the Southern Hemisphere Forestry Journal. *Southern Hemisphere Forestry Journal*. 69(1).

Martin, S., Brown, W.M., Klavans, R. & Boyack, K.W. 2011. OpenOrd: an open-source toolbox for large graph layout. in *Proc.SPIE* Vol. 7868.

Microsoft, A.T. for. 2015. [Online], Available: https://gallery.azure.ai/Experiment/Text-Classification-Step-1-of-5-data-preparation-3.

Notz, M., Grambau, J. & Hitzges, A. 2019. Evaluation of Sentiment Databases: A Comparison of Sentiment Databases through Social Listening Statements and Azure Machine Learning Studio. *ACM International Conference Proceeding Series*. 8–12.

Oosthuizen, R. & Pretorius, L. 2020. Bibliometric analysis of technology management research topic trends. *Towards the Digital World and Industry X.0 - Proceedings of the 29th International Conference of the International Association for Management of Technology, IAMOT 2020*. 701–712.

Owen, D.L. (in press). Editorial: Southern forests: How are you assisting in improving impact? *Southern Forests: a Journal of Forest Science*. 70(3):iii.

Owen, D.L. 2002. Southern African Forestry Journal -it's history. *Southern African Forestry Journal*. 193(1):81–82.

Owen, D.L. 2003. SA Forestry Journal on the internet. *Southern African Forestry Journal*. 198(1):84.

Owen, D.L. 2005a. New-look Southern African Forestry Journal – progress. *Southern African Forestry Journal*. 205(1):1–1.

Owen, D.L. 2005b. What is expected of the SA Forestry Journal in future? *Southern*

African Forestry Journal. 203(1):1–2.

Owen, D.L. 2006. Editorial Notices. *Southern African Forestry Journal*. 164(4232):1059.

Pliuhin, V., Korobka, V., Karyuk, A., Pan, M. & Sukhonos, M. 2019. Using azure machine learning studio with python scripts for induction motors optimization web-deploy project. *2019 IEEE International Scientific-Practical Conference: Problems of Infocommunications Science and Technology, PIC S and T 2019 - Proceedings*. (4):631–634.

Pritchard, A. 1969. Statistical bibliography or bibliometrics. *Journal of Documentation*. 25(4):348–349.

R Core Team. 2014. R: A language and environment for statistical computing/R Foundation for Statistical Computin. Vienna, Austria, 2014. *URL: http://www. R-project. org*.

Rajput, B.S. & Khare, N. 2015. A survey of stemming algorithms in information retrieval. *IOSR Journal of Computer Engineering*. 17(3):76–80.

Sarkar, S.D. & Goswami, S. 2013. Empirical study on filter based feature selection methods for text classification. *International Journal of Computer Applications*. 81(6).

Scharkow, M. 2013. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality and Quantity*. 47(2):761–773.

Smith, C. 2012. Sustainability of Forest Plantations. in *South African Forestry Handbook* 5th ed. B.V. Bredenkamp & S.V. Upfold (eds.). Menlo Park, South Africa: The Southern African Institute of Forestry (SAIF) B.V. Bredenkamp & S.V. Upfold (eds.). 271–280.

Stellenbosch University. 2022. *Dept. of Forest and Wood Science*. [Online], Available: http://www.sun.ac.za/english/faculty/agri/forestry/about-us [2022, May 05].

Du Toit, B. (in press). A reflection on achievements and opportunities for Southern Foresters in 2011, the International Year of Forests. *Southern Forests: a Journal of Forest Science*. 73(2):iii–iv.

Wang, X., Mccallum, A. & Wei, X. 2007. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. in *Seventh IEEE International Conference on Data Mining (ICDM 2007)* IEEE.

Wicht, C.L. 1958. TRAINING IN FORESTRY AND WOOD TECHNOLOGY AT THE UNIVERSITY OF STELLENBOSCH. *Empire Forestry Review*. 37(3):311–315. [Online], Available: https://www.jstor.org/stable/42600506.