

---

# Bibliometric Analysis of Information Communication Technology for Sustainable Development: A Machine-learning Based Approach

---

Dr. Rudolph Oosthuizen

Department of Engineering and Technology Management, University of Pretoria and Defence Peace Safety and Security, CSIR, Pretoria, South Africa

Prof. Leon Pretorius

Department of Engineering and Technology Management, University of Pretoria, Pretoria, South Africa

**Abstract:** Publication of research outputs is a method of researchers to capture their knowledge generated. Analysing the publication topics and trends in a research field can provide insight into the main research trends. A Bibliometric analysis, based on the topics from published literature, provides insight into the focus areas and trends of a research field. The objective of this paper is to extract the main research topics from papers on “Sustainable Development” and “Information Communication Technology”. The research topics are extracted from the abstracts and titles of papers using machine-learning for topic modelling. This paper identified the topics of Knowledge Management, Design Process, Social Change, and Smart Systems, as the primary focus of research into Information Communication Technology for Sustainable Development. A deeper analysis into Smart Systems identified quality of citizen life, solutions for the urban setting, energy, and the environment as key research concerns.

**Keywords:** Sustainable Development, Information Communication Technology, Natural Language Processing, Topic Modelling, Research, Bibliometric, Literature, Machine-learning, Knowledge, Research Trends, Research Roadmap, Abstracts, Titles

---

## 1 Introduction

Sustainable Development (SD) is an organizing principle for meeting the development needs of the world without compromising the ability of future generations to sustain themselves. The aim is to achieve a society where living conditions and resources continue to meet human needs while the integrity and stability of the natural system are maintained (UN, 2015). To this end, the United Nations has developed the Sustainable Development Goals (SDGs) (SDG, 2020). However, these goals are broad-based and interdependent. SD is required for maintaining a proper balance between development and the exhausting resources for human existence. The three pillars of sustainable development include (Punia, 2016):

1. Economic development to remove poverty and increase economic welfare.
2. Social development to improve the quality of education, housing, etc.
3. Environment development to reduce pollution and protect environment sources.

Information and Communication Technology (ICT) covers a combined set of physical, infrastructure and human resources. The spread and diffusion of ICT through societies and global interconnectedness have great potential to accelerate SD and bridge the digital divide by developing knowledge societies. ICT may contribute to achieving SD; however, it is a means not an end (Punia, 2016). Due to the geographical separation and multifaceted nature of SD, it also requires ICT as a critical component for the integration and exchange of information. Therefore, ICT has a major role in knowledge sharing, decision quality, inter-organizational links, and searching for the solution of the implicit conflict between sustainability and economic growth (Mohamed, Murray and Mohamed, 2010; (Ianioglo and Polajeva, 2017).

New ICT technologies and services enable public- and private sector organisations to perform their activities more efficiently. ICT can substantially accelerate the development progress of societies by constructing knowledge and bridging the digital gaps between communities. ICT can reduce the distance between individuals, communities and nations around the globe. This has a major impact on economic growth and social relationships. In developing countries, ICT applications can improve the quality of life for societies (Credé and Mansell, 2014; Punia, 2016). Examples of ICT contributions may include the following (Credé and Mansell, 2014):

1. Transport Sector. Advanced telematics can maximize road-transport efficiency, improve road safety, and counter the environmental problems of pollution, congestion, and resource consumption.
2. Health Care. Modern ICT can enhance the effective exchange of medical information, such as patient records, between health professionals. Education of health- workers isolated in rural areas can also be overcome.
3. Disabled and Elderly. The disabled and elderly can be assisted to achieve more independent lifestyles, with improved social integration. People with learning difficulties or disabilities can have access to education curricula for informal learning.
4. Environment. An ICT network provides access to environmental information about flooding, air- and water quality, industrial hazards, and forest fires.
5. Agriculture. Information systems can collate information on hydrology, soils, and rainfall to support socioeconomic decision-making and planning.
6. Science. New and improved tools are available to access, process, and share information and collaborate with other researchers.
7. Manufacturing. ICT supports factory automation, planning and control, and business management. Integration of process chains can accelerate production. Designing machine tools and parts can benefit from computer-aided design and interactive graphics.

8. Education. ICT provides more opportunities for learning activities for children and even adults. This provides opportunities for blended learning (Sanjeev and Natrajan, 2019).

Despite still having long-term sustainability issues to be solved, ICT could play a key role to support global economical, social, and environmental sustainability (Wu et al., 2018). However, new ICT-based opportunities require modern communication infrastructure and computer equipment matched to software applications. Successful implementation of ICT for SD requires scalability, integration, and sustainability in infrastructure. The implementation of ICT provides systems to store, access, manage and disseminate environmental data and information. Therefore, ICT could support the sustainable assessment of the situation to manage the negative impact of development on the environment. Meaningful content should be available to end-users (Credé and Mansell, 2014; Punia, 2016).

The implementation of ICT into the public and private sectors has different perspectives and impacts. ICT enables the public sector to create interaction, increase efficiency, improve quality of service delivery, reduce operating cost, as well as to provide increased transparency. On the other hand, the impact of ICT on the private sector includes expansion, economic development, increased productivity and efficiency, as well as growth. However, achieving impact in the public sector may take longer to materialise due to the lack of competition (Gatautis et al. 2015).

As the role of ICT in SD seems to be of importance, it should be the focus of researchers in the field. This leads to the question of how this research is progressing as well as what are the topics researchers are focussing on. This paper aims to analyse the bibliometric data of published papers on the role of ICT in SD. However, the number of publications published in most research fields continue to grow at a fast pace, the time and cost to perform an in-depth analysis are increasingly becoming prohibitive. Artificial intelligence-based machine-learning methods provide a useful method to extract latent topics from large corpora of text from research publications. Machine-learning methods require Natural Language Processing (NLP) to support data mining of text. NLP converts text into numerical data for the algorithms to process (Antons, Kleer and Salge, 2016). Automated NLP techniques, with topic modelling algorithms, can extract topics from a large corpus of text documents. Allocation of the most prominent topic per paper helps to generate the evolutionary trends over the publication period.

This paper therefore will first discuss the performing of bibliometric analysis using machine-learning algorithms and NLP. The contribution of topic modelling and its execution is explained. Next, the bibliometric analysis process is presented before executing the process on appropriate SD and ICT papers extracted from the Scopus database using selected search terms. This is followed by presenting and discussing the outputs and results of the bibliometric analysis process, before posing some conclusions.

## **2 Bibliometrics and Natural Language Processing**

The growth and development of a research field require communication and publication of research outputs (Valerdi and Davidz, 2009). In general, the research aims to understand, explain, and predict phenomena observed in a field to create knowledge that

stimulates growth. During the research, scientists codify their outcomes in publications. Academic publications are the building blocks of science. Validation of the published research is provided through peer reviews, which is a form of expert-based judgment. Scientometric (the “science of science”) analysis approaches are increasing in popularity for the assessment of this published research. Bibliometrics is an approach to perform a scientometric analysis that provides a quantitative, statistical, and systematic analysis to measure progress and trends in a research field (Van Raan, 2003).

Bibliometric approaches are reported to have been applied in behavioural science, engineering and medicine research fields (Jiang, Qiang and Lin, 2016). Bibliometrics, through visualisation of patterns in bibliometric data, informs researchers about the state of a research field. Patterns may occur in quantitative bibliometric data about leading authors, growth, performance, maturity, and intellectual mapping (Van Raan, 2003; Kalantari et al., 2017). The data for the bibliometric analysis may come from papers in a single journal, collection of journals, or any publication platform if a keyword search was performed on an academic database, such as Scopus (Keathley et al., 2015). Research progress in a field may be described in terms of the conceptual structure for scientific mapping or citations for performance analysis. Typical bibliometric indicators include publication statistics (number of papers), author statistics (citations), relational indicators (co-occurrence of words, co-citations, and co-publication) (Van Raan, 2003; Jiang, Qiang and Lin, 2016; Kalantari et al., 2017; Jia et al., 2018; Eker et al., 2019).

Another popular bibliometric analysis approach is to process a set of publications to extract and identify core topics and to visualise their evolution trends over time. Plotting the rise and fall in publication numbers of research topics over time provides valuable information for research investment decisions. Also, the emergence of new topics and concepts may indicate new directions of research (Lamba and Madhusudhan, 2019). Based on this discussion and motivation in this paper, topic modelling, with text-mining and NLP techniques, is implemented to extract, identify, and analyse the main topics from paper titles and abstracts.

With the fast-paced increase of academic scientific publications published each year, the traditional bibliometric methods to derive data from manually reading, analysing, and sorting papers by predefined topics may be prone to errors, time-consuming, and difficult (Eker et al., 2019). Manually building a predetermined topic list is difficult, as it can suffer from the subjective judgment of the researchers. It may also miss the latent topics from a large text corpus. Furthermore, being consistent throughout the analysis process may not be possible. Some articles may contain multiple topics while the predetermined topic categories could be missing new and emerging topics. Implementing automated software-based machine-learning methods will improve this process (Lee and Kang, 2018).

Data mining, using machine-learning methods, requires NLP to prepare the free text data for processing. The text needs to be converted into quantitative data for the algorithms for processing and deriving meaning. The NLP tools can also extract keywords, phrases, patterns, and relationships to interpret, cluster, categorize, summarize, and classify large amounts of input text (Agrawal, Fu and Menzies, 2018). With the ever-improving computer-processing power and capability of algorithms, more researchers start to apply these methods in bibliometric research. Topic modelling is one such method that implements unsupervised text classification with quantitative statistical algorithms to

extract semantic information from text. Supervised machine learning methods require labelled data to train algorithms. The advantage is that topic modelling is not dependent on prior knowledge of the topics in the text to be processed (Antons, Kleer and Salge, 2016).

The topic modelling processes unstructured (unlabelled) text from a body of documents to quantitatively discover underlying structures and latent themes. Topic modelling assumes that the mixture of words in a human-created document constitutes a set of latent topics aimed at conveying a message. The algorithm then analyses the occurrence relationships of these words to define different topics as a probability distribution. However, the method still requires human input with domain knowledge to interpret and name the extracted topics from the data (Kunc, M. and Mortenson, M. and Vidgen, 2012; Jiang, Qiang and Lin, 2016; Tong and Zhang, 2016; Ma et al., 2018)

The Latent Dirichlet Allocation (LDA) is a popular contemporary algorithm implemented for topic modelling. LDA processes the input text from many documents to generate a statistical model where unobserved groups of words explain the similarity of data (Blei, Carin and Dunson, 2010; Suominen, Arho and Toivanen, 2016). The LDA assumes that each document contains a set of words, where the order of the words does not influence the detection of latent topics, only the co-occurrence of the words per document in the corpus (Antons, Kleer and Salge, 2016; Lee and Kang, 2018). The topic-word and document-topic pairs are defined probabilistically to extract the required number of topics. Topics can then be assigned per document by the highest associated word probabilities (Tong and Zhang, 2016; Agrawal, Fu and Menzies, 2018; Eker et al., 2019).

The required number of topics for extraction is one of the inputs for the LDA algorithm. Multiple topics may be assigned to a document, each with a different calculated probability. The unique topic profile per document enables researchers to analyse the structure and temporal behaviour of the research field in the data (Antons, Kleer and Salge, 2016). Over the past five years, LDA as part of NLP has been applied to some extent to assess the evolutionary patterns of topics in the computer science, medicine and technology management research domains. However, the method is still young and improving; further applications are continuously researched as AI is improving (Jiang, Qiang and Lin, 2016; Tong and Zhang, 2016; Lee and Kang, 2018; Suominen, Arho and Toivanen, 2016). This methodology has not yet been extensively implemented and applied in the field of sustainability research (Tallón-ballesteros and Hutchison, 2019). This will therefore also be an additional focus in the rest of this paper.

### **3 Topic Modelling and Assessment Process**

As alluded to in the introduction, this aim of this paper is to analyse the topic landscape for research published on ICT as part of SD. These evidence-based insights should guide researchers to identify opportunities and plan their efforts (Antons, Kleer and Salge, 2016). The bibliometric analysis approach using topic modelling discussed in the previous section(s) and applied in this paper is shown in Figure 1.

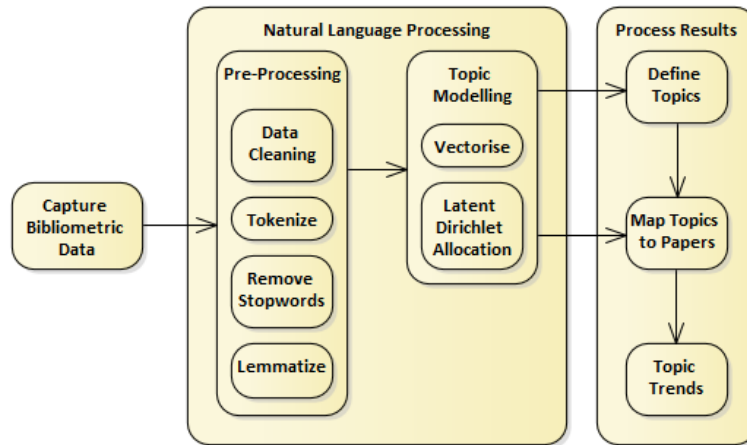


Figure 1: Topic Modelling and Analysis of Bibliometric Data

The entire research process is implemented here using data science principles executed with Python-based algorithms.

### 3.1 Import Journal Bibliometric Data

The process in Figure 1 starts with collecting the raw bibliometric data. The analysis in this paper processes the bibliometric data from academic articles published on ICT and SD as collected from a directed search in the Scopus database on 22 May 2020. Scopus was used as it is considered to be one of the largest databases on abstracts and citations of peer-reviewed literature. Only the text from the articles' title and abstracts was used as a proxy for full articles. Titles and abstracts provide a compact representation of the whole article, normally having the problem description and method along with some of the research results. It should provide enough representative important words about the research themes of the article (Agrawal, Fu and Menzies, 2018; Lee and Kang, 2018; Tallón-ballesteros and Hutchison, 2019).

The Pyblometrics library in Python enables the capture of bibliometric information on published journals listed in Scopus (Rose and Kitchin, 2019). The Pyblometrics tool also affords the capturing of papers from the result of keyword a search in Scopus. The results then will span over a multitude of journals, conference proceedings, and other publication platforms. Capturing bibliometric data for analysis of research activity on an interdisciplinary and evolving field, such as Sustainable Development, can be done through a basic search of keywords. However, selecting the best set of search terms may become difficult (Hassan, Haddawy and Zhu, 2014). The search terms used in the research for this paper is "TITLE-ABS-KEY ("Sustainable Development" AND (ICT OR "information communication technology" OR "information and communication technology"))". These search terms resulted in 2164 papers after papers with incomplete bibliometric data has been removed. The remaining papers cover the period from 1994 to 2020.

Python's Pandas library is also implemented to store the data to a .CSV file for further processing of the data in the subsequent steps. The bibliometric data captured for the analysis in this paper include the year of publication, authors, paper title and abstract. The data capture tool also provides access to other bibliometric data, which may typically include author affiliations, author country, author count, and citations. However, this is not required for the focus of this paper.

### 3.2 Natural Language Processing

- 1 Pre-processing. The SpaCy library in Python was applied to perform the NLP on the captured bibliometric data in this paper. The raw captured text often requires pre-processing for structuring, cleaning, and preparation for analysis. First, the titles and abstracts need to be combined into one document to increase the size of the text corpus for improved topic modelling. Clean and structured text is a prerequisite for extracting valid topics. The Pre-Processing steps ensure that the captured unstructured text is structured and transformed into a format suitable for analysis:
  - a. Data Cleaning. Most of the titles and abstracts of the captured papers for this SD and ICT research contain unwanted words and characters (e.g. Elsevier, Wiley, published, copyright, ©), numbers, punctuations, special characters, and dates. Because these elements tend to complicate the NLP processing algorithms, the data was manually cleaned in Excel, as well as with Python functions. Key abbreviations and acronyms, such as ICT, SD, and SDG, were replaced with the whole words to improve the accuracy of the NLP and topic modelling. Because the spelling of some relevant words differs between the United Kingdom and the United States English (e.g. 'behaviour' vs 'behavior' and 'modelling' vs 'modeling'), they were changed to the United States format. The search terms used to capture the bibliometric data were also removed as they are present in every paper. These terms do not add information on the topics present in the text corpus.
  - b. Tokenize. The tokenization function extracts the linguistic building blocks for sentences or paragraphs. If required, the key phrases having strong meaning, independent of the individual words, can be extracted using the algorithm (Patel and Soni, 2012; Lin et al., 2016).
  - c. Remove Stopwords. Stopwords are the common words without contextual meaning that need to be removed. Typical stopwords include "as", "and", "the", "if", "a", etc. They have a high frequency that adds noise when vectorising to the text. Also, the academic nature of the articles published in the journals provides other that do not provide information on the specific research topic. These typically include words for this SD and ICT research such as "paper", "research", "study", "describe", "example", "article", "literature", and "introduce".

- d. Lemmatization. Lemmatization also reduces the dimensionality of the text. It normalises the text by combining derivatives of a word, such as plurals and past tenses. Lemmatization is preferred to stemming, which truncates affixes off words to reduce them to their root. The lemma for each word is determined through morphological analysis and a vocabulary with part-of-speech information. The drawback of stemming is that the context of a word is not considered, causing to some stems not being an existing word (Lee and Kang, 2018; Eker et al., 2019).
- 2 Topic modelling. The Scikit-learn library in Python provides the vectorization and LDA to process the prepared text data from the previous step.
    - a. Vectorize. The CountVectorizer function transforms the text into a document-term matrix. The vectorisation algorithm implements a minimum and maximum document frequency setting (max\_df and min\_df) of the words to be included for processing. These parameters enable the algorithm to exclude uncommon and too common words throughout the papers in the corpus. The number of documents in the corpus, the number of terms per document, and the distribution of individual words over the documents affect the optimum values for max\_df and min\_df. The values were selected by calculating the perplexity score of each set of parameters for this SD and ICT research. Perplexity is a measure that indicates how well a probability model predicts a sample. The selected max\_df and min\_df should provide the lowest perplexity as well as satisfy other heuristic requirements, such as text sample size, the number of topics and interpretability of the topics (Jie et al., 2014; Agrawal, Fu and Menzies, 2018; Eker et al., 2019).
    - b. Latent Dirichlet Allocation. The LDA function processes the document-term matrix to extract the predefined number of topics. Again, the perplexity scores are calculated to determine the optimum number of topics suited for the size and diversity of the text data (Jiang, Qiang and Lin, 2016; Tong and Zhang, 2016). However, the LDA is a probabilistic algorithm, where every run of the code results in a different set of topics with different describing terms. A good set of parameters would typically result in an 80 per cent match of topics, with similar terms, between the different runs. Most of the extracted topics would be relatively easy to identify using prior knowledge of the field. Despite the probabilistic unsupervised machine-learning nature of the algorithm with minor inconsistencies, the extracted topics would still be useful to determine trending research topics and focus areas from the input text.

### 3.3 Process Results

The Python code used to perform the topic modelling also assigns the best fit topic extracted to each paper in the corpus. The output is again saved to a .CSV file for further



processing and analysis. Python, using the Pandas, Numpy and Matplotlib libraries, along with Microsoft Excel was used for in-depth processing and visualisation of the output data for this SD and ICT research. The next step is to define and name each of the extracted topics, as the algorithm only clusters documents by their topics, without interpreting and identifying them. Manual analysis, with domain knowledge and subject matter expert insight, is still required to interpret the LDA results, along with the associated bibliometric data, to assign a descriptive topic to each cluster of terms. Unfortunately, this remains a subjective process and may be prone to bias from the researchers (Antons, Kleer and Salge, 2016; Lee and Kang, 2018).

Word clouds, also known as tag clouds or Wordles, can be generated to support the interpretation of the topic terms. A word cloud is a weighted list of words that visually represents text data, where the importance of the words is graphically depicted using different font sizes or colours. Word clouds enhance the cognitive ability of a human, with suitable domain knowledge, to quickly form a visual image of the importance of terms relative to each other (Bashri and Kusumaningrum, 2017). In the case of SD and ICT presented in this paper, the probability distribution of terms per topic is used to depict the relative significance each word per topic.

Another way to interpret patterns in the bibliometric data is to review the trajectory trends of the extracted research topics over time. As all of the articles processed in this paper have a publication date, the topic occurrence can be plotted over time. First, the number of papers per topic per year is determined. Because the number of papers for this SD and ICT research theme increased over time, the number of papers per topic per year will not be useful to plot research topic popularity trends. Therefore, the counts need to be normalised by dividing the topic counts results by the total papers for that year.

Due to the fluctuations in the allocation of papers to topics per year, the data points become difficult to interpret. Therefore, for ease of interpretation, a fourth-order polynomial regression was used to smooth out the trends over time. The fourth-order binomial regression was calculated, using the Scikit-learn library in Python, on the data to derive the trends in the topics as a relative, not absolute, plot.

## **4 Results**

### **4.1 Parameter Selection**

Selecting the most appropriate values for the key vectorisation and LDA parameters for most effective topic extraction is not an easy task, but critical to the accuracy of the topics extracted from the text. As described in the process description from section 3.2.2., perplexity is calculated to evaluate the LDA model. Some heuristics were also used to help select the parameter values for vectorisation and topic modelling. An iterative parameter analysis was performed using a range of values for `max_df`, `min_df`, and number of ICT for SD paper topics, as seen in Table 1. These ranges of terms were found to be suitable for a large corpus of short documents.

Table 1: Monte Carlo Parameter List

Parameter	Start Value	End Value	Step size
Topic Number	7	20	1
Max df	0.85	0.99	0.02
Min df	0.1	0.25	0.01

A sample of the output values is presented in Table 2. The data show that a model with a small number of terms extracted from the texts, to describe the topics, produces a low (preferred) perplexity score. Min\_df affects the number of terms available for a topic generation as it determines the number and diversity of terms available for topic extraction in the document-term matrix. A min\_df value of 0.17 means that a specific word must be present in at least 17% (368) of the total number of captured ICT for SD documents to be included for topic extraction. However, a small number of terms may result in the subtle or hidden topics being missed.

Table 2: Sample of Parameter Analysis Output

Number of Topics	Min df	Max df	Topic Terms	Perplexity
8	0.24	0.87	12	11.51
7	0.19	0.87	22	20.87
11	0.17	0.95	37	32.71
13	0.17	0.99	37	32.88
12	0.17	0.93	37	32.91
15	0.17	0.87	37	32.96
13	0.15	0.93	48	41.05
18	0.14	0.87	61	52.39
9	0.1	0.87	99	77.47

As seen in Table 2, a higher number of terms result in an increased perplexity as the model is more prone to errors. However, a small number of topics may be too generic while a large number of topics becomes too vague, as well as difficult to interpret. For this paper, the number of terms was chosen as 37 to enhance the generation of clear word clouds. As an initial step, topic modelling was performed using values for 10 topics (min\_df = 0.17 and max\_df = 0.95) to generate five sets of topics. Evaluation and comparison of these different sets of topics resulted in at least 15 different unique ICT for SD topics. Therefore, the decision was made to select the parameters for 15 topics and 37 terms, which also have the lowest perplexity. These values are shown in the shaded cells of Table 2 as a min\_df of 0.17 and a max\_df of 0.87. However, the selection of parameters will ultimately depend on the focus and purpose of the overall analysis.

## 4.2 Publication Trends for ICT and SD Papers

The number of papers published per year, extracted with the search terms of “Sustainable Development” and “Information Communication Technology” for bibliometric processing in this paper, is shown in Figure 2. From the graph, it is clear publications on this theme only started to receive prominence from the beginning of the 21st century. It is interesting to note the sudden rise in 2019 after the number of publications stabilised from 2015 to 2018.

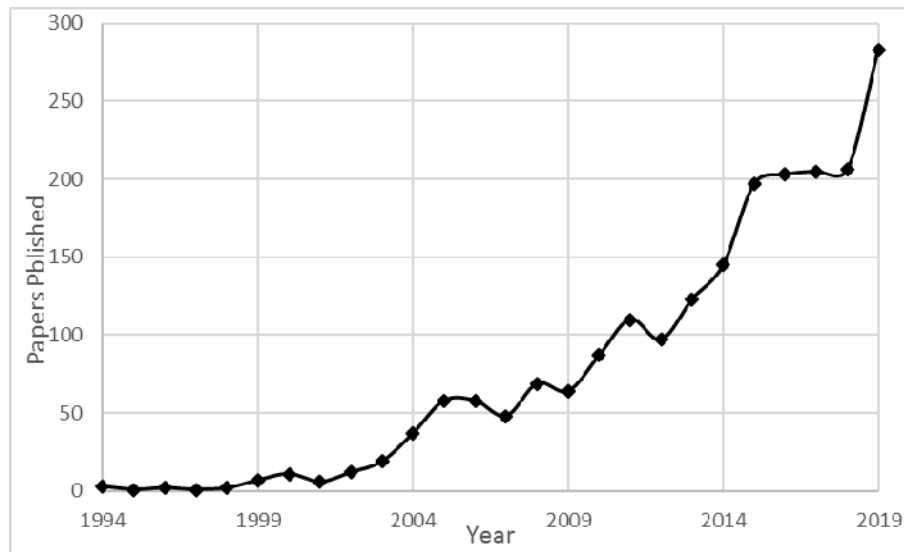


Figure 2: Papers Published per Year

The data for the year 2020 is excluded as the year is not yet complete and the actual number of papers not available. The primary publication platforms (top 10) for providing the research papers extracted in for this analysis performed using the techniques motivated and described in the previous sections are shown in Table 3.

Table 3: Publications Sources

No	Publication	Papers
1	IFIP Advances In Information And Communication Technology	65
2	Advances In Intelligent Systems And Computing	63
3	ACM International Conference Proceeding Series	50
4	Sustainability Switzerland	45
5	Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics	37
6	CEUR Workshop Proceedings	35

7	IOP Conference Series Earth And Environmental Science	30
8	Journal Of Cleaner Production	29
9	Communications In Computer And Information Science	24
10	ICT For Sustainability 2014	23

The figure and table highlight the growing importance of the field. However, this provides no insight into the structure and focus areas.

### 4.3 Sustainable Development Research Topics

The primary output of this paper is a list of SD and ICT research topics extracted from the corpus of the captured papers' titles and abstracts. The list of topics with their word clouds of defining terms is shown in Table 4. Since this is unsupervised machine learning, a slight variation may occur for each run of the algorithm. However, the accuracy is adequate for processing the trends of the research topics over the publication history of the journals. This process is preferred to manually read and classifying a total of 2164 journal abstracts.

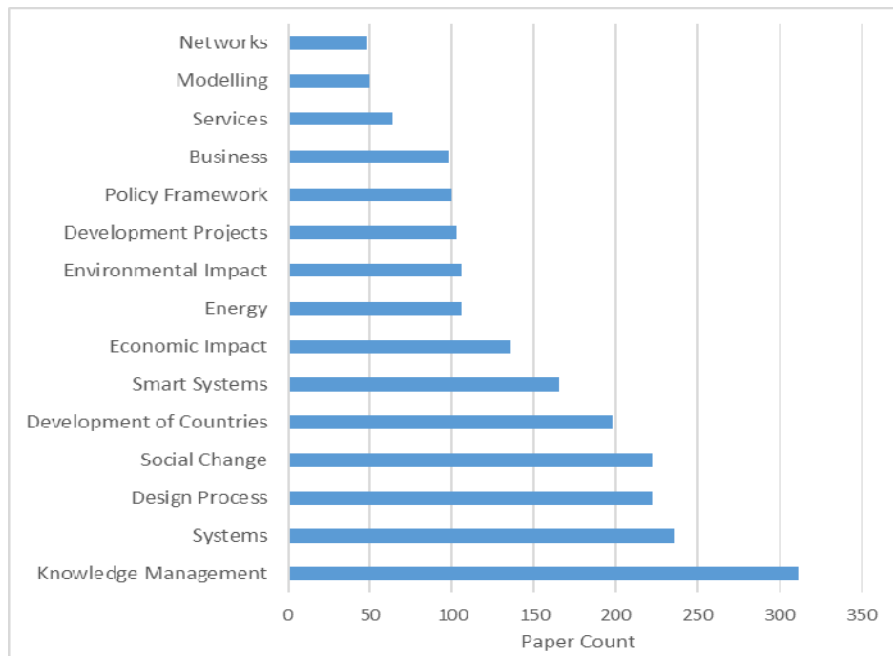


Figure 3: Papers per Topic

As the topic modelling algorithm also links the highest probability topic to each paper, the total number of papers per topic can be determined to indicate the focus areas within the research field. The number of papers per topic is shown in Figure 3. It should be clear in the graph that Knowledge Management is the leading extracted topic in ICT for SD

research. Other primary focus areas include the topics of Systems, Design Process, Social Change, Development of Countries, Smart Systems and Economic Impact.

#### **4.4 ICT for SD Topic Trends**

Because the dates of the journal papers and their assigned topics are available in the processed data, it is also possible to extract the publication trends of the topics over time. As significant numbers of papers per year only occur after 2003, the analysis was only performed for the period 2003 to 2019.

For simplicity and interpretability, the trends are split between increasing trends (Figure 4), decreasing trends (Figure 5), and stable trends (Figure 6). The characteristics of the trends to be considered are the shape and size of each graph.

The vertical axis indicates the proportion of the topics allocated to papers per the total number of papers published for each year in the corpus of documents processed. The values on the vertical axes of the three sets of graphs are kept the same to assist in the comparison of the trends. Even though the presented data only include documents from 2003, the total number of papers for the initial years remains low.

As a result, some of the topics may present a very high percentage of publications in the initial years. Therefore, interpretation of the topic trends should rather focus on the later years.

Table 4: Topics from the ICT for SD Papers

Number	1	2	3	4	5
Name	Design Process	Smart Systems	Social Change	Services	Networks
Wordcloud					
Number	6	7	8	9	10
Name	Business	Policy Framework	Systems	Energy	Environmental Impact
Wordcloud					
Number	11	12	13	14	15
Name	Modelling	Development Projects	Country Development	Economic Impact	Knowledge Management
Wordcloud					

In Figure 4, it is interesting to note that the extracted topics of Smart systems and Development of Countries are showing the biggest increasing trend. The topic of Economic Impact, which can be seen as being linked to the Development of Countries, also presents the same trend. Therefore, it can be deduced that research into the use of ICT in the context of SD, increasingly focuses on smart system technologies to have an impact on the economy and development within different countries.

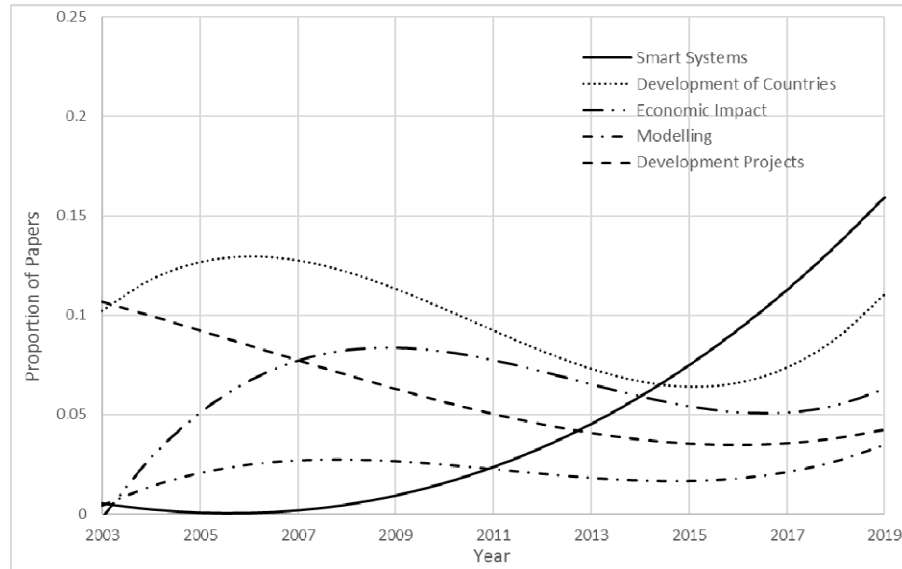


Figure 4: Upward Trends in ICT for SD Papers Published

From the trends in Figure 5, it can be seen that Knowledge Management, despite being the leading extracted research topic, has a decreasing trend over the last number of years. This could be the main reason for the flattening of the curve from Figure 2 over the same period. It seems that Knowledge Management is not seen to be a leading indicator of ICT for SD anymore. An interesting observation is a decrease in the Business topics relative to the increases of Development of Countries and Economic Impact topics from Figure 4. Also, the decline of the Energy topic should be noted. However, the graph does not imply that the absolute number of publications in these topics is decreasing, it is the relative importance (percentage) to the other topics over the same period.

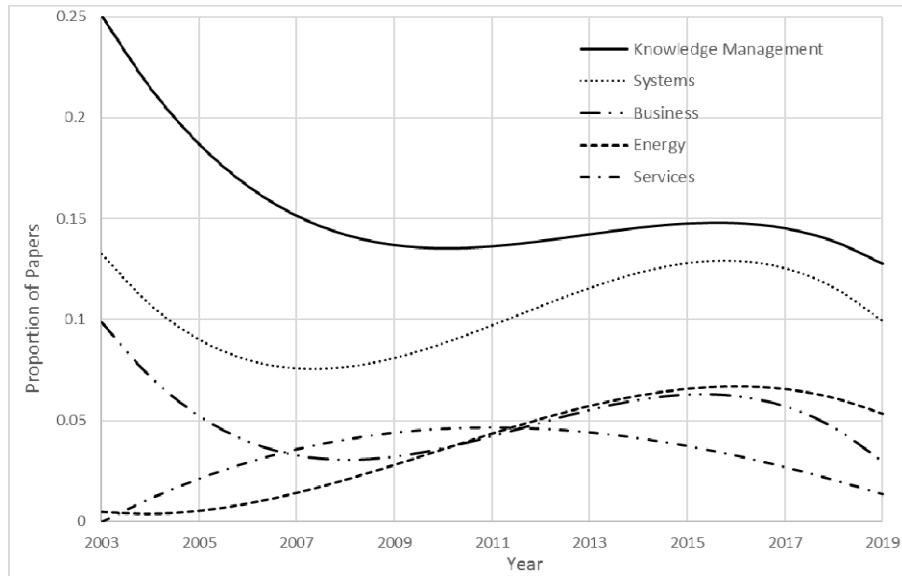


Figure 5: Downward Trends in ICT for SD Papers Published  
 The trends in Figure 6 represent extracted topics with only slight increases or decreases in publications relative to the other topics published of the same period.

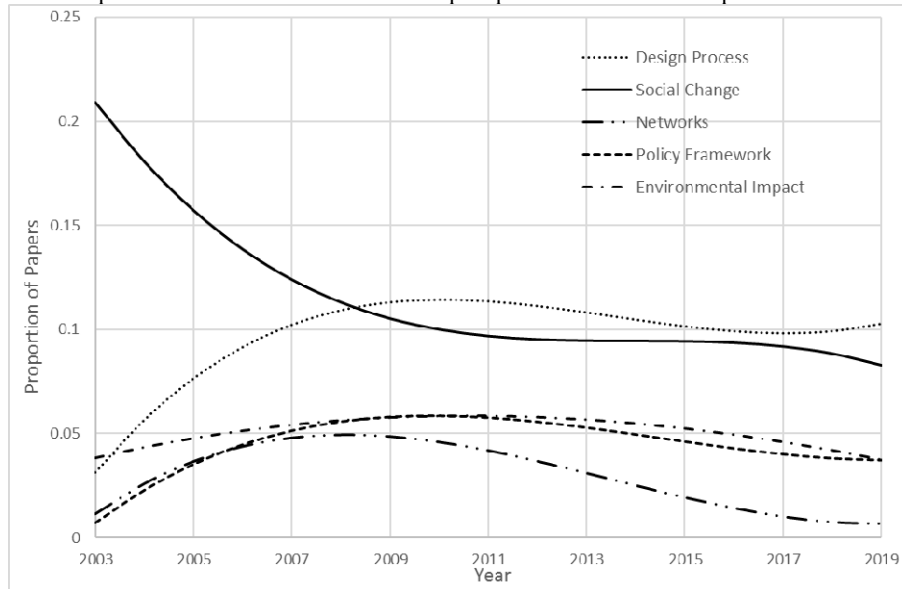


Figure 6: Stable Trends in ICT for SD Papers Published

Despite Environmental Impact being one of the main features of SD, it does not seem to have a high impact in the context of ICT. The same goes for Policy frameworks. These two topics may present opportunities for future research. The topics of Design Process and Social Change are some of the topics with the highest popularity, but they remain



constant relative to other topics. Therefore, future research may combine these with other topics to take a different angle into ICT for SD.

#### 4.5 Keywords Extracted from Papers Published on ICT for SD

Another output of a bibliometric analysis in ICT and SD is the frequency of keywords, as seen in Figure 7. The obvious keywords used in the search for the reports would be too numerous and is excluded from this graph. It is more informative to analyse the frequency of “other” words in the corpus of documents. Figure 7 provides the total count of keywords over all the papers. The keywords of “Smart” and “City” have a high overall count as well as “Energy” and “Services”. However, from the topics related to energy and services, from Figure 5, are decreasing. Also note the frequency of “Digital”, which is much lower than the other words. Therefore, the analysis here will focus specifically on these keywords.

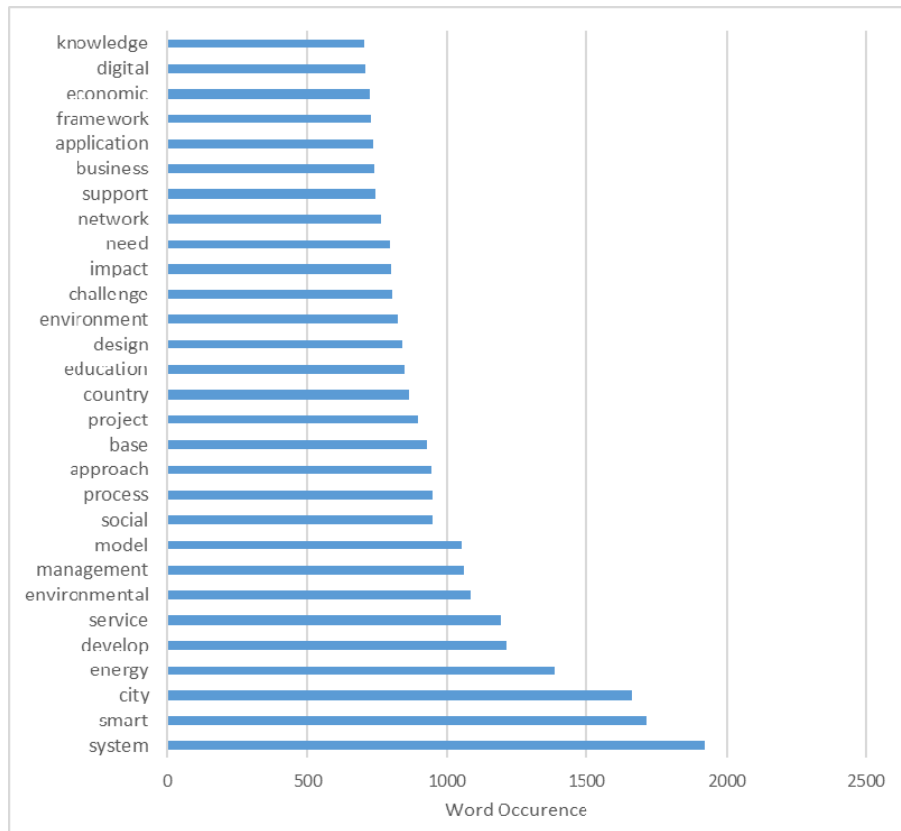


Figure 7: Keyword Frequency in Papers Published on ICT for SD

The trends of the selected keywords since 2005 are shown in Figure 8. The data for 2020 is excluded as the year is not yet complete and would skew the trends in the graph. Note the sharp increase of “Smart” and “City” relative to the other words of interest. However,

the number of papers published on ICT and SD has the same increase over the corresponding period, except for the period from 2015 to 2018 where a stagnation was experienced. This trend corresponds with the “Energy” and “Services” keywords. Despite the stagnation, the keyword occurrences of “Digital” also increased along with “Smart” and “City”.

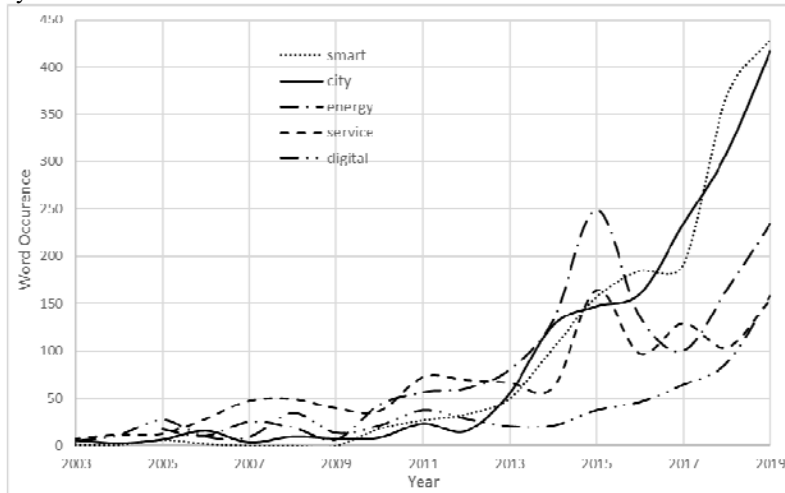


Figure 8: Selected Keyword Trends in Papers Published on ICT for SD

To factor out the effect of the increasing number of publications, the keyword frequencies were normalised in Figure 9. Still, despite being normalised the keywords of “Smart” and “City” still have an increasing trend.

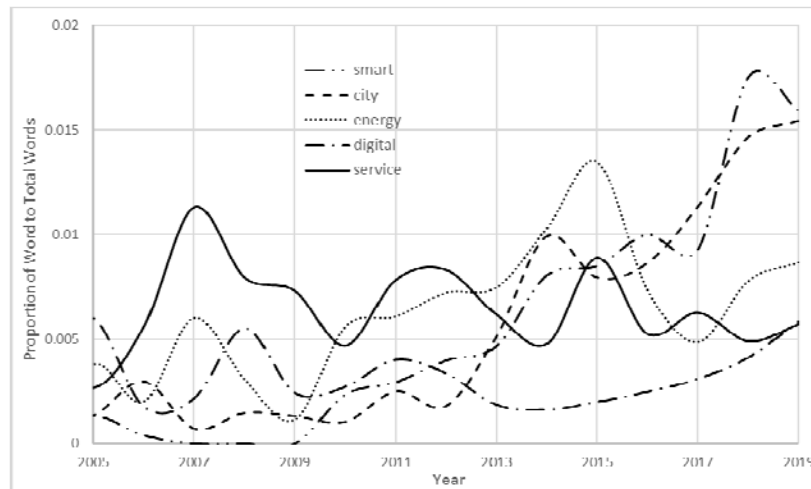


Figure 9: Normalised Keyword Occurrence in Papers Published on ICT for SD






This is in contrast with “Energy” and “Services” that are flattening out. This indicated that a greater portion of the published papers tends to be about the words “Smart” and “City”. The keyword “Digital” presents similar behaviour, albeit at a smaller scale.

Because the keyword frequencies present similar behaviour to the topic trends, it provides validation for the machine learning-based topic modelling approach implemented in this paper. Topic modelling provides a richer searching tool than only using selected keywords. Since the topic of Smart Cities seems to be of interest to the research field, a deeper level of analysis is performed by again implementing the process described in section 3.2.2. The aim is to gain a more in-depth understanding of what the focus topics within Smart Cities research are related to the context of ICT for SD.

#### 4.6 Expansion of the Smart Systems Topic

As seen from the outputs discussed in the preceding section, Smart Systems, and city, in particular, seem to be important topics, especially in the recent past. Therefore, the bibliometric analysis process, described in section 3.2.2, was redone on the papers allocated to the topic of Smart Systems as part of SD in ICT. The topics were extracted from the titles and abstracts from the remaining 165 papers. To enhance the ability of the LDA algorithm to extract topics about Smart Systems, the terms “smart” and “city” were removed from the processed text. Due to the smaller set of documents focussed on a single topic, the parameters were selected as five topics, min\_df as 0.24, and max\_df as 0.85, which produced the lowest perplexity. The resulting topics are shown in Table 5.

Table 5: Smart System Related Topics with Word Clouds

Number	1	2	3
Name	Smart Applications	Citizen Quality of Life	Smart Energy
Wordcloud	 <p>challenge, quality, life, require, need, environment, application, urban, infrastructure, service, develop, concept, economic, solution, base, process, quality, service, urban, infrastructure</p>	 <p>quality, life, citizen, urban, concept, infrastructure, service, economic, life, concept, infrastructure, service, economic, life, concept, infrastructure</p>	 <p>energy, system, solution, problem, develop, concept, base, economic, develop, concept, base, energy, system, solution, problem, develop, concept, base</p>
Number	4	5	
Name	Environmental System	Urban Solutions	
Wordcloud	 <p>environmental, system, process, challenge, economic, social, require, urban, service, need, solution, base, problem, develop</p>	 <p>urban, approach, need, base, require, economic, concept, solution, social, application, service, challenge, problem, develop, process, approach, need, base, require, economic</p>	

The publication trends of papers within the topic of Smart Systems as well as the allocation of papers to the new sub-topics are shown in Figure 10. It seems that the research within the context of ICT for SD on the topic of Smart Systems focusses on the quality of city life and to a lesser degree, solutions for the urban setting. Research into Smart Systems is also addressing energy, the environment as well as its applications.

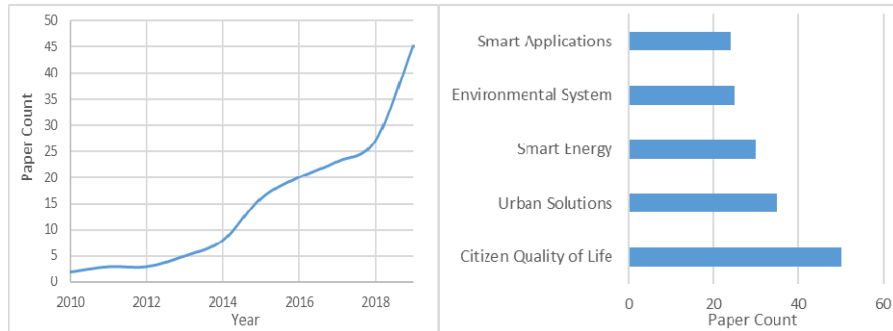


Figure 10: Smart System Publications

The trends for the research focus into the different topics, as explained in section 3.3, are shown in Figure 11. Despite being the topic with the highest total of papers allocated, Citizen Quality of Life is experiencing a decline in recent history. This is in contrast with Urban Solutions that present a steady growth over the analysed period.

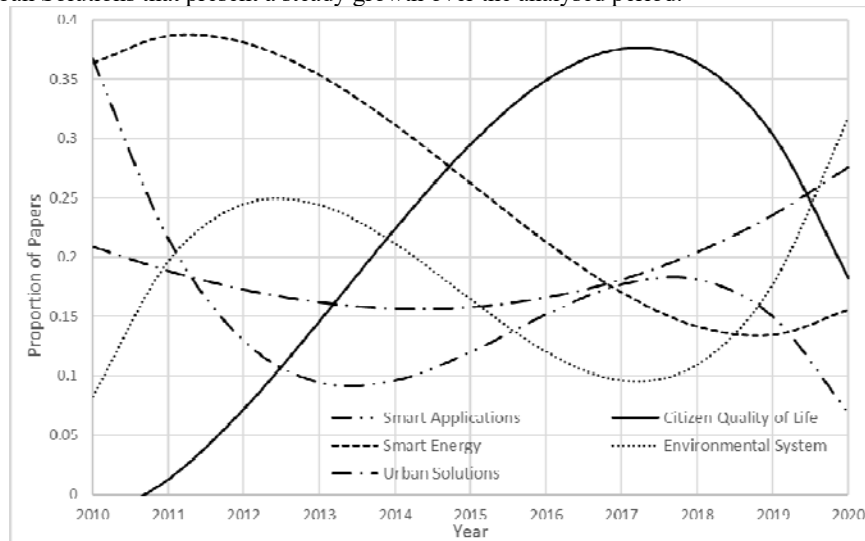


Figure 11: Smart System Topic Trends

Of late, the effect of Smart Systems in the environment is showing growing interest from researchers. Although Smart Energy and Smart Application dominated the early stages of the analysis period, they are falling in terms of researcher focus towards the end of the analysis period.

## 5 Discussion and Conclusions

This paper performed a bibliometric analysis to extract the research topics from papers with ICT and SD in their titles, abstracts or author keywords. The search to capture these

papers delivered more than 2000 papers for analysis from the Scopus library. This number is difficult to manually process effectively, therefore machine-learning NLP and topic modelling algorithms were implemented as a novel contribution to the field of research on ICT in SD. The process described in section 3 captures and prepares the text data to be processed and analysed gain an understanding of the research landscape.

Key to this process is preparing and cleaning the raw text data before processing. It is important to remove the search terms and other academic writing terms as they do not contribute to extracting the underlying research topics. Another crucial aspect is selecting the most suited values for the vectorisation and topic modelling algorithms. The set was selected by iteratively calculating the perplexity of various combinations of the parameter values as well as implementing heuristics about the preferred number of topics and describing terms for the topics.

A balance needs to be struck between too few or too many topics. A large number of topics is difficult to uniquely identify while a small number may miss some of the hidden or underlying topics. To assist in the manual identification of the topics, word clouds were used to support the cognitive process of assessing the relative importance of the topic terms. Plotting the topic publication counts and trends over time highlighted some interesting patterns in the research field of ICT for SD. The understanding gained from the total number of publications allocated to a topic and the temporal trends of the topic relative to other topics often tended to differ greatly. Both these views are required for a proper understanding of the characteristics of the topics in ICT and SD.

From the bibliometric analysis of the papers, the topics of Knowledge Management, Design Process, Social Change, Development of Countries, Smart Systems, and Economic Impact are the leading topics of research into ICT for SD. However, of these, only Smart systems and Development of Countries exhibit a strong increasing trend. This indicates that research tends to focus on smart system technologies to have an impact on the economy and development within different countries.

The increasing popularity of research into Smart Systems warranted a deeper look into the bibliometric data of this specific topic. Therefore, the papers allocated to the topic were further analysed using the same topic modelling based bibliometric process. However, due to the smaller sample of papers, the number of topics to be reliably extracted is reduced. The output of this second phase of analysis highlighted the topics of smart systems solutions for urban societies and the environmental aspects as popular research areas. The research topics of smart systems include quality of citizen life, solutions for the urban setting, energy, and the environment. Of these, it seems Urban Solutions increasing in popularity. The role of Smart Systems in the environment is also gaining prominence.

This paper demonstrated the utility of machine-learning-based topic modelling to support the planning and execution of research. Topic modelling presents a paper and topic search method that is far richer than a plain search using keywords. It can extract hidden and underlying topics from text. This paper also demonstrated that the bibliometric analysis process is scalable and can be applied in a cascading approach. The key is to start wide and focus with iterative steps until the sample of papers is narrowed down to a number that can be manually processed in detail.

## 6 References

- Agrawal, A., Fu, W. and Menzies, T. (2018) 'What is wrong with topic modeling? And how to fix it using search-based software engineering', *Information and Software Technology*. Elsevier, 98(February), pp. 74–88. doi: 10.1016/j.infsof.2018.02.005.
- Antons, D., Kleer, R. and Salge, T. O. (2016) 'Mapping the Topic Landscape of JPIM, 1984–2013: In Search of Hidden Structures and Development Trajectories', *Journal of Product Innovation Management*, 33(6), pp. 726–749. doi: 10.1111/jpim.12300.
- Bashri, M. F. A. and Kusumaningrum, R. (2017) 'Sentiment analysis using Latent Dirichlet Allocation and topic polarity wordcloud visualization', in *Fifth International Conference on Information and Communication Technology (ICoICT) Sentiment*. IEEE.org, pp. 1–5. Available at: [https://ieeexplore.ieee.org/abstract/document/8074651/?casa\\_token=53pCkXRfnuoAAA-AA:hKgXQTceapOd76\\_m3rZwwdy6RfbPBu2sc\\_4StDtMrMX8XuwEWC6iYz\\_94vR9BtakVMm7dkyMELuuNQ](https://ieeexplore.ieee.org/abstract/document/8074651/?casa_token=53pCkXRfnuoAAA-AA:hKgXQTceapOd76_m3rZwwdy6RfbPBu2sc_4StDtMrMX8XuwEWC6iYz_94vR9BtakVMm7dkyMELuuNQ) (Accessed: 17 May 2020).
- Blei, D., Carin, L. and Dunson, D. (2010) 'Probabilistic Topic Models', *IEEE Signal Processing Magazine*, 27(6), pp. 55–65. doi: 10.1109/MSP.2010.938079.
- Credé, A. and Mansell, R. E. (2014) *Knowledge societies--in a nutshell: information technology for sustainable development*. IDRC.
- Eker, S. et al. (2019) 'Model validation: A bibliometric analysis of the literature', *Environmental Modelling and Software*, 117(March), pp. 43–54. doi: 10.1016/j.envsoft.2019.03.009.
- Hassan, S., Haddawy, P. and Zhu, J. (2014) 'A bibliometric study of the world's research activity in sustainable development and its sub-areas using', *Scientometrics*, 99, pp. 549–579. doi: 10.1007/s11192-013-1193-3.
- Jia, Y. et al. (2018) 'Trends and characteristics of global medical informatics conferences from 2007 to 2017: A bibliometric comparison of conference publications from Chinese, American, European and the Global Conferences', *Computer Methods and Programs in Biomedicine*. Elsevier B.V., 166, pp. 19–32. doi: 10.1016/j.cmpb.2018.08.017.
- Jiang, H., Qiang, M. and Lin, P. (2016) 'A topic modeling based bibliometric exploration of hydropower research', *Renewable and Sustainable Energy Reviews*. Elsevier, 57, pp. 226–237. doi: 10.1016/j.rser.2015.12.194.
- Jie, L. I. et al. (2014) 'Bibliometric mapping of “ International Symposium on Safety Science and Technology ( 1998- 2012 )”', in *International Symposium on Safety Science and Technology Bibliometric*. Elsevier B.V., pp. 70–79. doi: 10.1016/j.proeng.2014.10.411.
- Kalantari, A. et al. (2017) 'A bibliometric approach to tracking big data research trends', *Journal of Big Data*. Springer International Publishing, 4(1), pp. 1–18. doi: 10.1186/s40537-017-0088-1.
- Keathley, H. et al. (2015) 'Bibliometric analysis of author collaboration in engineering management research', *International Annual Conference of the American Society for Engineering Management 2015, ASEM 2015*, pp. 679–689.

- Kunc, M. and Mortenson, M. and Vidgen, R. (2012) 'A computational literature review of the field of System Dynamics from 1974 to 2017', *Journal of Simulation*, 12(2), pp. 115–127. Available at: [papers3://publication/uuid/07C88B45-D01F-4FC3-AC91-3A9085095100](https://papers3://publication/uuid/07C88B45-D01F-4FC3-AC91-3A9085095100).
- Lamba, M. and Madhusudhan, M. (2019) 'Mapping of topics in DESIDOC Journal of Library and Information Technology, India: a study', *Scientometrics*. Springer Netherlands, 120(2), pp. 477–505. doi: 10.1007/s11192-019-03137-5.
- Lee, H. and Kang, P. (2018) 'Identifying core topics in technology and innovation management studies: a topic model approach', *Journal of Technology Transfer*. Springer US, 43(5), pp. 1291–1317. doi: 10.1007/s10961-017-9561-4.
- Lin, J. R. et al. (2016) 'A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM', *Computer-Aided Civil and Infrastructure Engineering*, 31(1), pp. 18–33. doi: 10.1111/mice.12151.
- Ma, T. et al. (2018) 'Topic based research competitiveness evaluation', *Scientometrics*. Springer Netherlands, 117(2), pp. 789–803. doi: 10.1007/s11192-018-2891-7.
- Mohamed, Mirghani, Murray, A. and Mohamed, Mona (2010) 'The role of information and communication technology (ICT) in mobilization of sustainable development knowledge: A quantitative evaluation', *Journal of Knowledge Management*. Emerald Group Publishing Limited, 14(5), pp. 744–758. doi: 10.1108/13673271011074872.
- Patel, F. N. and Soni, N. R. (2012) 'Text mining: A Brief survey', *International Journal of Advanced Computer Research*, 2(6), pp. 243–248. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.9006&rep=rep1&type=pdf> (Accessed: 16 May 2020).
- Punia, Y. (2016) 'Information and Communication Technology for Sustainable Development', *Voice of Research*, 5(1), pp. 2277–7733. doi: 10.4018/978-1-60566-026-4.ch306.
- Van Raan, A. (2003) 'The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments', *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, 12(1), pp. 20–29. doi: 10.14512/tatup.12.1.20.
- Rose, M. E. and Kitchin, J. R. (2019) 'bibliometrics: Scriptable bibliometrics using a Python interface to Scopus', *SoftwareX*. Elsevier B.V., 10, pp. 1–6. doi: 10.1016/j.softx.2019.100263.
- SDG (2020) SDGs .. Sustainable Development Knowledge Platform, un.org. Available at: <https://sustainabledevelopment.un.org/sdgs> (Accessed: 24 May 2020).
- Suominen, Arho;Toivanen, H. (2016) 'Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification', *Journal of the Association for Information Science and Technology*, 67(10), pp. 2464–2476. doi: 10.1002/asi.
- Tallón-ballesteros, J. D. A. J. and Hutchison, D. (2019) 'Studying the Evolution of the 'Circular Economy' Concept Using Topic Modelling', in *International Conference on*

Intelligent Data Engineering and Automated Learning. Springer International Publishing, pp. 259–270. doi: 10.1007/978-3-030-33617-2.

Tong, Z. and Zhang, H. (2016) ‘A Text Mining Research Based on LDA Topic Modelling’, in Proceedings of the Sixth International Conference on Computer Science, Engineering and Information Technology (CCSEIT), pp. 21–22. doi: 10.5121/csit.2016.60616.

UN (2015) Transforming our world: the 2030 Agenda for Sustainable Development, United Nations, General Assembly. doi: 10.1163/157180910X12665776638740.

Valerdi, R. and Davidz, H. L. (2009) ‘Empirical research in systems engineering: Challenges and opportunities of a new frontier’, *Systems Engineering*, 12(2), pp. 169–181. doi: 10.1002/sys.20117.

Wu, J. et al. (2018) ‘Information and communications technologies for sustainable development goals: State-of-the-art, needs and perspectives’, *IEEE Communications Surveys and Tutorials*. IEEE, 20(3), pp. 2389–2406. doi: 10.1109/COMST.2018.2812301.

Gatautis, R., Medziausiene, A., Tarute, A., & Vaiciukynaite, E. (2015). Towards ICT impact framework: Private and public sectors perspective. *Journal of Economics, Business and Management*, 3(4), 465-469.

Sanjeev, R., & Natrajan, N. S. (2019). Role of blended learning environment towards student performance in higher education: mediating effect of student engagement. *International Journal of Learning and Change*, 11(2), 95-110.

Ianioglo, A., & Polajeva, T. (2017). The essence and phases of the comprehensive system of ensuring the economic security of enterprise. *International journal of learning and change*, 9(1), 59-74.