Scientometric Analysis and Topic Modeling for Farm Financial Decision Modeling(FFDM)

Gedefaye Achamu (achamug@gmail.com)
Bahir Dar Institute of Technology, Bhir Dar University, Dire Dawa University Institute Of technology
https://orcid.org/0000-0002-4765-5982

Eshetie Berhan Atanew
Addis Ababa Institute of Technology, Addis Ababa University https://orcid.org/0000-0002-7125-4994

Sisay Geremaw Gebeyehu
Bahir Dar Institute of Technology, Bahir Dar University

Shimelis Tilahun
Bahir Dar Institute of Technology, Bahir Dar University

Systematic Review

Keywords: Farm, Decision, Scientometric, Topic Modeling, Cluster, Mapping

Posted Date: April 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1555059/v1

Scientometric Analysis and Topic Modeling for Farm Financial Decision Modeling (FFDM)

Gedefaye Achamu Meretie[1], Eshetie Berhan Atanew[2], Sisay Geremaw Gebeyehu[3], and Shimels Tilahun[4]

[1] PhD. Student, Bahir Dar Institute of Technology, Bahir Dar University,
Lecturer, Dire Dawa Institute of Technology, Dire Dawa University
achamug@gmail.com
[2] Associate Professor, Addis Ababa Institute of Technology, Addis Ababa University
berhan.eshetie@gmail.com
[3] Associate Professor, Bahir Dar Institute of Technology, Bahir Dar University
sisayg78@gmail.com
[4] Assistant Professor, Bahir Dar Institute of Technology, Bahir Dar University

Abstract.
Scientometric analysis and topic modeling deployed in this survey in order to map and structure knowledge in the state of art of farm financing decision making. Application of scientometric approach based on bibliometric analysis help to capture the interest of scholars and organization to the problem domain and also found essential in clustering keywords, authors and publication to some sort of schemes and hence mapping of knowledge through visualizing map. Topic modeling as complement of bibliometric analysis further extends and makes clear how those keywords semantically related and their contribution to topics while clustering based on bibliometric analysis is binary. Using both methods and approach refed as TAKe, publications from different source and type examined and analyzed for state of art of farm financial decision making. The survey then signals, in the state of the art of farm financial problem, financing has been treated for descriptive or exploratory purpose than as decision variable hence predictive consultancy than prescriptive advisory in farm financing investment.

Keywords: Farm, Decision, Scientometric, Topic Modeling, Cluster, Mapping

1. Introduction

With modelling as practice of capturing real phenomenon via, most of the time, abstraction, a yet important issue is system classification, modeling objective and approach or purpose (normative vs positive)(S. R. Johnson & Rausser, 1977, p. 164). Modelling objective might include either of (i) descriptive, (ii) explanatory (iii) predictive or (iv) decision model while the last two of this classification however in the contemporary research interest are highly emphasized due to the potential they provide in answering policy question in the problem domain say for instance in agriculture economics and financial problems. Since, surely to say that, every farm decision is constrained and most of farmers especially in developing country potentially exposed to various risks including production, financial and marketing risk (Harmon, 2018; Ruete, 2015), proposing specific model or framework is impossible and that is the case that makes farm investment decision challenging. Furthermore, from financial management (FM) perspective, decision generally varied in both impact and frequency to the economy (Hilkens et al., 2018).

The JRC's scientific and technical report on investment behavior in conventional and emerging farming systems under different policy scenarios considers the importance of reviewing literature to capture insights over (i) determinants (ii) effect of policy and (iii) classification of quantitative tool for analyzing farm investment decision (Vittori Gallerani, et al 2008). Literature in farm investment therefore has shown a progress in two versions: (1) in line with general economic literatures during 1950s and 1960s, (2) specific to agricultural economics literature explored during 1990s. Literatures, mainly starting 1980s, have focused on number of investment -related topics and finding from the report essentially reveals the gap related to various issues. These includes: (a) instruments to deploy, (b) model adaptation towards farmer preference and expectation; (c) closer attention to the connection between investment, technical change and learning; and (d) a more empirically relevant treatment of the decision maker's(farm household's, or farms) objectives. Moreover, failure in policy analysis and treating it separately even in the recent studies is a major area need intervention. Epistemological and ontological approach has been recommended in farming system to work as both interdisciplinary approach and multidisciplinary integration in order to incorporating the hard and soft version of the problem domain like farm financing decision model (FFDM). Practically, these all, however, are addressed separately instead of deploying principles and methods not only from various discipline but also from approaches at and from different perspectives. Note that, though drowning in information, academia is still starving for knowledge(Naisbitt & Aburdene, 1990) and implies that we are wanted to organize such bulk information and transform it to knowledge. This is essentially demanding as today's decision-making environment is influenced by such dramatically accelerated and bulk of data from science, technology, and innovation (STI) activities. One way of bringing to front the

selection over methods, tools, and approaches in one hand, and not to overlapping and repetition on the other hand, is evidence identification through systemic review and meta-analysis. Moreover, reviewing literature also help to capture development trends of discipline and how such temporal change has altered the entire topic(Han, 2020). Concerning to topical change and intellectual structure in Library Information Science (LIS), Han classified literature review methods as (i) continent analysis, (ii) bibliometric method and model based approach. Highlighting the first as the task of scheme classification of research content to detect research development and was focused around 1970s-1980s is sufficient here and readers are referred for detail to (Han, 2020 & refernces therein). Since, in one hand, bibliometric methods are prevalent approaches in evaluation studies through its techniques (see; Han, 2020, p. 2563), Keywords analysis, Citation analysis, Co-occurrence and bibliographic coupling. On the other hand, model based approaches are recent methods towards capturing intellectual structure of a scientific domain and overhands the remaining two in term of examining larger corpus, we give attention here. Consequently, bibliometric analysis, which has been brought to the age of big data (Chen et al., 2021) for mapping such evidence identification is the central scheme of this paper for the problem domain under investigation.
Objective in this survey therefore is to assemble a comprehensive library of literatures on the pattern of decision-making over farm financial decision problem in order to study the progress of the problem domain over time. Since the area under investigated is polycentric and is composed from wide array of disciplines and subject area among others, finance, business, economics, accounting, agriculture, assessment of literatures at perspective of domain analysis that most of the time has been demonstrated separately is motivational for topic modeling. A two-stage approach then followed at which the first stage of the survey is an initial bibliometric analysis based on bibliographic metadata and demonstrated using open access analysis software VoSviwer. Using the result of the bibliometric analysis, in the second stage, a topic modeling (TM) approach from contemporary machine learning(ML) paradigm and fundamental of natural language processing (NLP)considered to identify topics to such interlinked disciplines that are believed to show a correlation in the decision making process of a farmer and/or financial institutions for instance. Ultimately, deployment of TM is for discovering thematic structure from the corpus of documents (publication) to the problem domain. Keywords and abstract from retrieved publication respectively considered as vocabulary and corpus of document while publication extracted from Zotero reference manager used for comparison purpose since it was extracted purposely specific to the problem being studied.
Remaining sections therefore extended with setup of methods and materials in section 2, and result and analysis in section 3 to state of art of the interest while section 4 discussions and interpretation of the result obtained. Finally conclusion, and take notes presented in section 5.

2. Methods and Materials

Since the purpose of this scientometric and topic modeling is both to map and structure knowledge obtained from publication, survey design and description of tool and technique are prime tasks.

2.1 Survey Framework

Figure 1 presents the two-stage analysis approach that first demonstrates a bibliometric analysis and followed by topic modeling based on the findings especially on the three important components of a publication, title, abstract and key words abbreviated as TAKe . Our bibliometric analysis in the first stage therefore starts with selecting search engine through both generic and extended query terms supported by Boolean operator “OR”. Setting inclusion and exclusion criterion are also part of this step while the analysis step mainly focused on those two-bibliometric analyses: Citation analysis and co-occurrence analysis by highlighting to those remaining bibliographic analysis techniques. The second stage on the other hand begins with some preprocessing task to make ready data for topic modeling.

2.2 Search Engine Selection (SES)

In their investigation over Google scholar, Microsoft Academic, Scopus, Dimensions, Web of science, and Open citations’ COCI for a multidisciplinary comparison of coverage via citation, Alberto Martin –Martin et al.,(2021) ranked such six data source. The rank in descending order of citation percentage to 2,515 English language published documents with 3,073,351 citation: Google scholar (88%), Microsoft academic (60%) which share, however, with Scopus and Web of science (WoS) respectively as 82% and 86%. Scopus then place in the third rank while the fourth one is Dimensions (54%) than that of WoS. Dimensions still take the share of 84% with Scopus and 88% with WoS citation. Furthermore, it found more citation than Scopus in 36 categories, more than WoS in 185. According to their investigation, limitation regarding to Dimensions for that analysis period was its failure to cover humanity fields. It could be realistic to generalize that its editorial policy for Google Scholar to share higher percentage not only for this finding but also in general cases, i.e. Google scholar follows an inclusive and automated approach (Martín-Martín et al., 202)

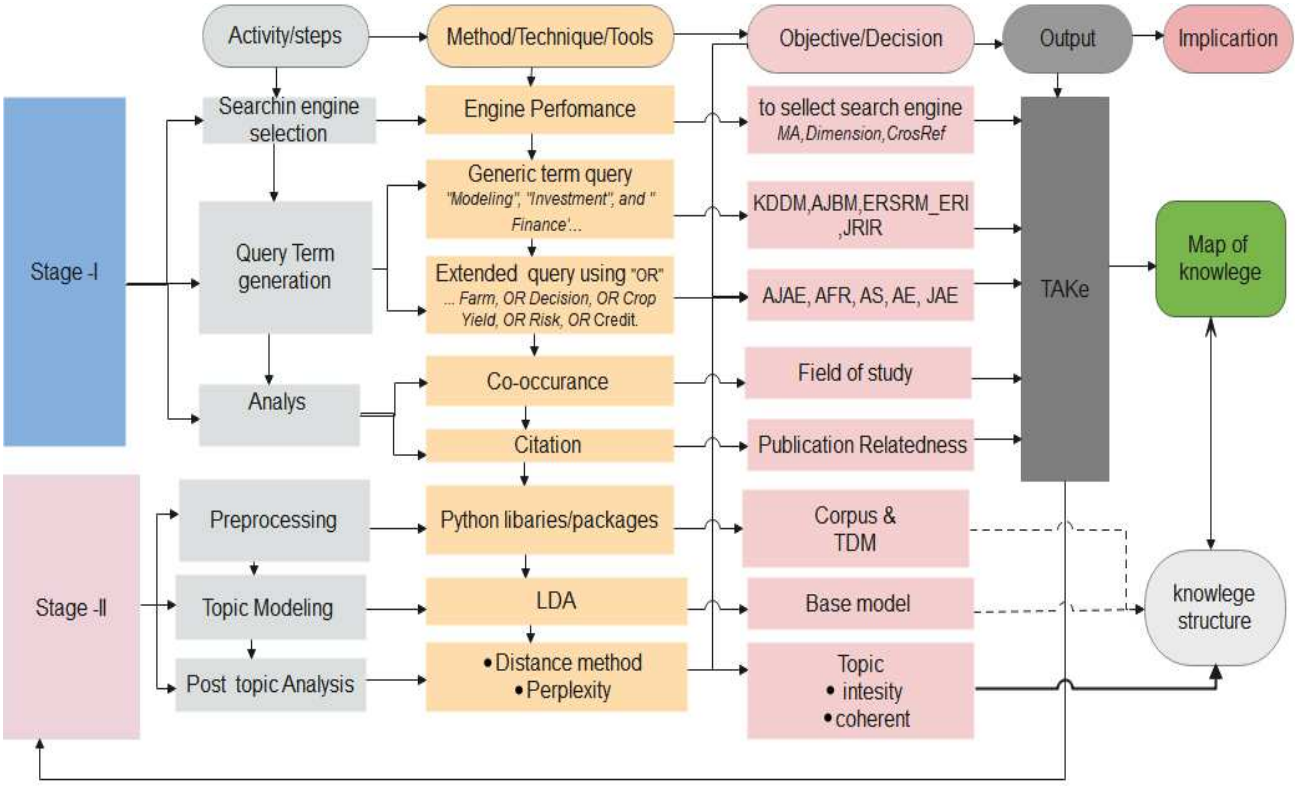


Figure 1 Survey Framework of Two stage FFDM topic modeling

Though, Microsoft Academic search (MAS) ended in 2012[1], a new platform called Microsoft academic (MA) launched in 2016 whereas new scholarly search database, Dimensions, launched by digital science (Hook et al., 2018) with a freemium model i.e. only advanced functionality like API (Application Programing Interface), which designed to facilitate bulk access in Microsoft Academic (Wang et al., 2020), According to Hook et al., dimensions has tried to include grants, patents and clinical trials besides of books, book chapters and conference proceedings in the publication index. As a newly approach towards data source, significance of dimension has been compared to other data source as has done in Martin et al , Vicent PGuerrero-Bote,et al., and reference therein. Comparison made by Martin et al is as discussed so far while investigation by Vicent PGuerrero-Bote,et al., is only between dimensions and Scopus at country and institutional level. We rather are ignorant for theses research interest, comparing databases at perspective of country and institution, while comparison based on coverage, still dimensions guarantees a 25% greater than Scopus (Guerrero-Bote et al., 2021). According to (Martín-Martín et al., 2021 and reference therein) WoS covers about 75(155) million records in its Core Collection (regional and subject specific) citation index, Scopus over 76 million records and Google scholar over 300 million records.

In general, as depicted in the approach, Google scholar search (GSS), Microsoft academic (MA), Crossref, and Dimension searching engines selected for this investigation.

2.3 Query Term Generation and Publication Retrieving

A two-step analysis methodology followed at which the first is based on using generic term *Modeling, Investment, and Finance* to capture the state of the general problem dimension while in the second analysis a further investigation demonstrated using additional query terms using the "OR" Boolean operator. This is since both the publication and source retrieved using such generic term does not warranty for drawing a conclusion specific to the case of FFDM, additional query terms Boolean operator "OR" as Farm, OR Decision, OR Crop Yield, included.

For both co-authorship and citation analysis a(2,1) inclusion and exclusion criterion followed to imply that an author and organization should have two documents with minimum of a single citation not to narrow down role of both organizations' and authors' in the problem interest and indeed this is further justified by imposing minimum citation for document to be unit. In all the retrieving process, a further restriction imposed is to retrieve publications from primary source and all the publication must have a DOI.

[1] https://web.archive.org/web/20170105184616/https:/academic.microsoft.com/FAQ

Since maximum number of literature that Vosviewe , an open accessed tool for publication retrieval, can analyze is 5000, a separate analysis made for the source using MA and Crossref in VoSviewer by selecting journals based on their performance rank obtained from both Dimensions analysis using VoSviewer and PoP. Moreover, VoSviewer also allow us to analyze our trial quest through the reference manager (Zotero) as a data source. Table 1 therefore presents data type and data source used in this investigation.

Table 1 . Data Type and Data Source Followed

Category	Items	Search engine /database			
		Microsoft Academics (MA)	Crossref	Dimensions	Google [scholar] Search (GSS)
Data Type	Bibliographic	✓	✓	✓	✓
	Network Data			✓	✓
	Text Data			✓	✓
Data Source	API download	✓	✓		
	Database File			✓	
	Zotero Reference Manger		✓		✓

According to Purnwok ibn Sangadi[2], bibliometric analysis as quantitative tool of assessing the academic publication, does not measure science, scientist, or scientific productivity rather help to map science(see stage-I in Fig.1), which is both complex and cumbersome (Bibliometric, 2017). According to(van Eck & Waltman, 2017) to cluster publication determining publication relatedness is the first task either based on citation relation or word relation. Citation relation generalizes direct citation (DC), bibliographic coupling (BC), and co-citation (COC) whereas word relation is about word sharing based on either title and/or abstract and/or full text. Since, BC shows relatedness between publications, that cites the same publication; and citation relation is about publication cited by the same publication (Shibata et al., 2009), DC better detects research fronts than COC and BC. Whereas for (Boyack & Klavans, 2010) DC is rather less accurate and these two generalization by themselves are true if long and short period (less than five year) respectively imposed as inclusion and exclusion criterion(Waltman & van Eck, 2012).

According to Waltman & van Eck, COC and BC requires two DC and hence indirect methods they are. Since aim of this survey is to explore the extent and depth of research history, approach and mechanism regarding agriculture and finance particularly crop production as subsystem of farming that is polycentric inherently where both multidisciplinary and interdisciplinary are attractive, analysis method based on co-occurrence and co-citation more preferred, which, however, doesn't mean others are not touched. For both co-authorship and citation analysis a(2,1) inclusion and exclusion criterion followed to imply that an author and organization should have two documents with minimum of a single citation not to narrow down role of both organizations' and authors' in the problem interest and indeed this is further justified by imposing minimum citation for document to be unity. In all the retrieving process, a further restriction imposed is to retrieve publications from primary source and all the publication must have a DOI.

2.4 Preprocessing

As classical text mining method, topic modeling helps to represent documents (publication) as space vector to compute and analyze similarity among vector and documents respectively. Left side of fig.2 gives topic modeling structure for LDA (Latent Dirichlet Allocation) at which a three-layer Bayesian probability model composed of N-words, k- topic (prior), and M-text or document. Purpose in LDA is to train for the output of ψ (the distribution of words for each topic K) and ϕ, the

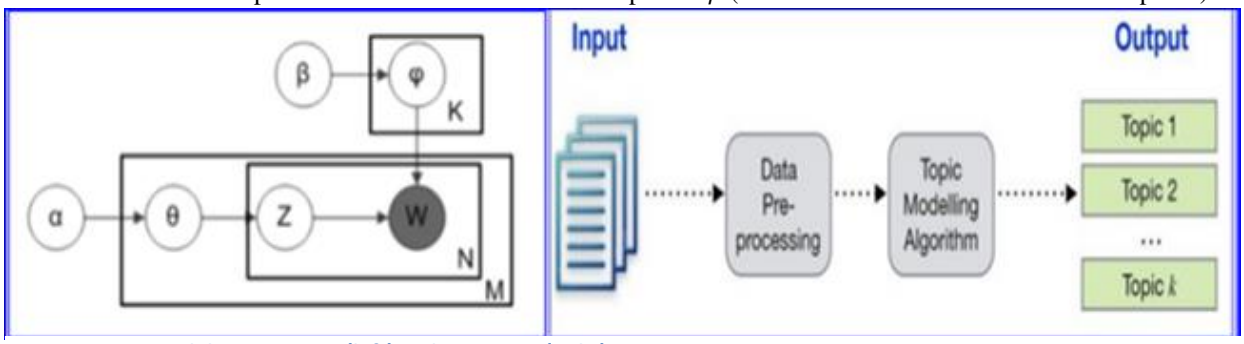


Figure 2.Topic modeling structure (left) and processes (right)

[2] http://kit.ft.ugm.ac.id/sp/subjects/guide.php?subject=br

distribution of topics for each document i using the two most Dirichlet prior concentration parameters that represents (i) document-topic density(α- $parameter$) and (ii) topic-word density (β -$parameter$). With a higher α (β), documents (topics) are assumed to be made-up of more topics (words) and result in more specific topic (word) distribution per document (topic).

Moreover, due to evolution of issues and concepts dynamic topic modeling is also available while in this survey we restricted ourselves Topic modeling with LDA and *BerTopic*, a topic modeling technique that uses transformers (BERT embedding) and class-based TF-IDF to create dense clusters and it also allows to easily interpret and visualize the topics generated. Three stages in *BerTopic* include (i) Embedding the textual data (documents), (ii) Cluster Documents and (iii) Create a topic representation. Implementation of BerTopic in this analysis is based on "paraphrase-MiniLM-L6-v2 sentence transformers since the semantic similarity is for single, i.e., English language publication only. In its best, *BerTtopic* uses the more preprocessing step called UMAP (uniform manifold approximation and projection) than LDA and scikit-learn and even is better to the competitive one in the state of art t-SNE2 to boost the performance of density based clustering.

By leveraging transformers and, c-TF-IDF, BerTopic helps to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions. The modeling followed here is using genism, Scikit respectively from, Natural language processing (NLP) and machine learning, and BerTopic. In any of the packages, the priori task is data preparation and preprocessing to obtain dictionary and corpus respectively to map word to unique id and bag of words-thus both dictionary and bag of words(corpus) used as input for topic modeling.

2.5 Dictionary and Corpus Preparation

Using title, abstract and keywords (TAKe) dictionary and corpus prepared for topic modeling. Those obtained 894 keyword from AJAE then forms 894x791 spares matrix with 1620 stored elements and transformed to their root using WordLemmatizer function and reduced to 894X760 with 1591 stored element. This small reduction was due to the sufficiently sparsity (0.998) of the original corpus and about 1591 unique vocabulary present in our word list (corpus). As indicated in the word cloud of fig.4, Economic, Market, Risk takes the higher weightage and followed by decision, model, management and capital.

On the other hand, topic modeling based on the analysis of title and (abstract) of publication from reference manager (Zotero), done using Scikit learn, genism and BerTopic. Using tfidf Vectorizer, CountVectorizer from *sklearn.feature_extraction.text*, topic modeling based on Scikit package deployed with test size of 0.2. Each of the documents first converted to list of words using *countvectorizer* and transformed into 351x1001(241x4865) sparse matrix with 3320 (22986) stored elements in compressed sparse row format to title and (abstract) respectively. Once again sparse matrix and obtained as 0.99 (0.98). The vectorized documents now converted to bag-of words (corpus) using *doc2bow* and a total of 194 (1555) unique words after removing infrequent and common words unique words in initial 351 (241) documents with unique word 881(4196). This is done by filtering out words that occur less than 3 documents, or more than 60% of the documents. Hence, pruning the common and rare words, we end up with only about 22.02 %(37.06%) of the words.

3. Results and Analysis

3.1 Results

Three important source of publication deployed, MA, CrossRef and Reference manager in-depth besides the Dimensions that simply helps to extract keywords in particular.

Microsoft Academics(MA)

Table2&3 presents result of bibliometric analysis for selected sources while the graphical illustration portrayed in Fig.3 displays the general bibliometric analysis result of FFDM. Since, a two stage query term regeneration approach followed, first using generic terms and followed by terms with Boolean operator, the priori gives no sound results to the problem questioned. Explicitly, using query term of *modeling, investment, and finance* only 185 publication retrieved on Microsoft Academic (MA) and of those 854 keyword for analysis method of co-occurrence with field of study as unit of analysis 129 meets the threshold, hence for the combination of (*occurrence, TLS*), economics as a keyword obtained to take the top occurrence (85,403) and followed by business (50,253), investment (41,226) finance (38,191), and econometrics (30,149). For the Boolean operator implementation, a cross validation approach followed since VoSviewer software potentially gives a maximum of 5000 publication only, those sources in table 4 like Journal of the American Association statistics (JAA); Journal of Financial Economics (JFE); Econometrica (Econ.rica);Machine learning(ML) also found essential from the PoP analysis.

Table 2. Analysis of Top four journals correlated with problem studied using dimensions database (total analysis)

Source	Documents	Citations	TLS
American journal of agricultural economics(AJAE)	70	1478	115
Agricultural finance review (AFR)	86	856	83

Agricultural systems(AS)	71	2349	67
Agricultural economics (AE)	35	666	53
Journal of Agricultural Economic(JAE)	9		

Note: *TLS=total link strength*

Table 3. Analysis of Top four journals correlated with problem studied analysis (between selected sources)

Source	Documents	Link	TLS	Av.norm.citation
AJAE	70	4	50	0.79
AFR	86	4	33	0.40
AS	71	4	23	1.35
AE	35	4	21	0.65
JAE	9	4	21	0.87

Moreover, as described in the methodology section, those results of table2&3 and fig.3 are results from databases and a total of 6858 publications retrieved at which AJAE takes the higher share (22.2%) and followed by AS (19.4%). consequently, based on the relevance to the problem domain, five journal as source of publication identified (see table.2) using the TLS, link and citation and a total of 271 publication with 894 keywords at minimum threshold of three, left. Figure 4 then display location distribution of AJAE and conveys that how agricultural research and publication has enhance in the developed country, particularly American universities has put their enormous contribution to the sector.

CrossRef

The result of so far discussed were based on the search engine of Microsoft academic (MA) based on API that is somewhat less restrictive and an alternative exploration made for search engines called Crossref. A crosref based publication retrieving done for further exploration of the bibliometric analysis in the problem domain. Since separate exploration, using Crossref in VoSviewer only possible for single term expected to appear only on the title of publication need to be retrieved. Since another inclusion and exclusion criterion is also required for impossibility of retrieving due to inherent restriction on maximum publication of VoSviewer, again search is restrict on the source that is identified earlier as better- AJAE. Figure 5 then presents the bibliometric analysis result of Crossref database for terms indexed and one important advantage of Crossref based exploration is the possibility to examine at single term which however is a limitation on the other hand.

Reference Manager

In this case those publication intentionally collected and stored in Zotero reference manager utilized.Fig,6 demonstrates that most of the publication are recent (see, left side of Fig.6) and of those 351 retrieved publication 228 are articles (64.96%) and followed by book section(article in series) (10.83%) whereas book, conference paper and thesis (dissertation) takes third, fourth and fifth position. Reports, webpage and blog posts are part of source though supplied little publications. Bibliographic analysis particularly to co-authorship to this source is as depicted in fig.7 at which a total of 733 authors contributed and the horizontal axis in fig.7 implies author publication relationship and arithmetically almost three authors are expected in each of the publication, whereas, as depicted in Fig.7, of those 733 about 50 authors have at least two and at most five publications. The remaining 668 authors, however, does not mean that each has a single publication as they would appear as co-authored of and referenced author. Since documents with more than one authors that accounts about 33.14 %(117/353) the knowledge comes from similar source of knowledge and of course similar research scheme indeed.

One important mechanism to classify to which of the research scheme of those publications can be categorized is to clustering in VoSviewer (see Fi.g.8) those 733 authors. Of those 733 authors, only 14 authors are connected as indicated by the non-gray colors and displayed in networks depicted in Fig. 8 and classified in to three clusters using documents as weight factors and average of publication year as score value of the visualization. Cluster one (Yellow range in Fig.8) therefore ranges averagely from 2013-to- 2014 and above and composed from 7 items (authors). Cluster two (green range) on the other hand ranges from 2008 –to-2013 and composed from 4 items including Lempert and Robert while cluster three (blue range) ranges between 2004 and 2008 averagely. The maximum total link strength (TLS) using full counting method obtained for Lempert and Robert, authors with number of documents equal to four, attains TLS of 16 and using binary counting the maximum TLS is 5.00 by Brige,Johon R. and Louveaux Francois, authors with five document while Lempert and Robert receives a TLS of 1.0.

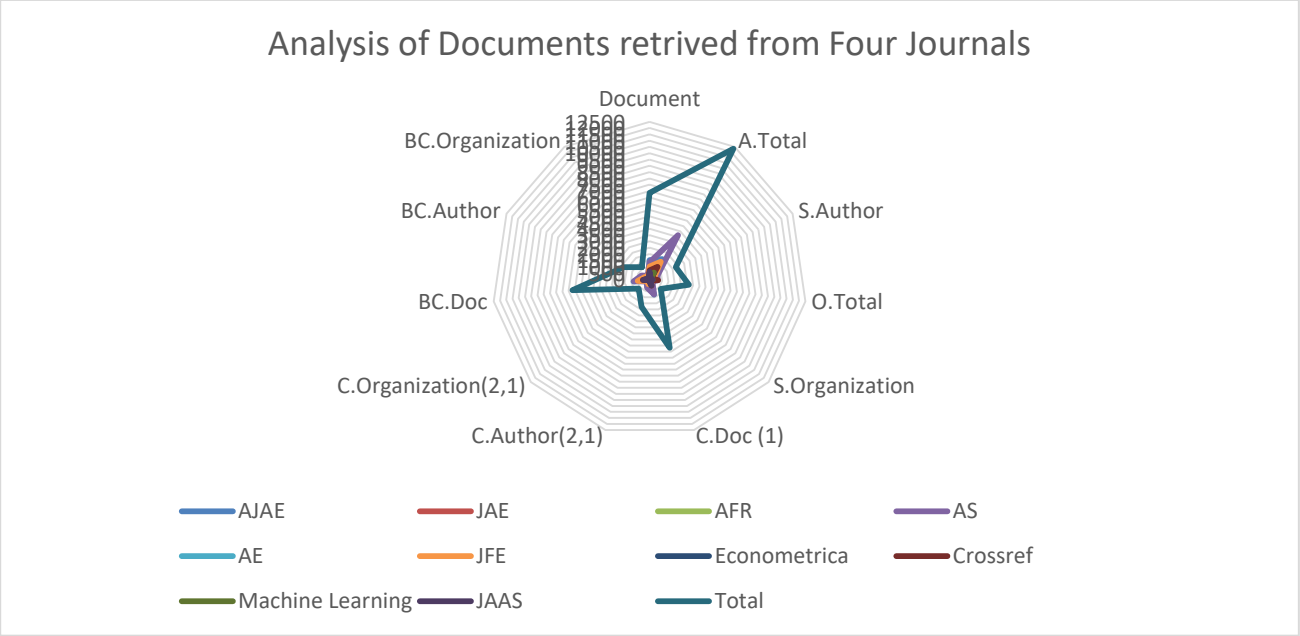


Figure 5.Summary of Analysis of documents form different source

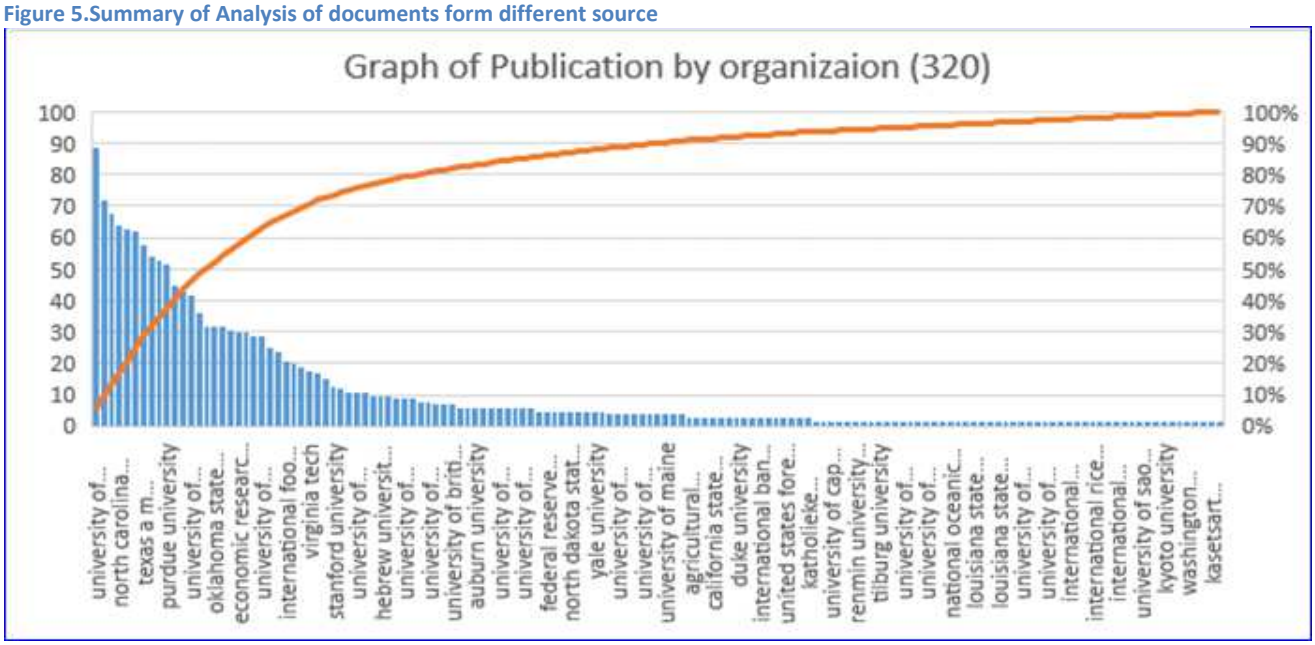


Figure 4.Publication distribution AJAE by organization

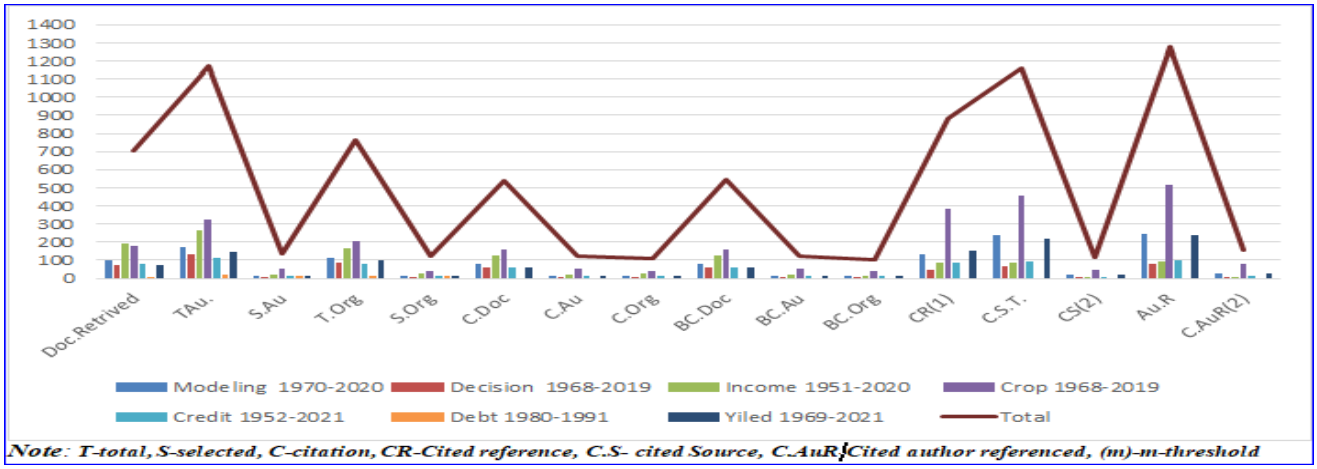


Figure 3.QUery Term Analysis Using CrossRef in AJAE

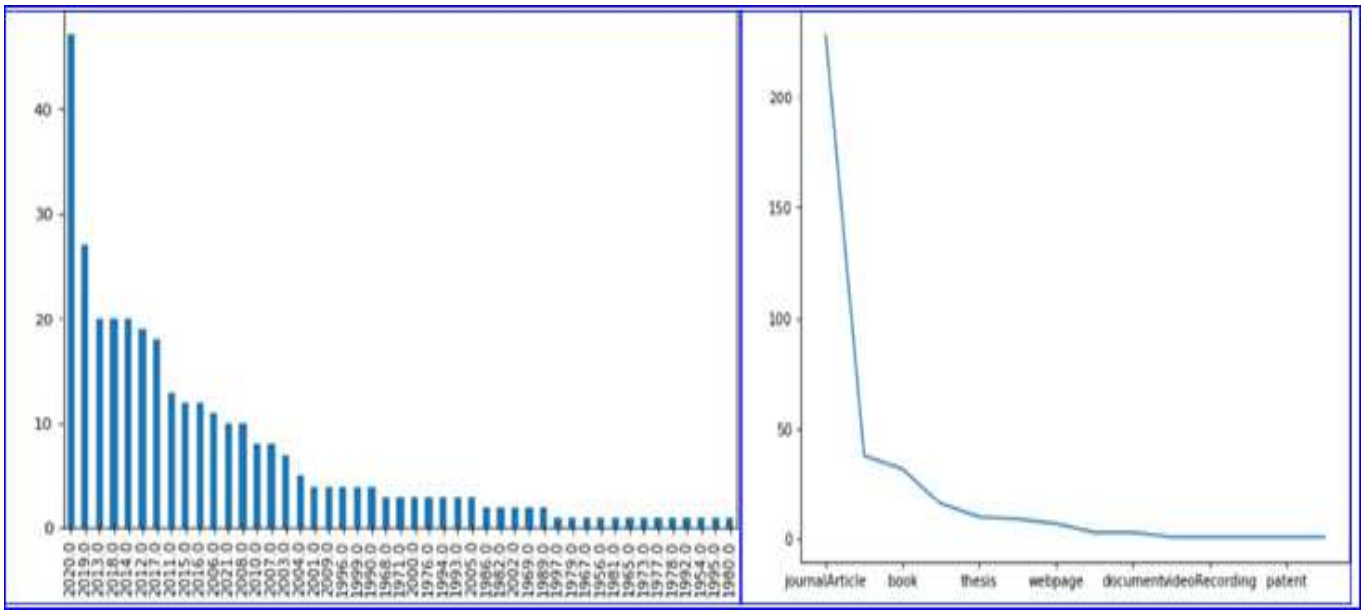


Figure 8.Distribution of Publication retrieved in year (left) and source Type for reference manager

This is shows how full and binary counting methods can be distinguished that can be only observed on the

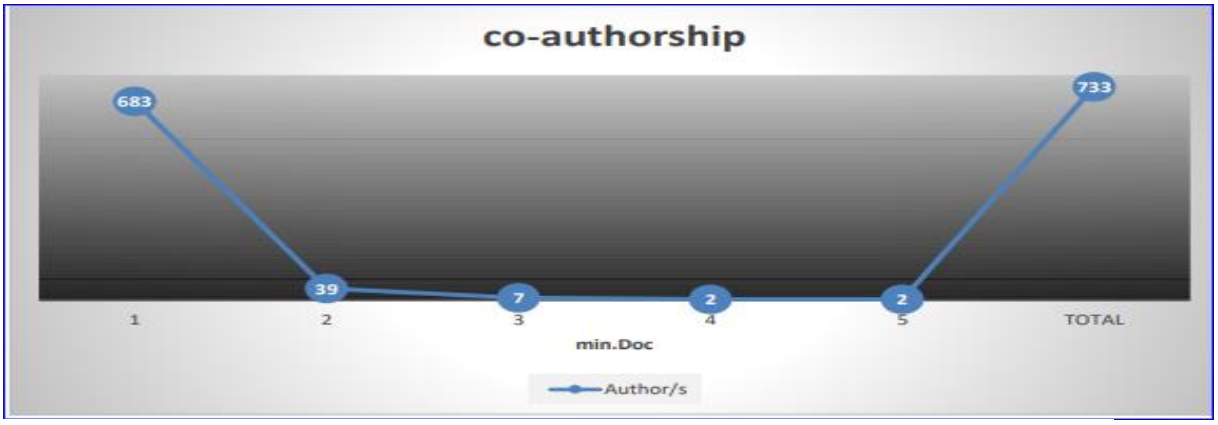


Figure 7.Cluster of Authorships from reference manager source

network link Strength at which the importance of fractional (Binary) counting method to reduce the influence of

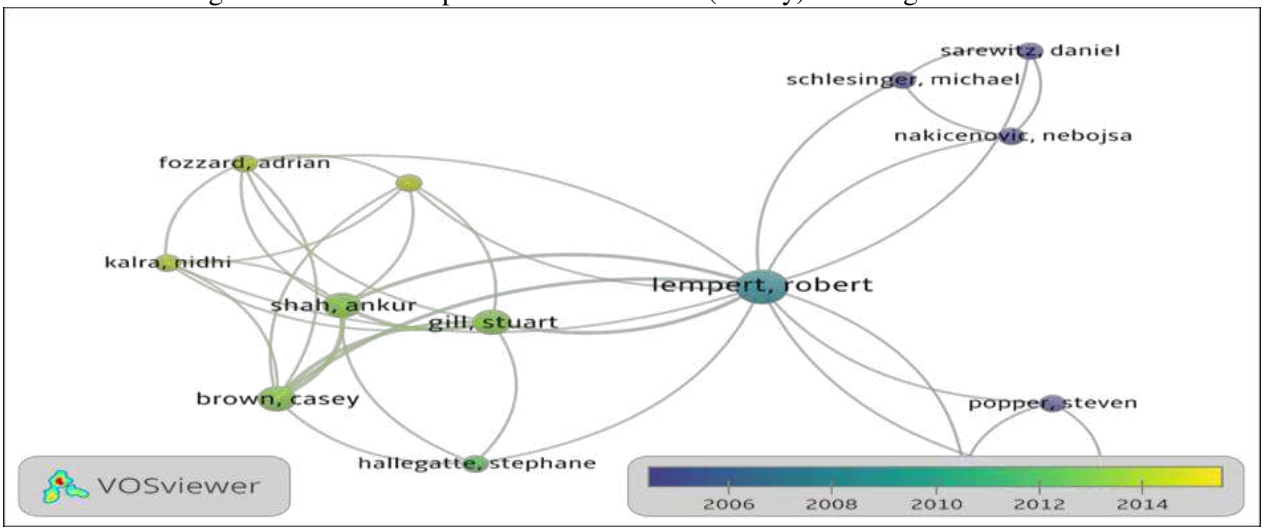


Figure 6.Bibliographic analysis for reference manager data source

documents with many author. The reference manger (Zotero) result shows (Fig.9) that the number of publication for abstract reduced by 104 since abstract for those publications like webpage, report and even some book section are not made available. Figure 8. Then summarize the result finding and both knowledge mapping and knowledge structuring now likely to drawn from the analysis followed and interpretation and discussion then make clear each of the result in the subsequent section.

3.2.2 Analysis

According to(van Eck & Waltman, 2017) to cluster publication determining publication relatedness is the first task either based on citation relation or word relation while major aim of the survey to acquire and structuring knowledge over various approaches and methodologies of farm financing decision making. Analysis of publication mainly did using citation and co-occurrence as both of these analysis helps to learn about a filed or topics. Alternatively, since aim of this survey is to explore the extent and depth of research history, approach and

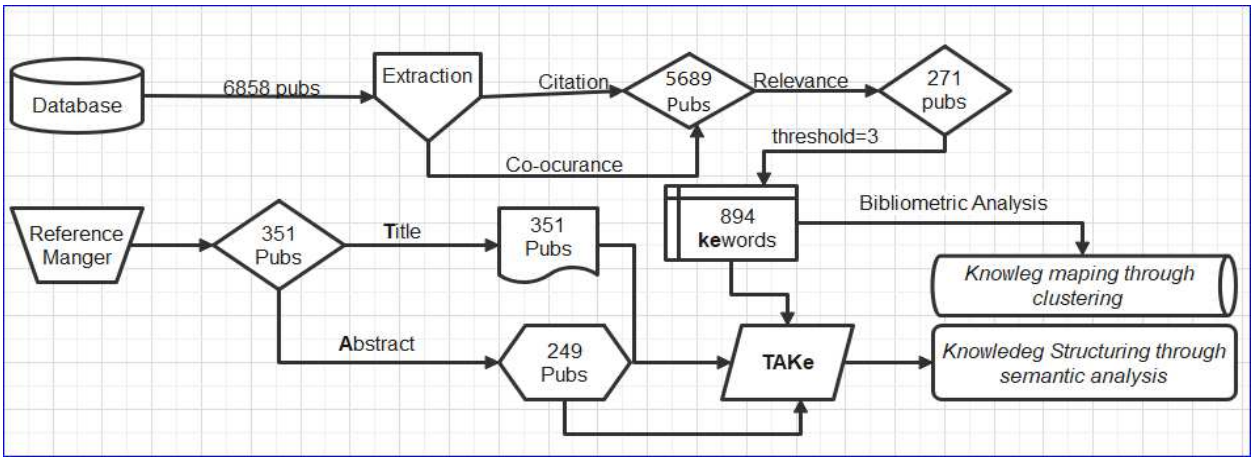


Figure 9.Result of Publications Retrieved for FFDM

mechanism regarding agriculture and finance particularly crop production as subsystem of farming that is polycentric inherently where both multidisciplinary and interdisciplinary are attractive, analysis method based on co-occurrence and co-citation more preferred. Since this, however, doesn't' mean others are not touched, co-authorship (authors vs organization) for example discussed based on a (2, 1) threshold inclusion and exclusion criterion followed to imply that an author and organization should have two documents with minimum of a single citation not to narrow down role of both organizations' and authors' in the problem interest a (as made available in Fig.3) though did not appear here due to space limitation.

3.2.3 Citation analysis

With threshold of unity as minimum citation for publication, of those retrieved 185 publication using generic query term, 113 documents identified and the maximum citation (407) achieved by Marks Glaser(2007). In fact, this figure is the second maximum citation as the top citation was scored by Dean T. Jamison (2013) but, since it doesn't create any link (link =0) to any others, it becomes seventh. Hence, what matters, however, is link and only six publication: Marks Glaser (2007, 407); Jing yuan Wan (2019,18) ,Lin Wiliam Cong(2021,2), Milan Lovric((2010,10),Huewin Lin(2011,29) and Taqadus Bashir(2013,14) found to make it(see Fig.7(a&b). The work of Dean T. Jamison (2013) entitled as Global Health2035: a world covering within a generation, "overconfidence and trading volume" is also the title of the document by Markus Glaser. These two paper, however, seems to have no direct implication to our problem interest FFDM and it is due to the generality of query term used.

On the other hand, if unit of analysis instead selected to be source of the literatures, from138 total source about 91 sources obtained, if minimum threshold for document of source and citation of source set to unit. Otherwise, list of source reduced to 9 if the threshold number of document increased to 2 or 3. With link as weight, only three sources found visible including Knowledge Discovery and Data Mining(KDDM), African Journal of Buses management (AJBM) and Erim Report Series Research in Management Erasmus Research Institute (ERSRM_ERI). Extending the analysis by score perspective that is normalized by citation, Knowledge discovery, and data mining source still rated as almost to 4.0, which is an impact factor for the journal actually according to VoSviewer manual generalization.

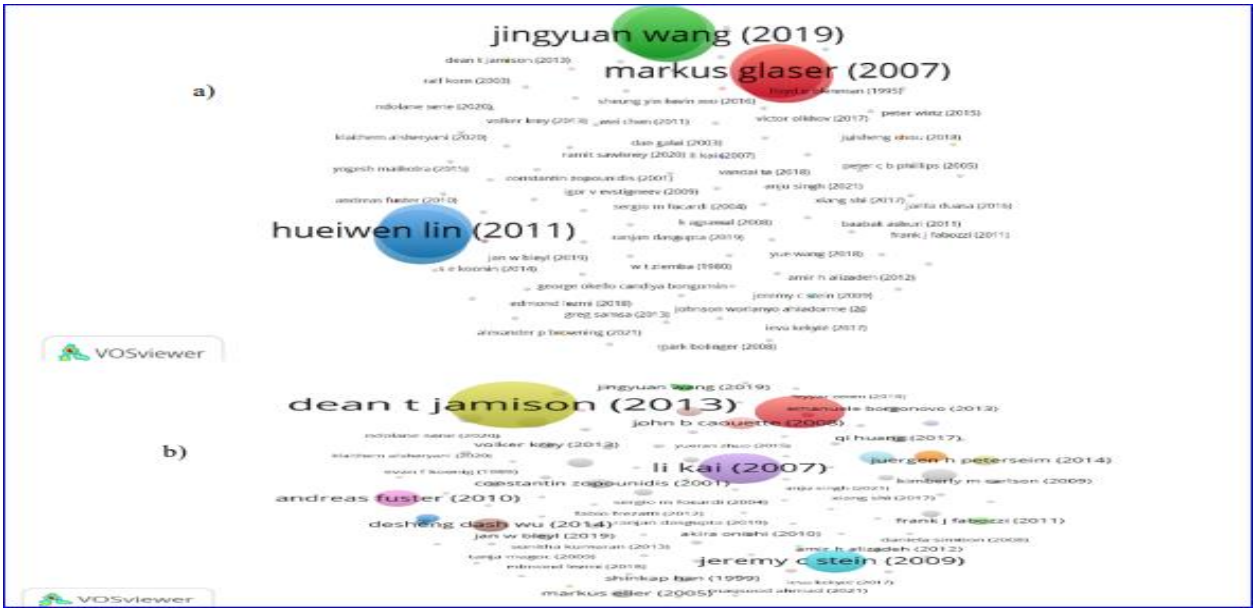


Figure 10.Documents using Generic term based on citation (a) and link (b)

Figure 11.Bibliometric coupling of source for MA with query term Modeling, Investment, and Finance

that is essential (B.S.Kade mani, 2011) which is closely related to co-citation but actually are about retrospective and forward looking perspective (Garfiled,2001) respectively. The bibliometric coupling analysis with source as analysis unit and minimum number of document for the source and minimum number of source citation respectively set to one and two to give 75 sources from 138 sources. Strength of bibliometric coupling to such 75 source with other source calculated at which Kybernets with two-document, five citation, and total link strength of 114 takes the top position and followed by ERSRM_ERI, Jeneva risk and insurance review (JRIR) and AJBM. The overly visualization (Fig.8b) of source with link as a metric of weights and average normalization of citation as a score value to indicate impact factor of journals (sources) as indicated by the color (blue, green and yellow) respectively for low ,medium and high impact of the source (see.Fig.9). Journal of Economic perspective at cluster seven for instance creates 17(though clusters are not made available) links and has total link length 37 with average normalization that is based on the association method is 4.96 (Yellow range) while Journal of finance has a link of 10 and total link strength 15 with scale value 7.77 and again in Yellow range (high impact). It can be continued to the entire source and two more important sources we wanted to mention here again are KDDM and AJBM that are both in the range of yellow and green with scale value 4.27 and 3.02 respectively. Since both the publication and source retrieved using such generic term does not grantee for drawing a conclusion, and it is why additional query terms then added with Boolean operator “OR” as Farm, OR Decision, OR Crop Yield, OR Risk, OR Credit for more specific analysis to the case of FFDM.

On the other hand, for Crossref database, using citation as a weight and normalized citation as score value, example of publication, those that are highly cited in their range: low (0.0-0.75), medium (0.76-1.5) and high (1.51>) identified as listed in table4.Using search term "income" that was not used in the previous analysis, 192 publications retrieved with minimum number of citation for a publication rather specified as five and 125 publications pass the threshold. Again these publications ranged from 1970 to 2010 and maximum average normalized citation as scale value obtained to be five (yellow range). These 125 publication, however, reduced to 70 due to disconnection between publications, and fall to 69 cluster, almost a publication as a cluster, hence exemplifying publication made here for those with in the yellow range only (4.5> of Fig. 8) and some of these publication are from sno-29-38.To demonstrate a comprehensive search using such terms in Crossref database, a Publish or Perish software used since it help to search publications other than specified. For example, with source specified AJAE on the space provided, the result gives other source including, agricultural economics, European Review of agricultural economics. One limitation of this technology is that it generates only 1000 publication and a total of 19032 citations with 271.89 and 19.03 citations per year and paper respectively. It reports author per publication as 2.11 with h and g index of 66.9 and 105 while the hi-norm obtained to be 47.

3.2.2 Co-occurrence

Using AJAE with minimum number of occurrence of a keyword at three, 894 keywords from 3077 analyzed to give economics with 946 occurrence and 6422 TLS take the top position and followed by business and agriculture. This result analysis in fact is seems indifferent with the analysis made using title only. It is also evidential that the schema of FFDM yet dominated by the field of economics, business, econometrics, microeconomics, and agricultural economics (see AJAE in Fig.12).

Moreover, as can be seen, financial economics, future contract, and finance it shown to have high correlation too. Capturing something essential to the schema of FFDM from this particular source of information is possible through the analysis of how those clusters classified. Cluster one with red ball (circle) representation in AJAE of Fig.7 and this cluster consists of 152 items that are mainly for modeling and solving methods, as it includes [none, linear, goal, dynamic and stochastic] programing, mathematical optimization etc.The second cluster, (green ball) with those 122 it emphasized to market and economic analysis including the financial analysis while third cluster, Blue ball (with 98 items) is especially to economic, environmental and ecosystem challenges of agricultural business. Similar to cluster three, cluster four, Yellow ball, also seems to deal about agricultural business, which however, emphasized on the economic growth particularly to food safety and subsidies and still highly dependent on the market economy and agribusiness. Moreover, agricultural, commercial, and financial policy, especially concerned to globalization and issues of hazard makes these and other clusters cluster linked. This can be justified further, if we take agricultural economics in cluster seven (orange colors) that also includes agronomy and agricultural science and agricultural engineering, is highly interlinked to other clusters as essential research interest of agricultural research. Cluster 5 with purple ball circle mainly are about resource managements, where those methods and approaches from both economics and econometric are demanding. It mainly composed with those types of issues focusing to economic model, economic evaluation and economic efficiencies including issues related to environment starting from essentiality of planning to policy matters including risk issues while cluster 6 (those with aqua ball) is more or less to deal about a process how decision-making is constructed and its constructs. It clearly constructed with terms that shows the importance of the As-IS approaches and business decision mapping along with data collection and methods of decision analysis including conceptual framework and decision support system. Issues related to theories and principles in the problem domain asserted in cluster 8 (Brown balls) including principle of information asymmetry, mainstream economics, managerial economics and positive economics as well as theory of firm. Economics as highest cited field of study in cluster 10 of the pink colors is more of about economic theories while subject of both macro and micro economic to deal both investment production decision are highly versatile to study. For instance production function, which is in cluster 8 for instance an essential one for microeconomics? Due to those theories and principles like production function and theory of firm, this cluster highly correlated to those most clusters. Those terms like production model, production risk, and simulation modeling in Cluster 9 for instance is highly correlated to many of other clusters including cluster 1, 10, and 12. Cluster, those with light green (#90EE90) 11 is more about risk and risk mitigation mechanism especially related to financial risks in the field of actuarial science, a discipline that assess financial risk in the insurance and finance filed using mathematical and statistical methods. Figure 10 provides snapshot of the query terms from each of the cluster formed by VoSviewer analysis to capture the linkage among those 13 clusters. For instance for the term "Modelling" three clusters found to consist it with 18 items. On the other hand, using another important term "Decision" about 17 terms identified only from cluster 1 and cluster 2.

This term, however, highly emphasized in cluster 2 as of those 17 it consists 16 items. Extending the filtering to the one that is emphasized in this investigation, “finance” only two clusters found with five items – cluster one with one item and cluster two with four items. Instead of the general term, Finance, an indicative term in this aspect instead is “Credit” and about nine items phrased in this case which, however, again obtained in two clusters- one and five. As observed in this filtering and from the general fact of the problem domain under investigation, an essential items that is central also to those research interest is “risk” which in this case it is to mean total risk found to take high shares in terms of items and clusters i.e., about thirty two items from five clusters. This conveys that about 3.6% of those keywords in one or another ways dealt about risk and its extension and create linkages of about 38.5 % of clusters. Similarly, co-occurrence analysis from the source agricultural system (AS) provides 668 keywords from those 2452 that are classified into ten clusters with: environmental (429,3464), agriculture(442,3462), agronomy(327,2936),and yield(251,2004) takes the first four position(occurrence, citation), as indicated by blue(C3); green(C2);C3; and yellow(C4) for each of cluster i of AS in Fig.7. Those red balls (circles) in this source denote C1 and include production, mathematics, and computer science to convey how to model agricultural economics and manage knowledge. It is visible among others knowledge management for instance strongly connected with relatively thick curved line, with agriculture and business in C2, to indicate high linkage among and between keywords in the network.

Generally, those items in C1 (red ball) are more about methods and tools of capturing agricultural problem while C2(green ball) deals about the subject matter and related theories in agriculture. C3 on the other hand composed of items that emphasize the science of agriculture, technologies as input while nature of agricultural outputs, and related activities, like for instance sawing, concentrated on C4 and have a strong linkage with C8 that deals about mechanisms for high input activities like cropping irrigation . Using JAE 187 keywords classified into eleven clusters, which is more similar to AJAE. Periodically those key terms concentrated ranging 1990 to 2010 with minimum of average minimum and maximum [0, 60] or [0.6, 1.4] average normalized citation. Agricultural economic (AE) based on proposed setup pro vides 1172 keywords of which 281 keywords meets the minimum requirement from those 400 publications by 863 authors from 272 organization. That 181-selected keyword then classified into nine clusters at which economics (see AS in Fig.6) takes the top in terms of occurrence of 293, with links of 270 and TLS of 1812. Predictably followed by agricultural economics with occurrence of 143 and TLS 926 and keywords like agriculture, production and yield takes next position with (occurrence, TLS) respectively (128,843),(85,584) and (55,330). A similar analysis for both coauthorship and co-occurrence for Agricultural Financial Review (AFR) that give a retrieval of 347 publication by 625 authors from 102 organizations performed and 232 keyword from those 917 clustered into ten clusters. The co-occurrence analysis still provides economics (volute ball in AFR of F.g6) appear first with (occurrence, Link, TLS) of (215, 219, and 1399) and agriculture (green ball)(111,175,730) and business (aqua ball) (114,195,748) for instance found in the first row of the analysis. These three keywords, in fact, are from different clusters as indicated by the color of each circle (ball) in Fig.6 at which read balls is cluster one, green ball cluster two while volute and aqua color indicates cluster five and six respectively.

A similar analysis for CrossRef database is possible but it is very cumbersome as the analysis is term by term. Using key term “modeling” as query term for example provides about 97 publications under the inclusion and exclusion criterion in term of publication ranging from 1950 until now. The retrieved publication reported, however, ranges from 1970 to 2020 while the co-aligned term “decision” retrieved ranged from 1968to2019 with72 publications. It is evident that issues regarding to decision modeling in farming activities were emphasized after the late 1967 whereas discussion and research related to income and related financial issues like credit backs to 1950s and are hot research interest still. Yes, it is true that yield is more related to modeling and is possibly affected by farmer’s decision hence crop modeling as demonstrated (Reynolds et al., 2018)is an important concern in farm decision. For CrossRef co-occurrence, fig.13 portrays network of 328 keywords at which only those non-grayed are connected that are 60 and categorized in 8 clusters. computer science - machine learning, statistics - machine learning and mathematics - optimization and control takes the first three top position in terms of occurrence and TLS with full (fraction) counting method respectively, 8,8 and 7, and; 15,16 and 11 (8,8,6). Almost about 90.8% of the keywords obtained to occur at a rate of unity but with different TLS value if full continuing method followed unlike that of fractional counting method that gives an equal value of TLS with occurrence.

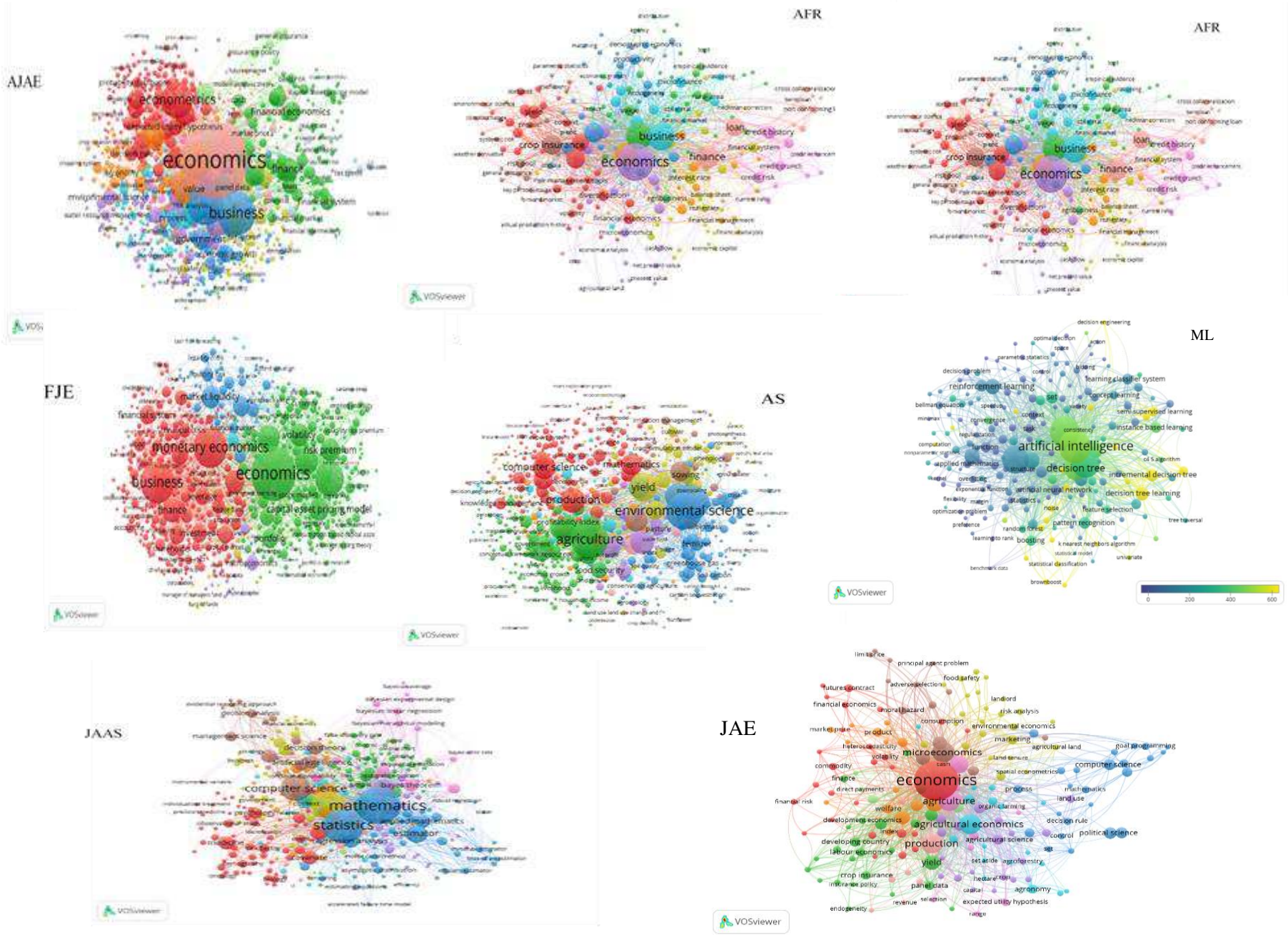


Figure 12. FFDM Schema for FFDM using Various Sources

Report from full counting methods to those low occurrences (unit), especially related to topics that are highly coherent to problem domain of the study as illustrated bellow rather implies they are cooccurred and signifies that how publications are inter related in the network of the knowledge domain specified. Consequently and as is demonstrated in the below list of some terms, though are in frequent each of the keywords instead are

Figure 13.Keyword from Reference managers to FFDM

not ignorant at full counting method, since the maximum TLS in this case from key word " simulation" is 19 and is not much far away to each of the examples given below next to fig.13

adaptive policymaking	1	5
agent-based modeling	1	5
agile analytics	1	6
agricultural advice	1	8
agricultural credit	1	2
agricultural entrepreneurship	1	8
agricultural finance	1	8
agricultural production management	1	5
agricultural production planning	1	3
agricultural productivity	1	3
Agriculture	1	3
agriculture and state	1	3
ecological model	1	4

Moreover, as captured from the overly visualization of fig.14, that is constructed from those 60 items and

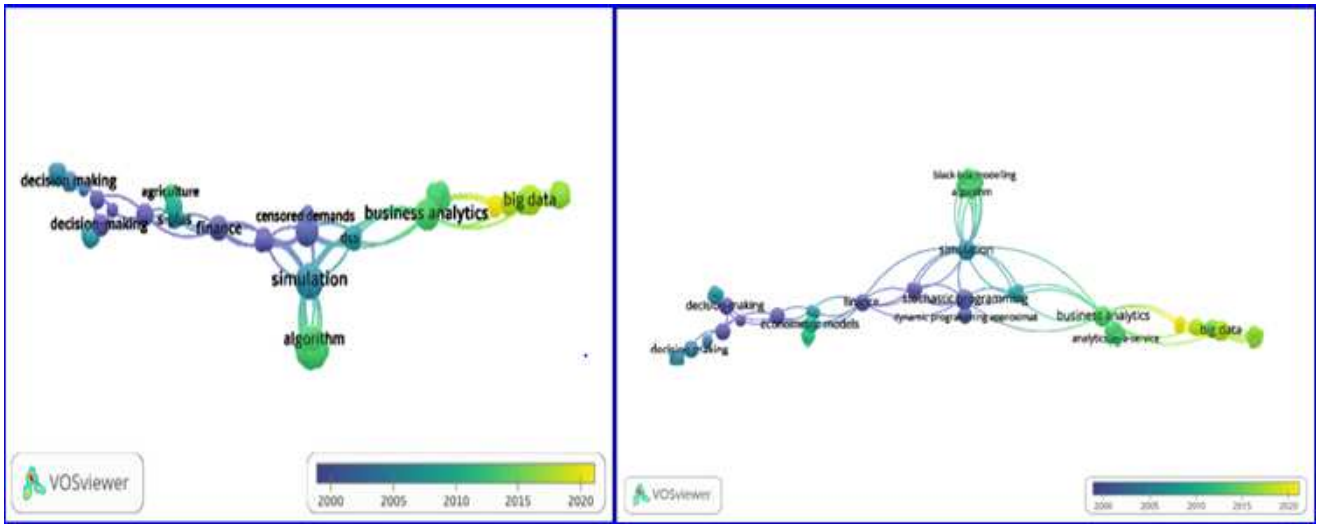


Figure 14.An overly visualization of keywords using bibliographic data type of reference manger data source (resolution for left=1, for right =0.5)

clustered to 8 clusters the using total link and publication year (average) as weight and score of visualization scale some kind of pattern can be visualized and understood in the evolution of the knowledge which furthers will be elaborated in topic modeling. Since each of the clusters scored based on average year of publication years, those publications indicated by blue color almost are about modeling of decision-makings particularly the normative approach, using methods and techniques from mathematics and statistics (econometrics, stochastic programing). While (light green) on the average around 2005 as indicated by the most occurred keyword, simulation, is a positive approach towards farm decision related to finance. Since normative and positive approach in farm decision ((See; G. L. Johnson, 1977 Fig.1) showed developments in methods and tools to each of the approaches and includes development of algorithms and introduction of data analytics. This development and evolution of tools and techniques in solving both normative and positive approaches now days, however, converge to the era of data driven approach as indicated by the yellow circles of Fig.14 and is hot research recent topic today. One advantage of VoseViewer in bibliometric analysis is it is flexibility related with number of clusters to be constructed through its resolution button of the analysis tab that sets default value for resolution to unity (left side of Fig.14). The higher the value of resolution at positive integer, the more the number of cluster formed to show shallowed and specificity and vice versa. Not only for better and for ease presentation, rather for precise generalization that would be re-evaluated in the coming section of topic modeling, those 8 clusters reduced to 6 by setting resolution to 0.5. In doing so, merging of keywords from those reaming two clusters reassigned to such six new clusters and the original structure now agitated, say for instance keyword“ decision making” in cluster one of the new grouping was in cluster three.

Using text data format from those 5338 terms obtained from those 353 publications, 659 terms only pass minimum occurrence threshold of three and using the relevance score default value of VoSviwere (0.6) 359 terms exposed for analysis. Despite of its less occurrence (3), term ‘*Continuous time financial models* (CTFM), takes the top withrespect relevance value of 3.11 while the least relevant term obtained in this analysis is ‘supply’ with relevance of 0.34. The highest occurred term as observed in the right most of Fg.20 is the term “Ethiopia’ with occurrence of 41 and relevance value of 0.477. Since minimum occurrence is 3.0 with average occurrence of terms equal to6.2, the occurrence value of term “Ethiopia’ signify many things that we would try to list some of them later, compared to that of less occurred but relatively high relevant term continuous time financial model (CTFM). These 395 terms then now classified into six clusters as depicted by various colors of network visualization in Fi.g 16 while statistics of clusters as summarized in table 4 and portrayed in Fig.17.Using most top terms in terms of occurrence and relevance for each of the clusters table 5 tries to illustrate to which research schema of the clusters are most persuaded. Terms those are general and common like ‘book’, ‘end’, ‘section’, ‘task’, in cluster for instance ignored for consideration. One important term that might not have any semantic information is term “ackoff” in cluster one that instead is an author who contributes a lot to the theory of system modeling than reductionist approaches. Therefore, though it does not help to extract semantic information, its repeated occurrence in the cluster along with other terms like theory and prescriptive analytics signifies something that important for the central scheme of the of cluster. With respect to occurrence of term or keyword like ‘prescriptive analytics (22)’ and ‘theory(29)’ in cluster one respectively implies the importance of bridging prediction and optimization approach and how to follow approaches and methods in the farm decision making. On the other hand, as relevancy of term used for evaluation, CTFM and learned policy in cluster one again suggests how dynamic is the financial modeling and it is essential to having flexible environment of financial policy respectively. With similar fashion to reaming terms and clusters and using implication from each of terms in the cluster, drawing the generalized implication is given for each of the clusters as listed in table 4.

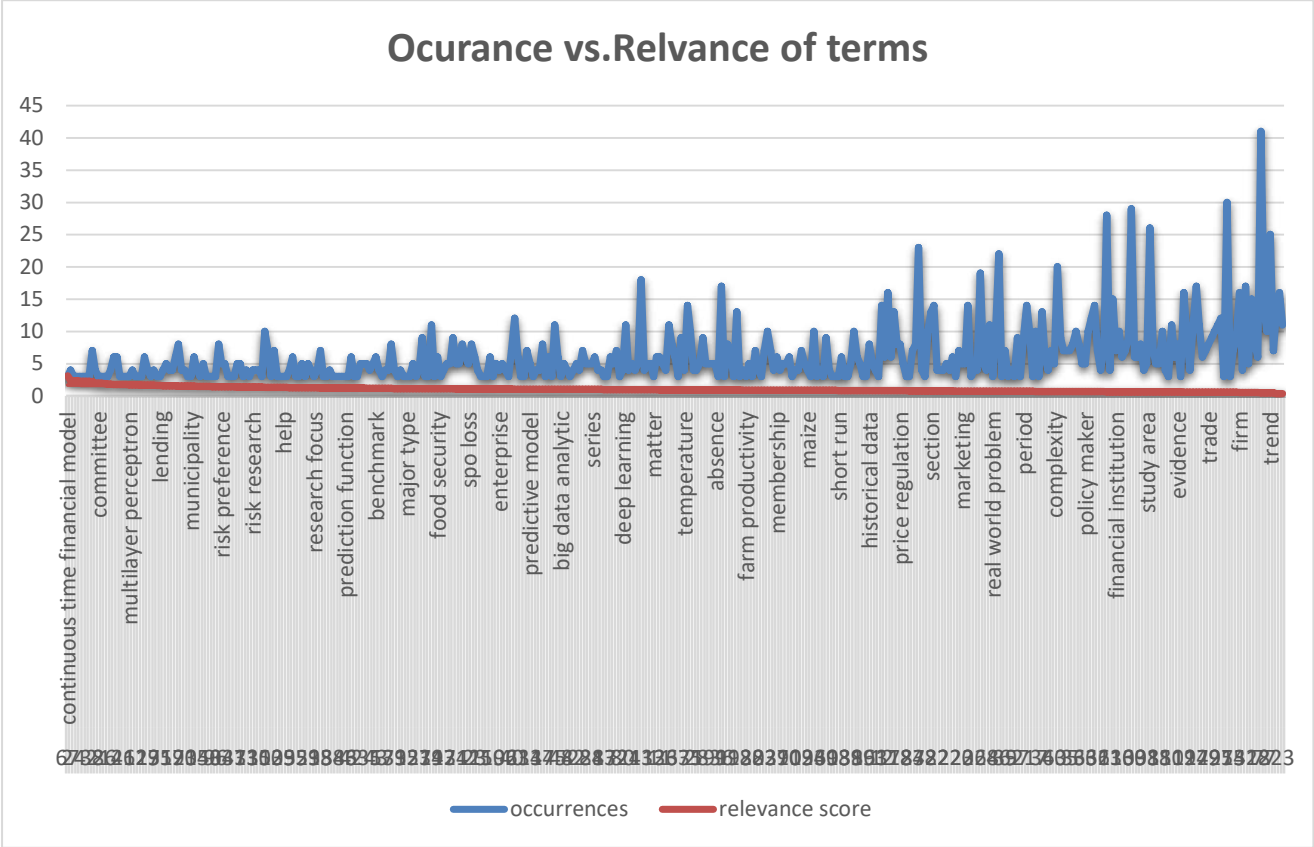


Figure 15. Termed occurrence and frequency for text analysis of FFDM

Table 4.Statistic of Clusters

	#Items	max .Ocu	Min.Ocu	max.rel	min.rel	ave.ocu	av.rel	%
C1	143	30	3	3.114	0.451	6.373	0.967	36.20%
C2	133	41	3	3	0.339	6.409	0.920	33.67%
C3	48	30	3	2.373	0.550	5.511	1.156	12.15%
C4	28	25	3	2.268	0.464	5.857	1.116	7.09%
C5	26	20	3	1.749	0.564	6.923	1.023	6.58%
C6	17	12	3	2.258	0.644	5.176	1.241	4.30%
	395	158	18	14.762	3.0108	36.24966	6.42317	

Note: Max (Min).Ocu=maximum (Minimum) occurrence, Max (Min).rel=maximum (minimum) relationship, ave(ocu).rel=average(occurrence) relationship.

C1: Modeling approach and procedures, starting from descriptive modeling to the most recent prescriptive approach
C2: Farm modeling and implication of financial leveraging
C3: Farm decision using recent approach in the domain of artificial intelligence (AI)
C4: Theory of financing and its attributes in the theory of firm
C5: The need of exploratory modeling in policy analysis
C6: The spatial evidence how farming activity is crucial to communities' livelihood.

3.3 Topic Modeling

This section is not for comparative analysis of bibliometric analysis made so far instead to complement and strengthen it. This is because of that topic modeling is more efficient than that of bibliometric, at which (co) word mapping is don through clustering, to evaluate exogenous variables and even the endogenous variable from the semantic nature they composed of. Based on the "Moto", TAKE (see also Fig.1&9), publication title and abstract analyzed taken from the reference manager while the last one from AJAE. With this premises using 894 keywords from AJAE in dimensions database, thirty topics generated by LDA (see below

some of them) and word cloud for those keywords given in fig.18. As, word cloud implies those publication

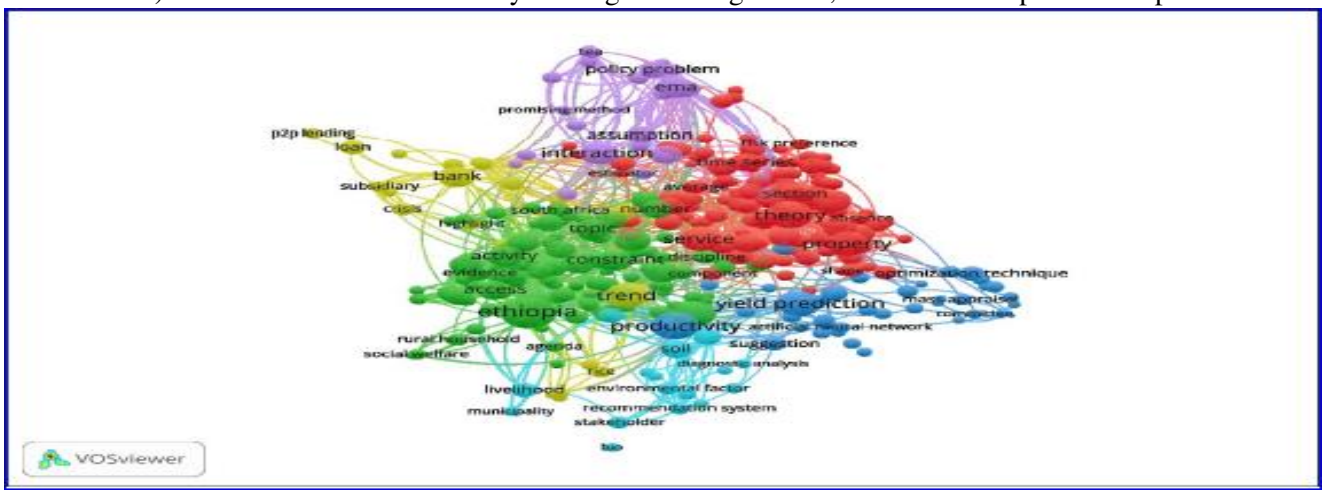
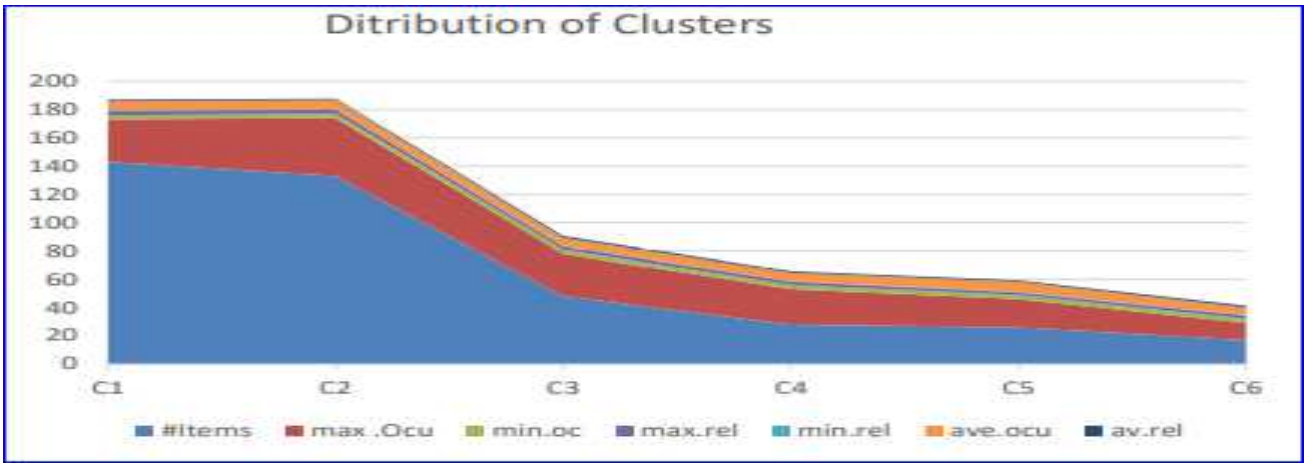


Figure 16. Network Visualization of terms for FDMA

can be characterized by topical contents that are very coherent with words like economic, market, risk



decision, model, agricultural etc and the importance of topic modeling lays on communicating of most salient

Figure 17.Distribution of cluster based on occurrence and relevance

points or themes between publication.

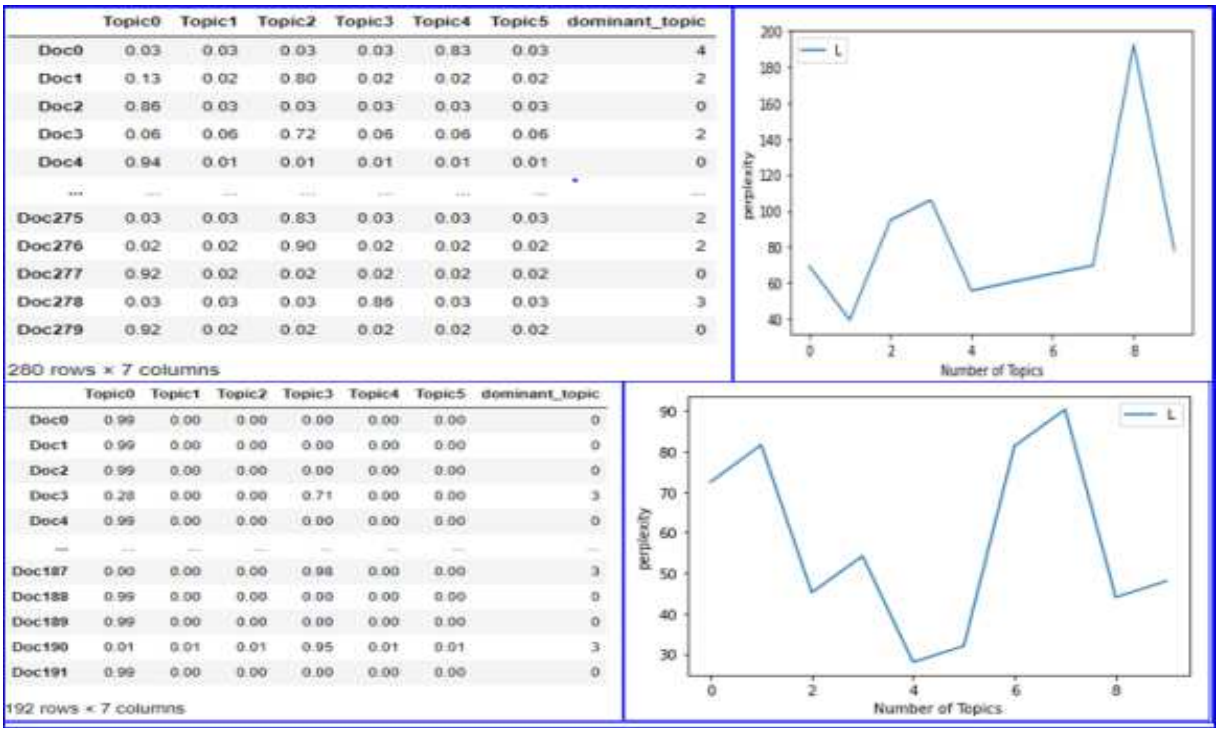
Topic 1 ['cost' 'supply' 'decision' 'process' 'demand' 'marginal' 'choice' 'chain' 'consumer']

Topic 16 ['rate' 'capitalization' 'mathematical' 'interest' 'corporation' 'context' 'microfinance' 'identification' 'loan']

Topic 30 ['time' 'factor' 'service' 'cost' 'outcome' 'phenomenon' 'hectare' 'sensitivity' 'wage']

Figure 18.Word cloud of keywords

Since, keywords are dependent to the topic of interest (Wang et al. 20163), mimicking past topic of interest using keyword needs time treatment and besides the frequency of keywords, length of words as demonstrated by Term-frequency invers document frequency (TFIDF).
Evaluation of topic model is based on some metrics as has been demonstrated in 'tmtoolkit' and this can be used to find a good hyper-parameter set for a given dataset, e.g. a good combination of the number of topics and concentration parameters (alpha and beta) defined in section 2.4. Since keywords here are simply list of word and fail to define texts making topic evaluation is nonsense while for title and abstract it makes sense. Given k number of topic, the prior concentration parameter over the document-specific topic distributions, α, is then equal to 1/k and the document-topic density in this case is 0.033 and implies that documents (here keywords) are with fewer topics as would expected and with no surprising the topic-word density(beta/eta) also small.
Topic modeling using LDA for Scikit implementation therefore gives six topics for those 280(192) training dataset publications as demonstrated in fig.19&20). LAD performance using Scikit learning determined by calculating perplexity or predictive likelihood for α and β equal to 0.01 that gives 65.109 (32.0) if topics are six(left side of fig.20) otherwise perplexity is 192(90.32) if eight topics propose (see right side of fig.20). Though it helps to determine optimal number of topic by measuring in what way model is able to predict, perplexity is less correlated with human opinion and for a model to be satisfactory, predictive likelihood should be low in contrast to log-likelihood score, which are essential to compare different models at large value. The'pprint' package gives model parameters for values to log-likelihood to be -15628.26 (-151455.78) and perplexity of 3399.04(6767.768) with learning decay rate that control the learning rate as 0.7 for those number of components/topics (six). The learning method was online with learning offset ((down weigh early iteration) of 50, none document-topic, and topic-word prior with total sample size 1000000 and 0 verbose in both case. The result obtained from perplexity plot implies that only the lower limit(for title) gives optimal topics, though is not yet smooth, hence, coherent score instead is preferable while since Scikit-learn package implementation of LDA does not provide this method to measure coherence score(Tijare & Rani, 2020), genism package from NLP deployed. Fig.19 therefore displays distribution of topics to words obtained from those 353(241) publication at which, for instance, terms like "research" + *"risk" + *"agricultural all contribute equally to second topic (topic 1, since python starts counting from zero) with weight of 0.027 to each in the title case.



	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic
Doc0	0.03	0.03	0.03	0.03	0.83	0.03	4
Doc1	0.13	0.02	0.80	0.02	0.02	0.02	2
Doc2	0.86	0.03	0.03	0.03	0.03	0.03	0
Doc3	0.06	0.06	0.72	0.06	0.06	0.06	2
Doc4	0.94	0.01	0.01	0.01	0.01	0.01	0
...	...	...	...	...	...	...	...
Doc275	0.03	0.03	0.83	0.03	0.03	0.03	2
Doc276	0.02	0.02	0.90	0.02	0.02	0.02	2
Doc277	0.92	0.02	0.02	0.02	0.02	0.02	0
Doc278	0.03	0.03	0.03	0.86	0.03	0.03	3
Doc279	0.92	0.02	0.02	0.02	0.02	0.02	0

280 rows × 7 columns

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	dominant_topic
Doc0	0.99	0.00	0.00	0.00	0.00	0.00	0
Doc1	0.99	0.00	0.00	0.00	0.00	0.00	0
Doc2	0.99	0.00	0.00	0.00	0.00	0.00	0
Doc3	0.28	0.00	0.00	0.71	0.00	0.00	3
Doc4	0.99	0.00	0.00	0.00	0.00	0.00	0
...	...	...	...	...	...	...	...
Doc187	0.00	0.00	0.00	0.98	0.00	0.00	3
Doc188	0.99	0.00	0.00	0.00	0.00	0.00	0
Doc189	0.99	0.00	0.00	0.00	0.00	0.00	0
Doc190	0.01	0.01	0.01	0.95	0.01	0.01	3
Doc191	0.99	0.00	0.00	0.00	0.00	0.00	0

192 rows × 7 columns

Figure 19.LDA-Topics generated using Scikit Learning for title (above and abstract (underneath)

[3] Keyword-based Topic Modeling and Keyword Selection

These terms/words for topic modeling based on abstract however contribute differently to each topic, like for instance term risk weights about 0.022 for topic 2, research weigh 0.007 and 0.006 in topic 0 and 4 while term

```
[(0,
  '0.023*"model" + 0.014*"datum" + 0.012*"use" + 0.008*"decision" + '
  '0.007*"research" + 0.007*"system" + 0.006*"approach" + 0.006*"farmer" + '
  '0.006*"agricultural" + 0.006*"value"'),
 (1,
  '0.020*"crop" + 0.013*"system" + 0.011*"agricultural" + 0.011*"use" + '
  '0.011*"yield" + 0.009*"model" + 0.008*"predict" + 0.007*"base" + '
  '0.007*"agriculture" + 0.005*"soil"'),
 (2,
  '0.022*"risk" + 0.013*"farmer" + 0.010*"datum" + 0.010*"farm" + '
  '0.009*"model" + 0.008*"method" + 0.008*"decision" + 0.008*"management" + '
  '0.007*"financial" + 0.007*"system"'),
 (3,
  '0.018*"model" + 0.011*"use" + 0.010*"crop" + 0.008*"optimization" + '
  '0.007*"datum" + 0.007*"decision" + 0.005*"well" + 0.005*"paper" + '
  '0.005*"study" + 0.005*"analytic"'),
 (4,
  '0.012*"price" + 0.012*"farmer" + 0.010*"land" + 0.009*"area" + '
  '0.008*"production" + 0.007*"crop" + 0.006*"increase" + 0.006*"spatial" + '
  '0.006*"research" + 0.006*"datum"'),
 (5,
  '0.016*"problem" + 0.012*"model" + 0.011*"optimization" + 0.008*"credit" + '
  '0.007*"base" + 0.007*"strategy" + 0.006*"farmer" + 0.006*"solve" + '
  '0.005*"use" + 0.005*"technique"')]
[(0,
  '0.028*"input" + 0.022*"agriculture" + 0.022*"case" + 0.021*"approach" + '
  '0.020*"base" + 0.018*"indicator" + 0.016*"agricultural" + '
  '0.014*"uncertainty" + 0.014*"food" + 0.013*"econometric"'),
 (1,
  '0.027*"research" + 0.027*"risk" + 0.027*"agricultural" + 0.020*"model" + '
  '0.019*"datum" + 0.016*"approach" + 0.016*"opportunity" + '
  '0.015*"smallholder" + 0.015*"use" + 0.014*"new"'),
 (2,
  '0.045*"crop" + 0.033*"yield" + 0.029*"risk" + 0.027*"use" + '
  '0.026*"prediction" + 0.021*"machine" + 0.021*"learning" + 0.017*"farmer" + '
  '0.016*"technique" + 0.015*"agriculture"'),
 (3,
  '0.059*"model" + 0.030*"decision" + 0.026*"agricultural" + 0.024*"make" + '
  '0.021*"new" + 0.021*"system" + 0.018*"farm" + 0.018*"time" + '
  '0.017*"economic" + 0.017*"regime"'),
 (4,
  '0.064*"machine" + 0.036*"learn" + 0.031*"optimization" + 0.030*"learning" + '
  '0.025*"algorithm" + 0.023*"application" + 0.019*"agriculture" + '
  '0.017*"method" + 0.016*"yield" + 0.013*"scale"'),
 (5,
  '0.040*"risk" + 0.032*"analysis" + 0.029*"farm" + 0.022*"policy" + '
  '0.019*"model" + 0.016*"management" + 0.015*"optimization" + 0.015*"apply" + '
  '0.015*"stochastic" + 0.014*"level"')]
```

Figure 20.Topic distribution for publication using genism (upper for "Abstract", underneath for "Title"

agricultural contributes to topic 0 and topic 2 with weight of 0.006 and 0.0011 respectively. This is one advantage of topic modeling in obtained single term with different topics, but with different contribution, compared to clustering in bibliometric analysis done so far. Furthermore, due to having, different distribution for topics in a document, obtaining topics that are dominant in topic modeling is an easy task at which topic 2 is most frequent and dominant topic for title-based topic modeling as observed in fig 26. In the same token, dominant topics for abstract based modeling reports that topic 0 and topic 3 exclusively dominates to all documents. For validation purpose, coherence score now easily determined as 0.4221(0.28) by importing Coherence_Model from gensim.models. An essential thing in this analysis is there is no any outlier for publication, since no negatively indexed topic, as experienced in '*BerTopic'* package, whereas these six topics can be coined to some scheme of research to the problem surveyed. For instance, topic-0 (in the case title based modeling, TM_T) tries to signify how to model agricultural problems particularly to food security at which various inputs highly affect modeling process (input takes relatively higher weights, 0.028).It farther conveys loosely importance of econometrics (0.013) modeling methods to handle uncertainties regarding to the problem indicators (0.018) whatever the approach (0.0021) is positivistic or normative. On the other hand, topic 1, can be generalized as how to agricultural research should be conducted particularly at the farm level than sectorial level where risk whether at systematic or unsystematic and or at perspective of finance and idiosyncratic risk for crop yield due to output uncertainty to both case. With similar fashion to remaining topics and topics from abstract (TM_A column),a rough generalize made to those six topics as table 5 which, however, further solidified by pining terms that are more prevalent.

Table 5.Generalization of topic to their central scheme

Topic	TM_T	TM_A
0	How To Model Agricultural Problems Particularly To Food Security	Systemic approach and modeling in agricultural decision
1	How to Conduct Agricultural Research Predominantly at Farm Level and approaches	Agricultural System Modeling and Crop Yield Prediction(ASMCP)
2	Role of Machine Learning in Agriculture to Predict Crop Yield	Farm Risk Modeling and Farmer Financial Decision(FRMFFD)
3	System Modeling in Agricultural Decision Making and Challenges of Farm Economic scenarios	Crop Production Optimization Modeling and Analytical Decision(CPOAD)
4	Application Of Machine Learning And Optimiz	Farmer Crop Production Acreage Allocation and Spatial Pr

	ation Methods In Agricultural Economic Scale Improvement	ices (FCPAASP)
5	Farm Management And Risk Optimization Modeling For Policy Analysis	Farm Optimization Model Under Credit Constraint (FOMUCC).

3.3.1 Model Improvement and Justification

Those topics obtained from genism package are based on the default value of LDA parameters (α = 0.1, β = 0.01)that are actually either symmetric or asymmetric distribution at which for the first case a higher alpha (beta) documents (topics) are made up of more topics (words) and vice versa. In the case of asymmetric distribution, higher alpha (beta) results in a more specific topic (word) distribution per document (topic). In general, higher alpha values mean documents contain more similar topic contents. The same is true for beta, but with topics and words: generally, a high beta will result in topics with more similar word contents and a general recommendation has been forwarded as asymmetric alpha is helpful, than asymmetric beta. In the case of genism, the default value for alpha is 'symmetric. This means that the value for alpha is uniform for each topic and each topic is evenly distributed throughout a document unlike asymmetric distribution (as measured by skewness) where certain topics would be favored over others. The formula which genism uses to calculate the symmetric value for alpha is to divide 1.0 by the number of topics in the model. For this and as improvement of genism based LDA implementation, improving the LDA topic modeling by defining supporting function as $defcompute_coherence_values(corpus, dictionary, k, a, b)$for k-number of topic, and hyper-parameter $a = alpha\ and\ b=beta$. This supporting function then runs by setting the minimum and maximum range of topic (2,11). Table 6 then displays an optimal number of topics with respect to asymmetric and symmetric hyper-parameter values ($\alpha = asymetric$, $\beta = symetirc$) with coherence score of 0.4315 and number of topic therefore now become five (see also fig.21). As can be seen from the snipsheet, one essential contribution of asymmetric alpha in contrary to LDA that assume common Dirichlet prior distribution is to identify dominate topics along their percentage contribution in the document. While distribution of document word counts seems uniform in fig.22, distribution of document word counts by dominant topic instead is skewed as portrayed in fig.23. This signifies how topic distribution rather distributed disproportionately in publication, and is expected actually, since, there always exist no indifference to central scheme of publications certainly. Dominant topic in a document implies the central theme of the publication that is latent in fact while publication in this analysis, however, are truth sets, since number of relevant themes can be known a priori and hence implementation of LDA therefore confirmed valid in this analysis. This is because literatures in agricultural decision-making have been relatively structured particularly related to approaches and purpose of modeling. For instance, regarding to purpose of modeling in agriculture, the two most approaches are (as confirmed by fig.26) Normative and Positive approaches (G. L. Johnson, 1977; Petit, 1976 For example) that can, however, be decomposed into various models. Whereas to account nature of problem domain leads classification of agricultural modeling either as deterministic or stochastic while incorporating adaptive behavior of farmers as agents mostly recommended through using theory of utility function.

Table 6.Coherence score and hyper-parameters for topics using genism-LDA using title and abstract

Topic	**Coherence Value**		**alpha value**		**beta value**	
0	TM_T	TM_A	TM_T	TM_A	TM_T	TM_A
	0.3956	0.2529	0.01		0.01	
1	0.4031	0.2441	0.01		Symmetric	
2	0.377	0.264	Symmetric		0.01	
3	0.3968	0.2685	Symmetric		symmetric	
4	0.4179	0.2657	Asymmetric		0.01	
5	**0.4315**	0.2571	**asymmetric**		**symmetric**	
6	0.3946	0.2499	0.01		0.01	
7	0.3996	0.2454	0.01		Symmetric	
8	0.4026	0.2548	Symmetric		0.01	
9	0.4076	0.2586	Symmetric		Symmetric	
10	0.411	0.2448	Asymmetric		0.01	

This is, however, should not be considered as sufficient classification of literatures in agricultural decision-making at which most of the time studies completely focused on investigating either on the agent's decision making preference and or production function. This can be justified using keywords of the various topics

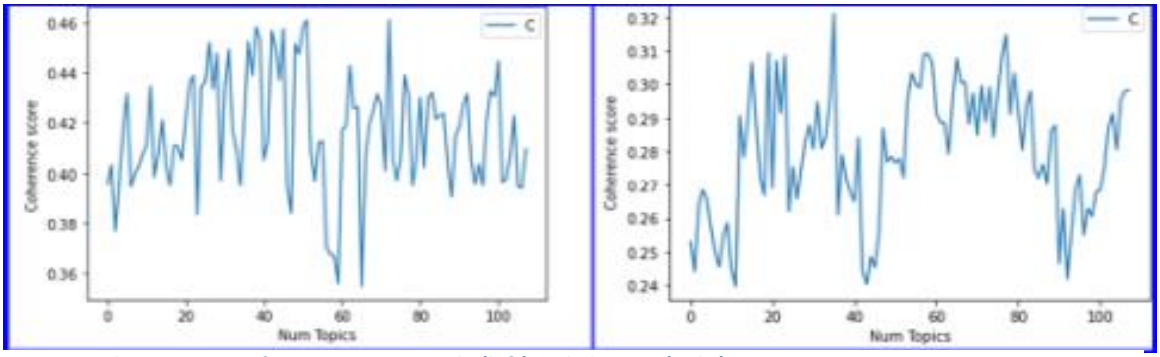


Figure 22.Coherence Score for # topics using title (left) and abstract (right)

extracted, as for instance the term 'theory' in topic 0 clearly signifies the importance of various theories in the problem domain that includes among other, theory of Firm , production theory and consumption function all which designed for the purpose of making viable decision making. This particularly expected in agricultural decision making that best characterized by risk as demonstrated in Topic 1, which is highly weighted in the topic. Furthermore, the difference in term's (keyword's) weight clearly convey themes of the publication say for instance, though keyword "model" appears in both topic 0 and 1, it receives different weights due to the orientation of underlying to pics. Explicitly, in the first case it is theory that more matters than models, though it is an immediate issue to be considered, for general case while it comes next to farm when risk is specified to agricultural decision. Similarly, analysis for other terms in and reaming topics can be mad while the important term issue especially to this investigation is the term 'credit' in topic1 that is composed from term starting risk to stochastic, highest to lowest weightage, that implies, when compared to other terms, something essential to examine critically. This is because that most of studies in agricultural decision making are more ignorant for direct and or explicit consideration of financial problems despite of its sever impact especially to households. This is justified by fig.25 that demonstrated most discussed topics in the document or publication retrieved and no terms that indicate financial decision like credit or debt appear.The 2D plot of topic using pyLD Avis in fig.24 is based on the dimensionality reduction methodology, principal component analysis

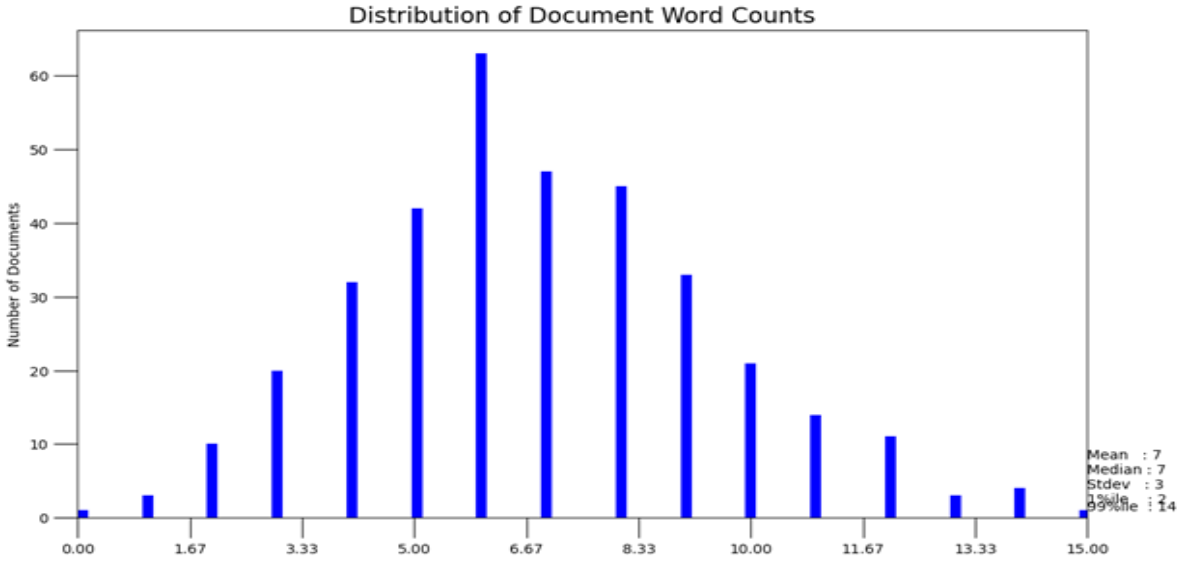


Figure 21.Distribution of Document Word Counts

(PCA) and there is only one overlap of topics (Topic 2& Topic4) whereas topic one found as more prevalent one as it makes up biggest portion of topic being talked about amongst documents (38.8%, upper part of fig.24). Similarly it is topic one again (but different topic here) that is more prevalent (37.3%, lower part of fig.24)

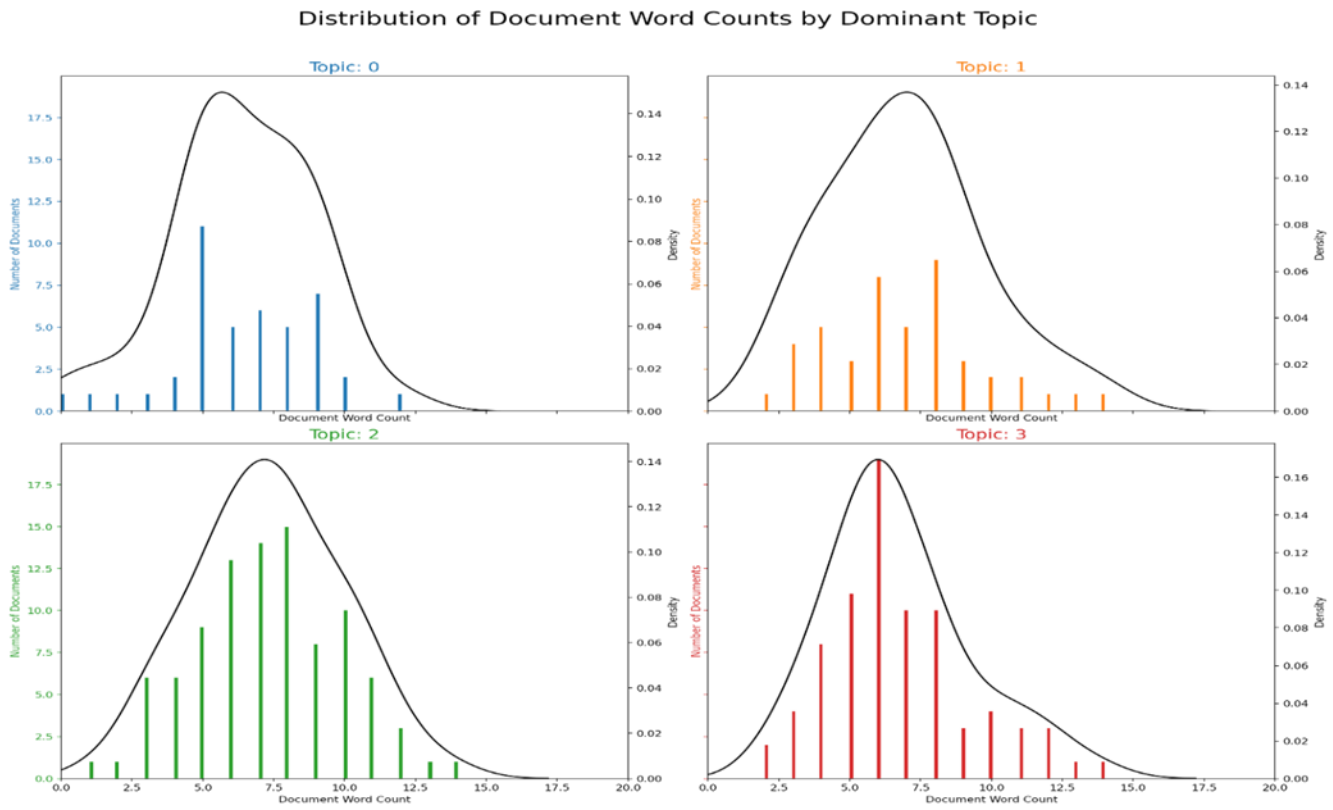


Figure 23.Document Word Counts by Dominant Topic.

4. **Discussion and interpretation of clusters as knowledge mapping and structuring**

As reported by (ROSSI et al., 2012, p. 458) the integrated farming (IF) as the whole farming approach and integrated crop management (ICM) or integrated production (IP) as holistic approach rooted from in Integrated Pest Management (IPM). Without losing generality, this can be generalized by the taxonomy proposed by(Reynolds et al., 2018a) using building blocks of $G \times E \times M \times S$, for $Geno\ type = G, Environment, E, Managment, M,$and $Socioeconomic, S$ paradigm of international crop production. Conceptual model G×E×M is based on biophysical variables that directly determine crop growth, and their interaction whereas since these biophysical variables are under highly influence of socioeconomic factors (S) like supply and demand of input/outputs, finance and credit, agricultural policies and the adaptive practice. Hence, G $\times E \times M \times S$ *can be a special case of bio-economic model.* Thematically, research activity in agriculture, however, broadly generalized (Hoffmann & Kleynhans, 2011) as either technological improvement or informational. Distribution of clusters enhancement with the first is mainly through agronomy, soil science, pathology, and entomology while agricultural economics and farm management contributes to the latter and this fits with the paradigm of the three way interaction $E \times M \times S$. This approach, almost but not completely, similar to the two main strands of David Gibbon regarding farm system research (FSR), one is about the fundamental to the field of FSR while the second and more emphasized was the methodological element seen from LERN group and agricultural knowledge and information system (AKIS) group(Gibbon, 2012). Another perspective of farm modeling from the perspective of cluster (C1) is underlying of interaction and relationship that leads scope of farming to either farm level or territorial or sector level(P. G. Strauss et al., 2008). According to Strauss et al, the latter is more facilitated by econometric modelling to assess market price and policy and hence are instruments of strategic decision, despite statement given so far, this an optimization methods and the normative approach it is. From purpose of modeling to C1, normative and positive approach has been frequently cited in agricultural decision literatures while Csaki(1976) mentioned :mathematical programming, mathematical statistics, production functions, input-output analysis and network analysis to it . Richardson on his book, Simulation for applied risk management,on the other hand describe positive approaches as a non-optimizing approach to farm simulation models to answer the positive question of what is the likely outcome than the normative answer what ought to be (Richardson, 2008, p.2). With Regard to FSR,(Feola et al., 2012) proposed three, stage of generation while Richard Bawden(2012) and the fourth one as : (i) the nature of reality (ontological beliefs);(ii) the nature of)

knowing and knowledge (epistemological beliefs);(iii) the nature of human inquiry (methodological

Figure 25.Visualizing Term score for topics

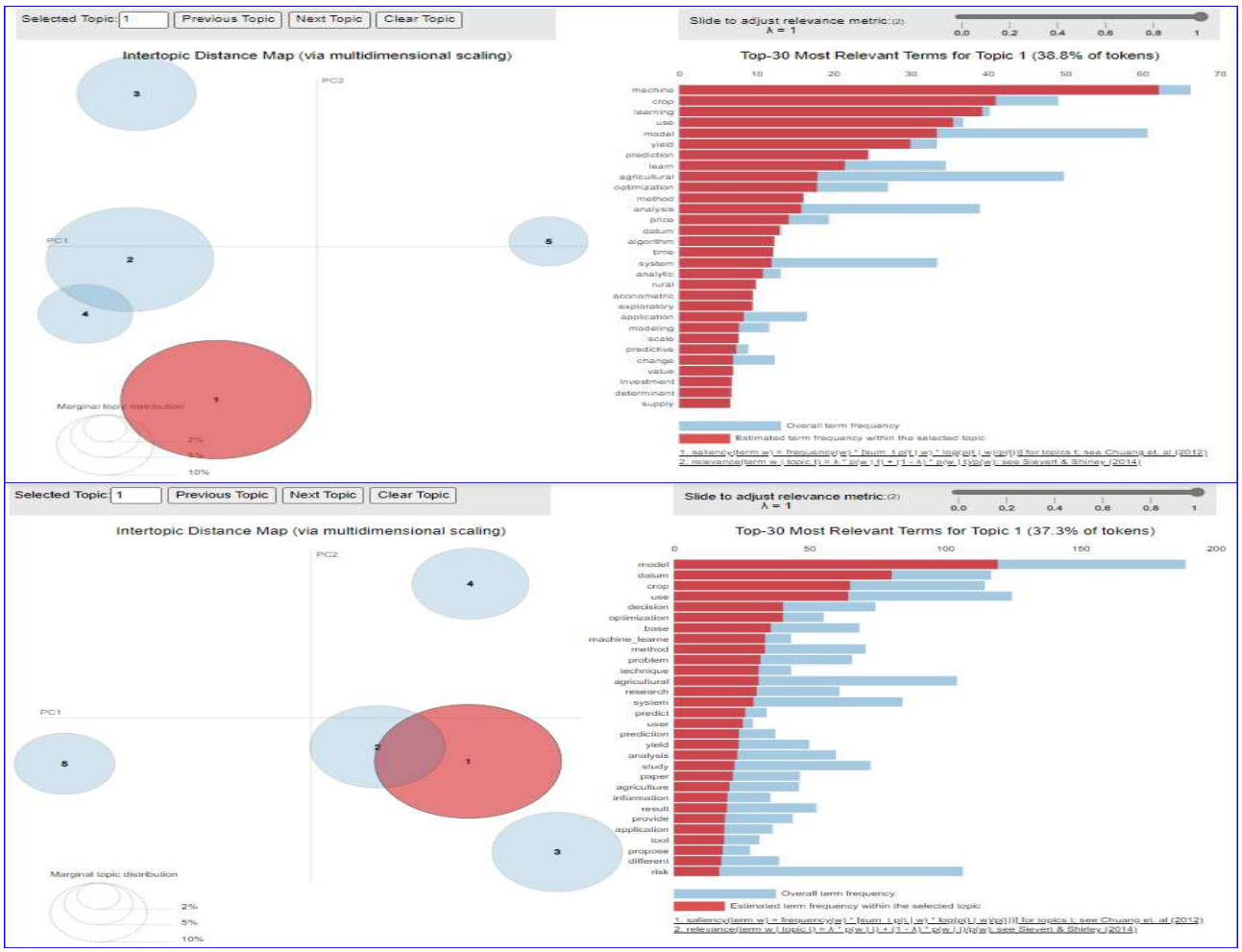


Figure 24.PyLDAvis visualization of topics for Titles (above) and abstract (underneath)

assumptions) and (iv) the nature of human natureassumption of preference. Particular to adaptive decision in

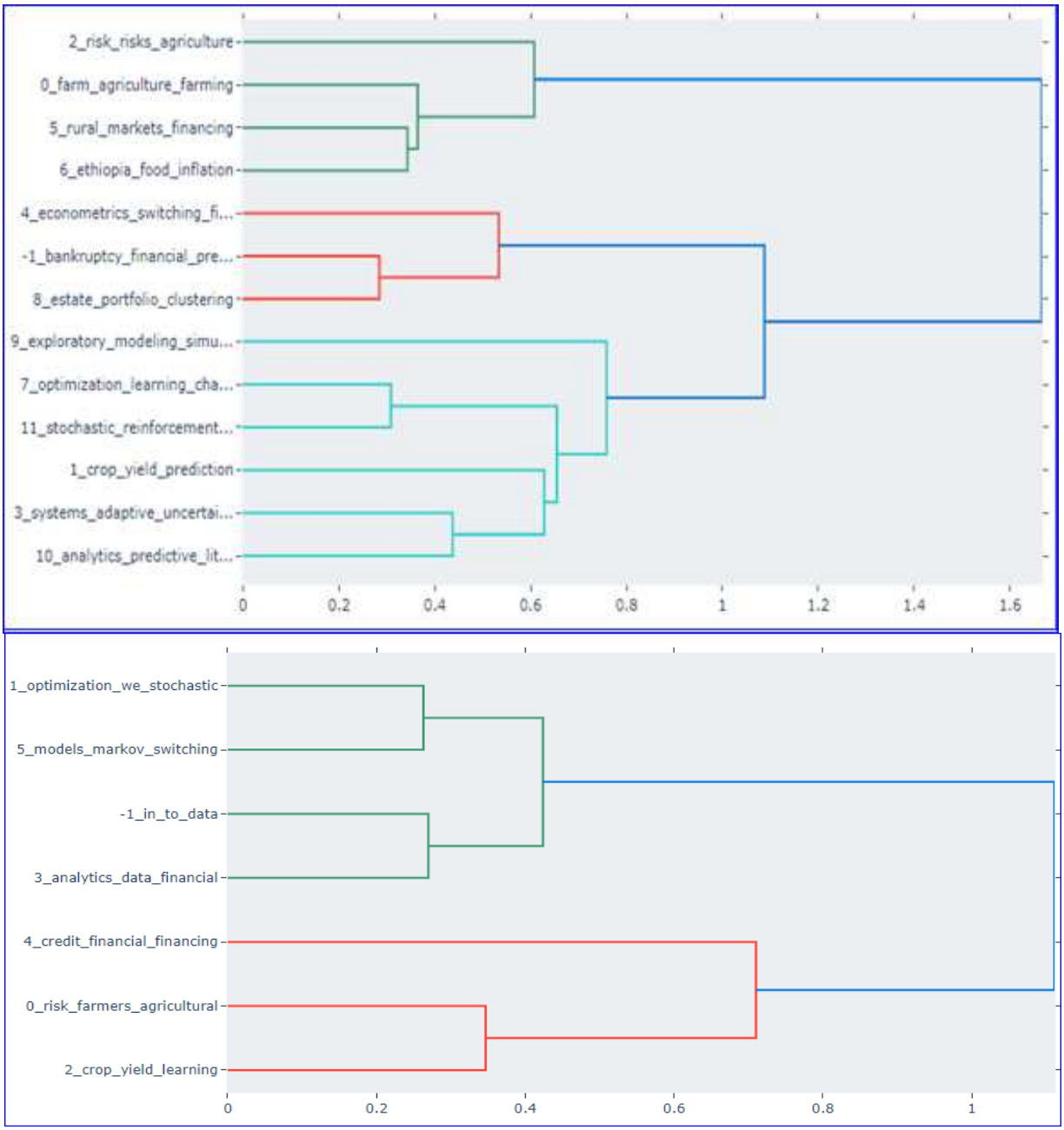


Figure 26. Hierarchical Clustering

farming activity, bio-economic and bio-decisional approaches have been device (See ; Robert et al., 2016) while, as noticed by (PG Strauss et al., 2008) unless utilizing models that potentially capture salient feature of the uncertain farming environment, making efficient decision and recommending for viable policy direction is imposible. According to Robert et al., both tactical and strategic decision should be adaptively addressed to take into account the dynamic nature of the problem and as described in the introductory section to the best of policy direction both prediction and decision modeling are worthwhile. Moreover, the operational decision is more complex in agricultural decision making to reach on common agreement due to variation in managerial skill and cognitive knowledge to operational decision making (Martin-Clouaire, 2017). It is openness along it being polycentric, when seen from financial relationship and institutional perspective, agricultural the nested hierarchy of governance affects the operational decision. For example, rules defining the amount and timing of fertilizer application on a field and the timing of debt return and credit receiving even contained in and affected by the rules at a higher collective-choice level of decision-making of course higher collective choice

rules are also contained in and affected by higher level of decision-making, the constitutional choice level. This operational decision challenge in modeling approach of C1 along with the essence of C2,shows how each of the cluster linked and it is risk as a triggering factor whatever the form of risk it be, production, risk, market risk, or financial risk for instance, has challenging farm level decision making. Since risks are due to those uncertain events in farming activity, the usual understanding of risk modeling in this case has dominantly been practiced by attaching probabilities to those uncertainties. Besides the pitfall of attaching risk and uncertainty respectively to known and unknown probabilities, the subjective nature of probabilities to decision maker where the attitude of ambiguity along with concept of ignorance has been considered as a measure of degree of confidence in the estimate of probability. Based on the desk review of the working paper authors of these survey generalize the issue of risk and uncertainty into the case of 2P(Achamu et al., 2022 Unpublished) to account both probability and possibility in the decision making process. One important attribute of adaptive modeling therefore is to realize the ignorance when new information imputed to the instrument as it helps to establish close relationship between reflection and action (Petit, 1976). Another critical issue in this schema is the possibility of incorporating financial risk especially to those that are credit-constrained farmers and hence accounting the two most, keeping dynamics in belief and preference of farmer as decision maker, risk in agriculture: risk aversion and downside risk (Hardaker et al., 2004). Farm financing as strategy of risk sharing on the other hand, however, magnifies risk unless optimal and viable financial structure exist, and two common problems in this regard, adverse selection and moral hazard due to informational asymmetry, therefore should be addressed during modeling in order not to bear both Type-I and Type-II. It is theory of utility from lender and borrower preference perspective seems viable in this case which however been elicited through the concept of certainty equivalence (CE) that better defines the problem at quadratic programing and a normative approach. C3s are more about advanced optimization methods and techniques than the importance of optimization problem in C2, as indicated by term Ethiopia and Malawi, in the problem domain. This is can be further justified by the terms coined in the cluster and including neural network (NN),deep uncertainty, hyper parameter in the area of Machine learning (ML) and hyper parameter optimization(HPO) to imply how problem in the agricultural study are being addressed in developing countries (see also fig.27). It further tries to show status of agriculture in general, the era of Agri4.0, and farming in particular where role of internet of thing (IOT) have been emphasized and generalized as precision agriculture. Similarly an in-depth analysis for remaining clusters may not be economical as far as each of them ain one or another way a touched theoretically by those discussed while C5 in its especial case, however, is very critical as far as policy direction is demanded. This is because of the potential exploratory modeling (EM) in giving robust formulation that might lend itself for flexible analysis of the decision process compared to the consolidative aproach(Bankes, 1993).

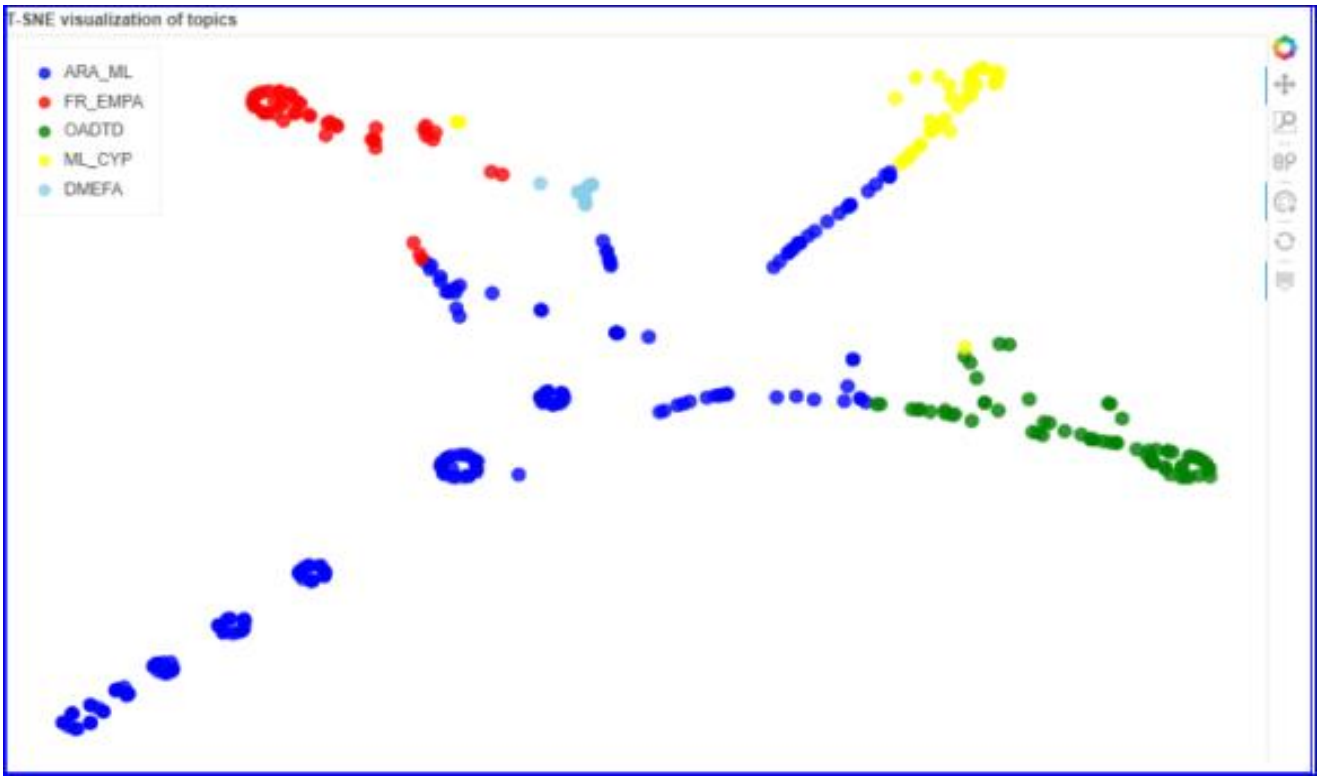


Figure 27 Structuring of topics for Title

ARA_ML: *Agricultural Research approach using Machine learning*,**FR_EMPA**: *Farming research and the importance Exploratory Modeling for Policy analysis* ,**OADTD:** *Operational analysis and Decision to Technological Development* ,**ML_CYP**: *Machine Learning Based crop yield prediction,* **DMEFA:** *Decision Making based on Explanatory Factor analysis*

With In this regard,(Langley, 2019)discussed, in depth, for three important agents: explainable agency, normative agency and justifiable agency. Each respectively mean that agency (model) (i) can provide, on request, the reasons for its activities ;(ii) if, to the extent possible, it follows the norms of its society; and (iii) if, it follows society's norms and explains its activities in those terms. This and the scientometric analysis result depicted in fig.14 demonstrates trend of solution approach, whatever the approach a decision maker has to follow, today is the era of big data and it is data science , mining and information extraction through the application of artificial intelligence AI matters. This is a remarkable development in decision making particularly for breaking the blurt boundary between positive and normative approach, i.e., neither purely normative nor positive approach exist. These two stream of standard branches, however, are methodological vector on the orientation and process to the theory of the firm (See; McKenna, 1996, p. 13) whereas this investigation is to the most three theory of firm: (i) managerial economic;(ii) behavioral economics; and(iii) transactional economics which all acknowledges uncertainties to nature of environments and concept of empirical study. Since, focus in this investigation is for farm financing decision particularly to crop production, of the available literatures generalization of production modeling approach for farming by(Antle & Capalbo, 2001) Petit(1976) as (i)utilization of representative farm model, commonly known as Representative farm aggregate (RFA) model,(ii), econometric models (iii) econometric based neoclassic models seems sound. Nevertheless, we rather found better insight from Petit (1976) discussion that generates three broad generations for both labeling as(i) econometric model based on statistical inference (EMBIS); (ii) Programing models (mathematical programing, MP) and(ii) General simulation models (GSM). Starting from their realization in the 1920's EMBIS have shown an impressive progress and appeal for some objectivity with care for not overstating to not account as optimization methods. On the other hand, as is obvious that mathematical programing emerged formally in the 1940's their utilization for agricultural problem began in the 1950 and vaguely continue tills 1960s. Comparatively, MP model help to organize a huge mass of information coherently better than EMBIS whereas GSM model that appeared in the late 1960 is with primary advantage of entailing both data flexibility and mathlemmatical structure.

Table 7.Clusters and expected schema

Cluster	Top occurred terms	Top relevant terms	Implication of top occurance	Implication of top relvant	Generalized implication
C1 - red balls	Theory ((29)	CTFM(3.14)	How approaches and methods should followed	Dynamics of finacial modeling	How Modeling approaches should be. Procedures and attributes of modeling.
	service (22)	systems practice(2.36)	Role and contribution of the deliverables	How practics at system level sould be realized	
	Prescrptive analytics (19)	learned policy(1.7644)	How briding prediction and decions is important	The importance of having flexiable policies	
	Scinece (15)	Iso (1.76)	The need of exploring new paradigm	The importance of standardizaion	
	Ackoff (14)	Sdp(1.723)	Wholnes is greater than sum of individual	How real world problem is both random and dynamics	
	Time series (13)	causal effect(1.6734)	Ternd effect and cusiality	The importance of studing exlanatorly and exloratorly	
	Modeling (12)	agricultural enterprise(1.647)	Realizing and capituring realities	How agricultre coulb be a potential source of busines	
	Reigm (11)	financing efficiency (1.5575)	The importance state of natures	Issue off finacila problems	
		scenario tree(1.4844)		There always exist more than one way	
C2 - Green balls	Ethiopia (41)	distressed debt(2.4176)	How agriculture stiill plays vital role in ethiopians economy	How likely the farm financing is banckptcy and the need of hedging	Farm modeling and and implication of finacial leveraging
	Crop yield (28)	price fluctuation(2.2638)	An explanatory variable	How House holds income highly dependnet to crop yiled	
	Topic, constraint (17)	Bankrapcy (2.1876)	How decisions modeling is imortant and is topic specific infact	Implies how uncontrolable factor and unforseen circustances should be treated in farm financing	
	Firm,inflation,innovation (16)	Representative farm(1.776)	The imporantance of exogenou varial in theory of firm	Indicates the research approache for farming in the 1960s at firm and aggrigate supply perspectives	
	Credit, acesse advisor (14)	Small scale irregation(1.6953)	Farming input, their acess and means	Mitigation approach of production risk	
	Adaptation (13)	Farm modeling(1.6762)	How natural variation influne farm decision	The need of	
	Household, risk mangement, extension and supply (11)	Farm level modeling (1.6555)	Micro level economic modeling	The importance of farm fiancing modeling at microlvel	
	Productivity (30)	portfolio management(2.3726)	General objctive of any study	How farm decision can correlated to invetsment decision	How recent appriaches in AI are
	Yield prediction (26)	Anfis([4]1.8765)	Farm income determination is most likely	The contemporary	

[4] Adaptive neuro-fuzzy interface system

C3 – blue			dependnet to yiled determination that is exogneous to faramer		atracting the farm finacing deision
	Neural network(10)	livestock production(1.7938)			
	Optimization technique (8)	food production(1.7734			
		multilayer perceptron(1.725			
		machine learning algorithms(1.6442)			
C4- Yellow	Trend(25)	p2p lending(2.2682)	Both crop yield and financial results can be attributed with time variation	Besides formal source of financing, informal fincing also common	Source of farm financing and major attributs of financing
	Bank (17)	Slice(2.2682)	Major source of formal financing is through banking		
	Support (10)	interbank market(1.6232)	Farm financing is polycentric and is open system that should be facilitated through advising for instance	The importance of global networking in financial institution	
	Loan (8)	Lending (1.6232)	An alternative source constrained farmers	Primary busines line of financical instiutes	
		Subsidiary(1.5656)		Agriculture as busines unit needs a policy that considers income sustainability	
C5- Purple	Interaction (20)	Mediterranean region (1.7492	Conveys precence of multiple actors in problem domain	Spatial and temporal perspectives of farm moeling	Exploratory modeling and policy analysis
	Adaption (13)		This is to implie how agronomic perspective of farming activities		
	Exploratory modeling, policy problem(11)		The importance of robust/ flexiable modeling in policy design		
	Deep uncertainty(8)		Farming decision is more than revealing risk based on probablistic nature of the state the world		
C6- Aqua	Soil (12)	Bio(2.258)	The need of corralating spatial and temporal componet of agriculture	Most suffics to the natural attributes and analogy	Representative Farm aggragate modeling (RFA-Model)
	Community(9)	agricultural production planning(1.9561)	Farm decision making process is highlu dependnet on the communities economic status	In todays farming farm level decision more is more of specific and hence farm level than RFA	
	Livelihood(8)	amara regional state(1.487)	Critically farmers lives majorly correalted with farm productivities	It is clearly evidencial for the region to relie the livelihood of the community on farming activity	

5. Conclusion

Using the scientometric and topic modeling as methods of extracting knowledge to the state of art of farm financial decision distinct publication analyzed. For the analysis publication source retrieved from database, API and reference manager in terms of bibliometric, network and text data.

The analysis majorly done for title, abstract and keywords of publication for the purpose of extracting knowledge and identifying research schema on the problem and it is the scientometirc analysis deployed for this case mainly based on bibliometric analysis. Using bibliometric analysis and the resulted visualization map through co-authorship, co-occurrence and citation interest of scholars and organization captured to the problem domain captured particularly word co-occurrence analysis emphasized since it uses field of study as analysis unit that helps to distinguish the focus or center of publication under investigation.

Importance of co-occurrence analysis found essential via cluster output of keywords to indicate the schema or taxonomy of research interest while by cluster analysis any keyword supposed to single cluster and hence to avoid such binary effect a topic modeling deployed.

Topic modeling mainly as method of unsupervised classification of publication utilized not only for clustering of keywords but also for semantic analysis. Since it can scan set of documents and detect word and phrase patters words in this approach found to any proportion than either unity or null.

Among popular topic modeling types utilized in this survey includes Latent Dirchilate Allocation (LDA) and BerTopic topic modeling at which the first is mainly for topic generation and the latter is for post topic analysis. In both cases, the analysis made with those most packages like genism and scikit-learn from python platform to facilitate the analysis and make clear the visualization topics.

Started with dictionary and corpus preparation, the generated model from LDA evaluated with perplexity and coherent analysis and followed by improvement of topic using BerTopic particularly for similarity and hierarchical clustering.

Finally, using clusters obtained from both analyses, discussion and interpretation made to state of art of farm financing and the finding signifies that besides the traditional bio-economic models that is very prominent particularly to agricultural decision modeling, incorporating the emerging technology like internet of things, Machine learning, data mining found critical to today's agricultural development where precision agriculture has been demanded. With this development progress and issues related to financial decision found with little concern and few papers tried to incorporate it as direct decision variable as confirmed from the result of both Scientometric and Topic modeling analysis. It is straightforward to acknowledge finical problems to consider as decision variable that describing their effect and making exploratory analysis to such financial turmoil environment.

Reference section is intentionally left blank for the moment