# A text mining analysis of proceedings of the New Zealand Agronomy Society 1971-2017 using Natural Language Processing

J. Liu, A. Hunt and L. Jesson
New Zealand Institute for Plant & Food Research Ltd, Private Bag 1401, Havelock North, NZ
Jian.Liu@plantandfood.co.nz

## Abstract

There has been a long history of agronomic research in New Zealand, some of which has been summarised through the 47 years of Agronomy New Zealand, the proceedings of the New Zealand Agronomy Society (ASNZ). Over this time, agronomic research has likely changed substantially. Text mining provides an opportunity to systematically interrogate these proceedings to reveal some of the trends that have emerged. For papers published from 1971-2017 we asked what proportion of papers refer to the key statistical methods ANOVA and Regression, what was the representation of research institutes among submitting authors, what two-word combinations were most often associated, how did the type of crop and change by decade and could we distinguish decades by the frequency of use of particular words? The application of statistical methods, such as ANOVA, increased from 5.4% in 1970s to 61% proportionally in present decade. The number of papers published each year decreased from a high of 38 to a low of 6. Researchers from Department of Scientific and Industrial Research (DSIR) contributed the most of papers before it divided into Crown Research Institutes (CRI) in 1992; since then Lincoln University, New Zealand Institute for Crop and Food Research and Massey University have presented the most papers, Wheat, maize and pastures remained the three main crop species reported over all decades, however decades showed different trends in combinations of words used. Over 10% of the words used in the 2010s differed from those in 1970s. These results show that text mining methods are a powerful tool to gain understanding form large literature databases such as proceedings and can be used to gain insights of how the methods and focus of a scientific disciplines changes over time.

***Additional keywords:*** text similarity, R software, key trends, two-word association and word clouds

## Introduction

The forward to the first published set of agronomy proceedings in 1971 outlines a need for agronomic research as farmers increasingly moving from pastoral to crop production and the urgent need for the formation of the Agronomy Society to provide a means of consultation amongst scientists (Lynch, 1971). Over the 47 years since, 898 articles have communicated the work of more than 1300 scientists. Conservatively this represents hundreds of thousands of hours of research and millions of $NZ in agricultural R&D investment. It is therefore worth pausing to review some of the trends that have emerged from this considerable investment.

Much has changed in New Zealand crop production over the past 47 years including the development of new cultivars, advances in machinery, changes in the type and price of available agrichemical inputs and fertilisers, increased emphasis on environmental footprint and diversification and deregulation through the removal of farm subsidies (Catriona *et al.* 2006).

Recent advances in the field of text mining now make it feasible to systematically evaluate every word in a large body of digitised documents and extract terms on the basis of frequency and association with other terms (Rani *et al*. 2014; Günther and Quandt, 2016; Silge and Robinson, 2016; Salloum *et al*. 2017). Many of the text mining packages are also available on open-source platforms such as R and Python, enabling a global community to reproduce and adapt the packages for their own purposes (Bird *et al*. 2009, Munzert *et al*. 2014).

Here we present the some of the trends extracted from the last 47 years of published proceedings of the New Zealand Agronomy Society. While the questions that can be extracted from the analysis of text are numerous, we have chosen here to focus on five aspects:

1. Key statistical methods to understand changes in the use of statistical methods.
2. Institutional representation to understand the breadth of organisations contributing knowledge to ASNZ.
3. The frequency of combinations of words (for example two words that are often written together) to reveal potential research topic changes.
4. Crop species that are reported on in the titles of papers to ask how main research crops in New Zealand have changed.
5. Text dissimilarity for evaluation of word usage over time to help understand research area changes.

## Methods

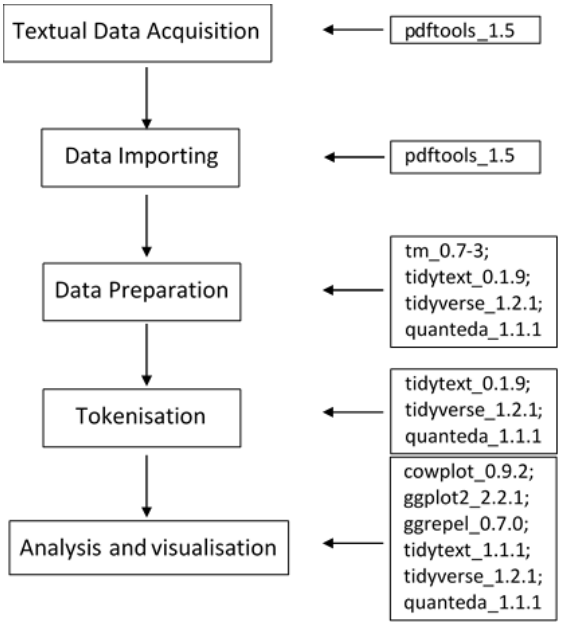
There are three key approaches for text mining: rule-based feature extraction and supervised and unsupervised machine learning (Günther and Quandt, 2016; Welbers *et al*. 2017). The first approach was used for the purpose of complexity reduction - i.e., a user can retrieve information from a text collection with the help of regular expressions based on pre-set rules. In contrast machine learning approaches can classify information and predict rules based on associations within the text. We focused here on rule-based feature extraction for simplicity. We made the key assumption that words changed across decades – but it is important to note that the choice of decade is an arbitrary metric. Hence, we pre-defined 5 periods: 1971 to 1979 as 1970s, 1980 to 1989 as 1980s, 1990 to 1999 as 1990s, 2000 to 2009 as 2000s and 2010 to 2017 as 2010s.

The flowchart below illustrates a workflow of textual data analysis and R packages which have been used in present study.

### Data acquisition and importing

We successfully downloaded 891 articles from Agronomy Society New Zealand website for text mining. The data acquisition process was carried on the R platform (R 3.4.4 Core Team 2018) with the R package pdf tools (version 1.5 Ooms, 2017). The package read PDF format papers into plain texts which then were stored in a list linked to the published year and unique identification number. Seven papers could

not be downloaded. However, we were able to extract the titles and authors of those papers from the website resulting in a complete dataset of titles, authors, published year and unique ID number of 898 papers.




**Flowchart** of text mining and packages applied in each step.

### Data preparation

We used three R packages: tm (version 0.7-3), tidytext (version 0.1.9) and quanteda (version 1.1.1), to prepare the text data (Feinerer and Hornik, 2017; Silge and Robinson, 2016; Benoit, 2018). The preparation consists of four common practices:

1. Lowercase all text;
2. Remove numbers;
3. Remove punctuation;
4. Remove stopwords (i.e. words that don't contribute meaning such as "the" and "when").

We used additional stopwords (see full list on github) to reflect words specific to agricultural science publications such as kg/ha and g/m$^2$, which were frequently used in the papers. A crop name dictionary (see full version on github) was created to extract crop types. The dictionary is developed from the crop name list from FAO website. In addition, the statistical method, organisation names, abstracts and keywords were extracted by applying regular expression on preliminary processed text data. For example, the regular expression that we used for ANOVA was "(anova)|(analysis\\sof\\svariance)".

### Tokenisation

Tokenisation is the key part of text mining, which converts the unstructured text into structured data (Silge and Robinson, 2016; Welbers *et al*. 2017). Tokenisation describes the process of extracting meaningful strings from text data (Bitam and Mellouk, 2008). Tokens can be words, phrases, sentences, chapters or other combinations of characters and symbols depending on the purpose of analysis (Bilisoly, 2008). We used one-word tokens in this study to identify crop type and analyse similarity between articles within and between decades. We used the "dfm" function from the R-Package quanteda to tokenise words by using its stem word feature. For topic extraction from paper titles we used the package tidytext to identify two word association tokens.

### Analysis

To assess how key statistical methods or terms may have changed over decades we analysed the relative frequency of terms to eliminate the effects of unequal elements such as number of paper each year/decade and number of words in each paper. In particular we examined changes in the frequency of words related to key statistical

methods, and institutional representation in each decade by detecting the presence of the term in the whole paper. To investigate trends in crop type we examined frequency of a crop in the title of an article. For text similarity, whole papers in each period were tokenised and were the unit for analysis.

To examine how combinations of words changed or remained the same across the decades we used similarity analyses to group papers by decade. We used cosine measure for text similarity as Strehl et al. (2000) and Huang (2008) found that cosine measure performed better results of similarity text (p-value < 0.05) compared to the other common Euclidean Metric measure regarding to human language analysis. We used dendrograms to visualise the dissimilarity (1-similarity) matrices: decades grouped together have more similar combinations of words than decades found further apart.

## Results

### Statistical method

A total of 222 or 24.7% of papers mentioned ANOVA. The overall number of mentions of the term ANOVA increased in the last 47 years from approximately 2% in 1971 to 80% of papers mentioning the use of ANOVA in 2015 (Figure 1).

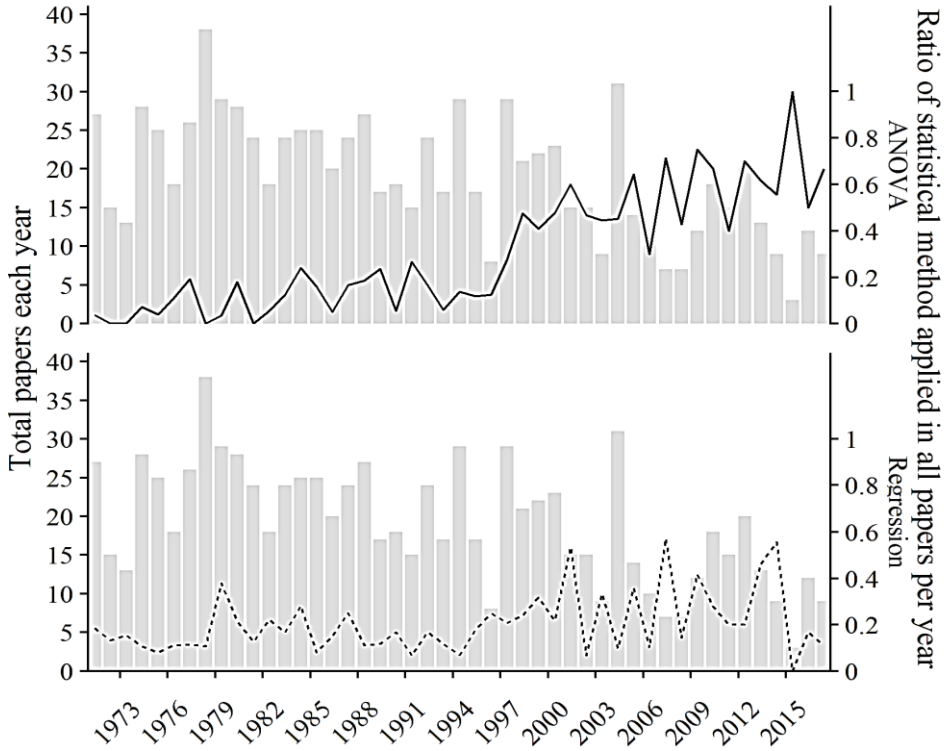



**Figure 1:** Total paper number (bar) in each year and mention of statistical method (line) in each year.

In the period from 1971 to 1993, the proportion of papers applying ANOVA each year varied from 0 – 22% with an average percentage of 11%, while total number of published papers in this period was 525. Since 1994, ANOVA was increasingly used in more papers. Only 14% of papers implemented ANOVA in 1994, and this

percentage increased to 50% by1998. Since 2000, approximately 56% of papers mention the word ANOVA.

Overall, 20% of all papers (898) used the term regression at least once (Figure 1). The proportion of papers that used regression was nearly double (28%) after 1994 than it was before 1994 (16%). There were 68 papers that used both ANOVA and regression. To explore if we were missing key statistical tests we sampled 562 papers that were not identified as using ANOVA or regression. Some (70 out of 562) used the term LSD and/or P-value without mentioning a statistical test. Therefore, it is likely that we missed papers that used statistical tests due to a lack of reporting of the statistical methods used in the paper. It is possible that the quality of reporting of statistical methods increased over time.

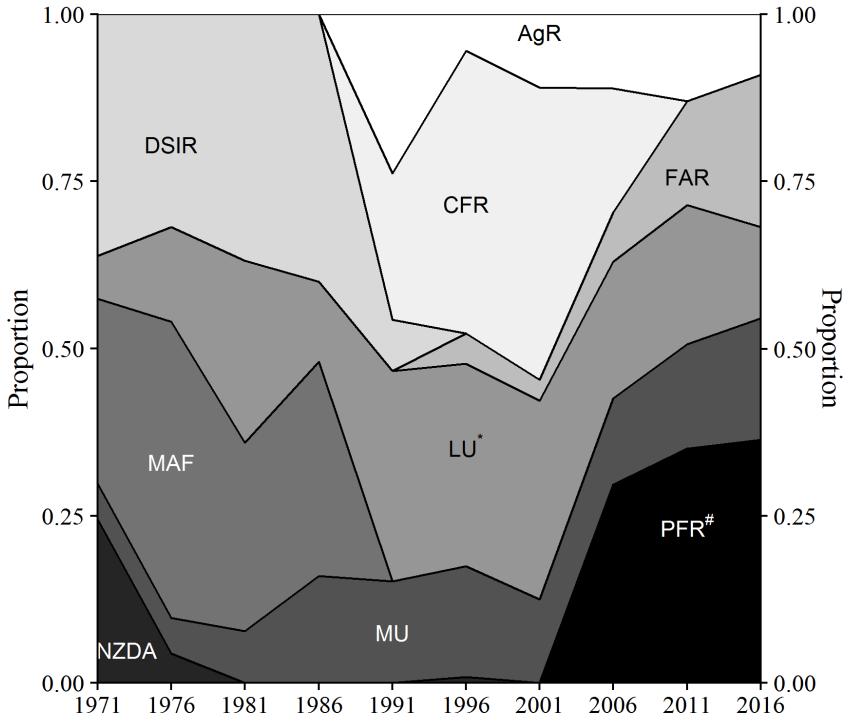



**Figure 2:** The proportion of papers published from top 10 organisations. It indicates the proportion of papers published by each institution for each in five-year interval. (A colour version can be found on the public github page). AgR, AgResearch; CFR, New Zealand Institute for Crop and Food Research; DSIR, Department of Scientific and Industrial Research (New Zealand); FAR, Foundation For Arable Research; LU, Lincoln University; MAF, Ministry of Agriculture and Forestry (New Zealand); MU, Massey University; PFR, New Zealand Institute for Plant and Food Research; NZDA, New Zealand Department of Agriculture. * In 1990, Lincoln University was formed from the previous Lincoln College. # In 2008, PFR was formed from the merger of CFR and HortResearch.

## Institutional representation

There were 57 institutions which had contributed papers in the New Zealand Agronomy Society (full list on github). Ten institutions contributed the majority of papers (801 or 89%). Changes in the proportion of papers published under each institution name changes are shown in Figure 2. It indicates the proportion of institution counts of total institution and paper counts in five year intervals.

Researchers from Department of Scientific and Industrial Research (DSIR) and Ministry of Agriculture and Forestry (MAF) published the majority of papers (69%) in 1970s and 1980s, followed by Lincoln College (LC) (15%), New Zealand Department of Agriculture (NZDA) and Massey University (MU). NZDA was the name of MAF pre-1970s, however, authors referred to NZDA solely till late 1970s.

In 1992, DSIR divided into Crown Research Institutes (CRI) which had AgResearch (AgR), New Zealand Institute for Crop and Food Research (CFR) and five others institutes (sciencenewzealand.org). AgR and CFR contributed approximately 50% of articles in 1992, while Lincoln University (LU) and MU contributed the other half. In the 1990s, CFR and LU were the two main contributors, 31.4% and 31.8%. Early in the 2000s, CFR was the largest contributor. The merger of CFR with Hort Research in 2008 into New Zealand Institute for Plant and Food Research (PFR) resulted in an average of 30% of papers published associated with this institution. LU and MU - contributed similar proportions of papers after 2006, 16.1% and 16.6% respectively. The proportion of papers from the Foundation for Arable Research (FAR), which was formed in 1995, increased from 4.4% to 18.9% in the 2010s.

## Two word associations

In the 1970s, the two-word combination of "Plant breeders" and "Lucerne" combined with its latin genus appeared in titles more than other word combinations, such as "pea cultivars" and "maize silage". In the 1980s, beet crop and nitrogen fertiliser started to draw researchers' attention along with white clover and peas (*Pisum sativum*). There was a wide variety of different word association in 1990s and 2000s, which suggests that research topics extended from crops themselves to the interaction of crops with the environment. White clover, maize/sweet corn and ryegrass were the main crops that were focused on in 1990s. In terms of agronomic research, the usage of dry matter accumulation with plant population, sowing method, nitrogen and water were commonly high in 1990s. Papers about seed production increased from 2000s to 2010s compared to the previous years. In 2010s, forage crops, such as fodder beet, were back after 20 years of lower occurrence. However, the diversity of expression in 2010s decreased. Overall, crops were addressed more using their common name in the 1970s than in the current decade where latin names of crops are often used. Dry matter has been a key element in the titles since 1990s. Research relative to ryegrass emerged from the 1990s and remained an important part of research until the present decade.


**Figure 3:** Word clouds of two-word associations that occurred over five times in paper titles by each decade. The size of words indicates the number of papers with these associations.

### Crop type

Figure 4 shows the proportion of crop species mentioned in titles over total paper number in each decade. Only the top five crops in each decade are presented here. Wheat, maize and pastures were the three most common crops mentioned by papers. Maize crop was predominant in 1970s. There was 15.1% of papers in the period of 1971 to 1979 that specified maize in the titles, which was twice that of wheat crop (7.3%), and triple Lucerne (5%). In the 1980s, the arable crops wheat and barley (20.2% combined) were the crop mentioned most followed by beet crop (5.6% including sugar and fodder beet) and Lucerne 5.2%. Vegetable crops such as peas and asparagus appeared equally. In the 1990s, pasture and arable crops were the main crop types and the proportion of papers for each crop type spread similarly with ranges between 6 and 7%. The other 93% of papers did not mention a crop name in the title. However, there were more papers that concentrated on maize (15.4% including sweet corn) and wheat (9.6%) in the 2000s. In the present decade, one third of total papers summarised research on ryegrass, cereals crops, potato and sugar or fodder beet.

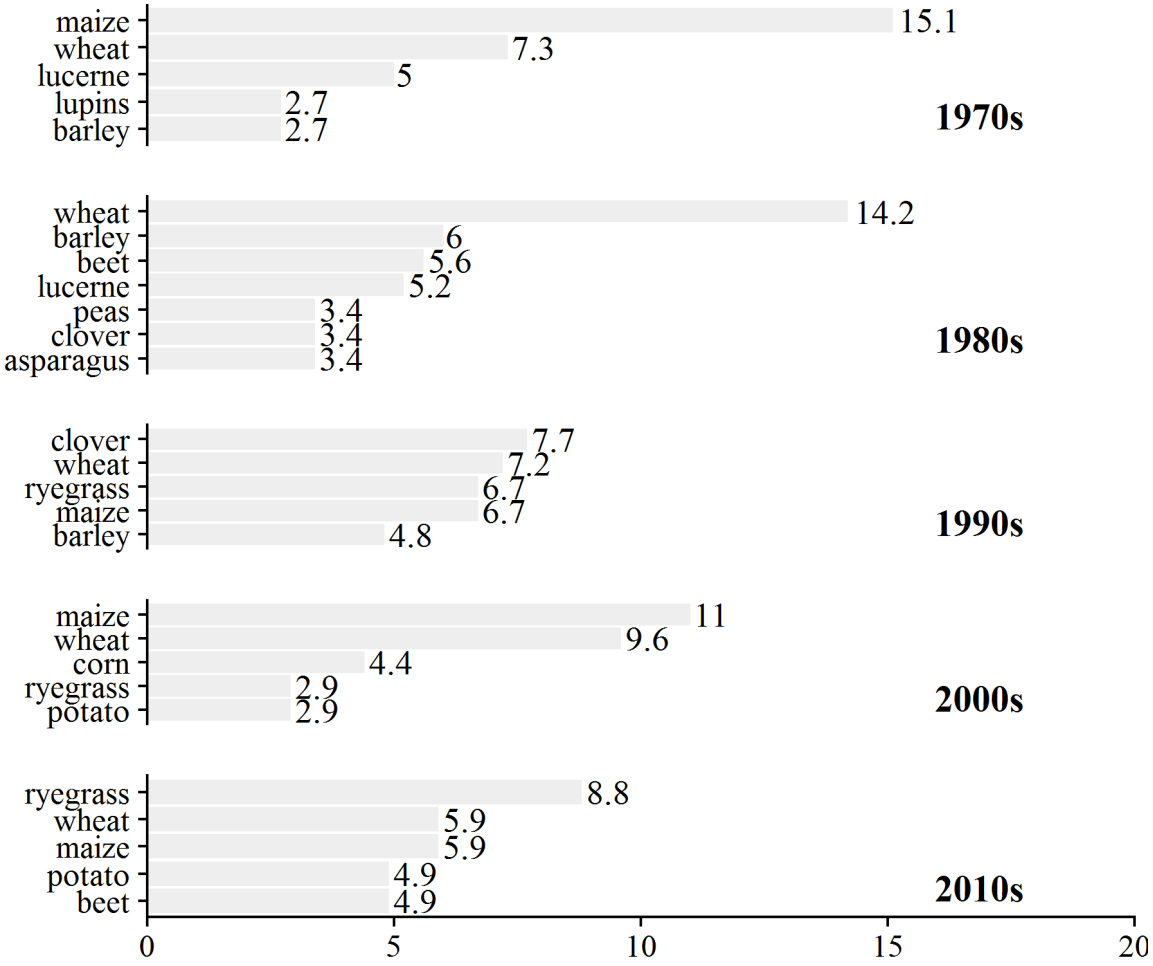



**Figure 4:** Proportion of crop types (top five) as mentioned in titles during each decade.

### Text dissimilarity

The similarity of word usage in different decades is visualised as a hierarchical clustering in Figure 5. Papers in the present decade had 10% of words that differed from words used in the 1970s. The contiguous decades show lower dissimilarity than the non-contiguous ones. For example, there was 4.6% dissimilarity of words used between the decades 2000 and 2010. The changes of the similarity of word usage likely reflect differences in research emphasis over time. Hence, the word cloud in Figure 6 provides the insights of dissimilarities among the five predefined periods. Words closest to each decade are words differentiate each decade from the others. These word clouds suggest that "maiz", the stem word for maize, "feed" and "variety" represent the words that differentiate the 1970s from other decades. "nitrogen" was distinguishing papers in the 1980s. seed and clover were the representative words for 1990s. In 2000s, there were fewer stem word compared to other decades, which likely indicates that papers in 2000s shared more common word vocabulary with other decades. The words or stem words "forag", "crop", "water", "treatment" and "agronomi" differentiated the present decade.

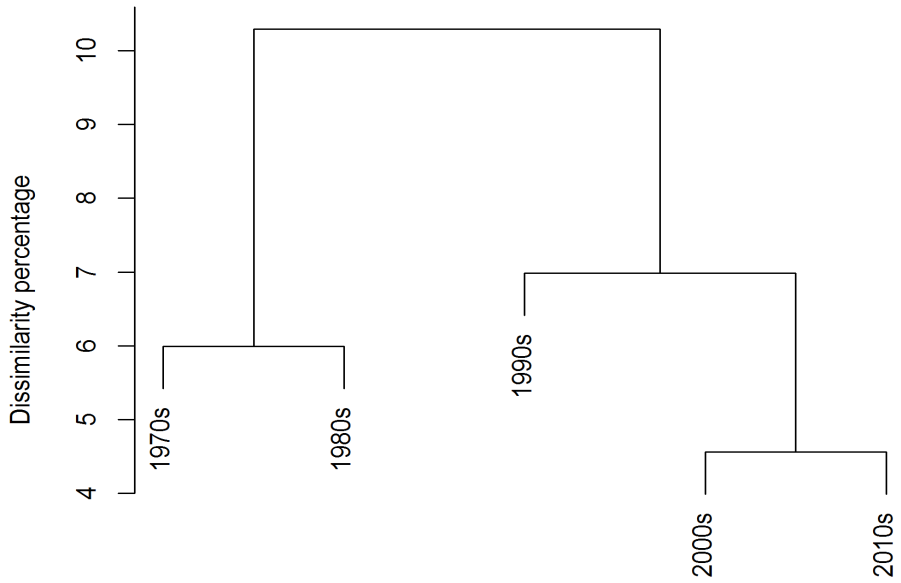



**Figure 5:** Dissimilarity (1-similarity) of word usage for all papers in each decade. Decades clustered together hare more similar than decades clustered further apart.

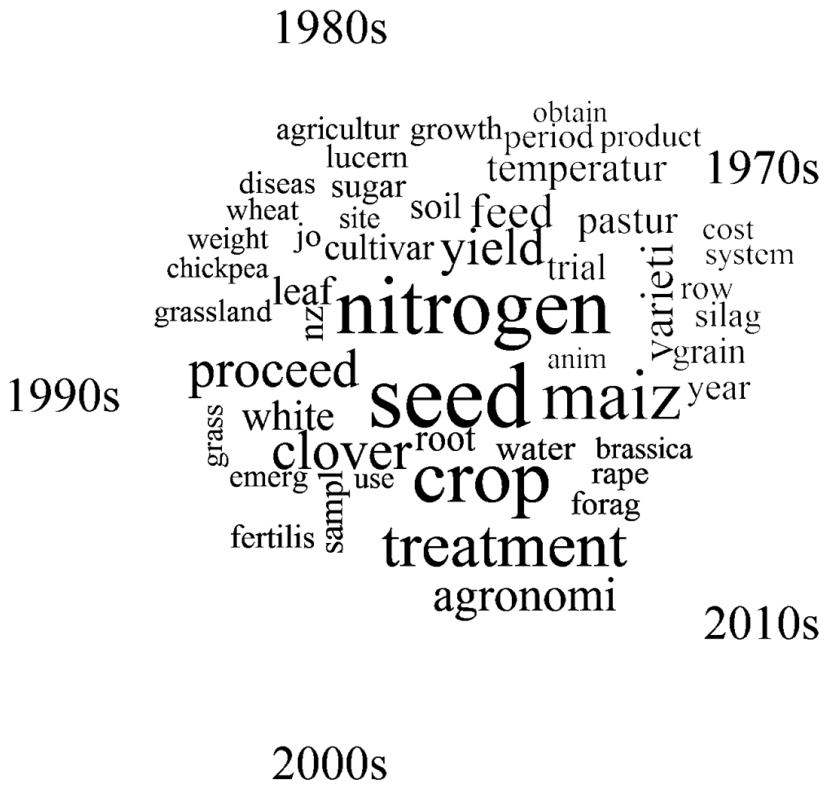



**Figure 6:** A word cloud representation of the 50 words in the papers included in each of the last five decades of the proceedings of the New Zealand Agronomy Society for the years 1971-2017.

## Discussion

Our text mining analysis has shown that the use of words combinations, analytical techniques and the institutional makeup of authors have changed dramatically in the proceedings of the Agronomy Society of New Zealand in the last 47 years. These likely reflect changes in the scientific emphasis of researchers as well as changes in funding of research for more different types of crops. These results show how text mining can be an important tool for researchers assessing not only the direction of research interest and perhaps could be investigated further to ask how funding and other research networks can influence peer reviewed published proceedings.

Changes in trends in analytical techniques are clearly shown using a text-mining approach. The frequency of the term ANOVA or analysis of variance shows that use of statistical tests is now common place in the majority of journal articles, and ANOVA is the most likely application. Researchers tended to use the full term expression "analysis of variation" before 1990s. The abbreviation of ANOVA replaced the full term completely during 1990s and afterwards. It is likely that after a few initial years researchers and the industry accepted the abbreviation ANOVA as the common expression of the method. The growing trend differences in frequencies between regression and ANOVA likely reflect changes from observational studies (1971 to mid-1990s) to more experimental study (late 1990s to present). This is also supported by the term "treatment" use after 2000 differentiating it from the two previous decades 1920-2000. However, further investigation may reveal alternative explanations. For example, the term linear model can apply to both continuous and categorical data, but was a word combination that was not highlighted in our two-word associations. In addition, while ANOVA is often used for experimental studies that apply a treatment to observational units, there is no guarantee that the terms were always applied in the same manner. Supervised machine learning approaches may prove informative for examining trends in different types of statistical tests that are finer grained than just ANOVA or regression.

The domination of maize-related papers in 1970s (Figure 4) resulted in that word being part of the key words that differentiates that decade (Figure 6.) However, maize was less frequent in the two word association results (Figure 3) than we expected. This is likely because maize was broadly associated with many different terms such as production or densities those associations occurred fewer than the appearance threshold of five times. A similar situation can be found in other decades (Figure 3). In 1980s, wheat was associated with the words quality and cultivars, and these association occurred less than white clover, even though wheat was the number one research crop type during the period. The words clover, maize and ryegrass were the three top crops but had very different term associations in two word association analysis in 1990s, 2000s and 2010s respectively. The emergence of these pasture relevant terms indicates that the expansion of dairy industry which demanded more research support regarding to seed production and agronomic method of growing pastures.

While pre-defined periods may not be able to capture the full potential of text mining on

the ASNZ papers, it nevertheless, seems adequate to demonstrate the intention of this study – to find informative trends over the history of 47 years of crop research. We cautiously draw three main conclusions:

1. Experimental studies or studies with treatments increased over the whole period.
2. Research crop types have changed from a dominance of research in arable and pasture crops to also include forage crops.
3. The papers published in different decades can be separated based on the use of words. In addition, more words contribute to the differences in the papers published in the later decades than earlier ones.

These results show just one aspect of the potential for text mining to understand research trends in science. Blei (2012) describes the future direction of text mining when one could easily extract the necessary information about a particular theme with various format data from the internet. There is a long way to go, however, as conventional domain searches take researchers more and more time to search key words and read abstracts and papers; techniques such as semi-automated text mining technique are likely to be increasingly employed for literature reviews and background searches. Further interesting avenues of research could ask how environmental issues, disease outbreaks and funding priorities change research directions.

## References


Benoit, K. 2018. quanteda: Quantitative Analysis of Textual Data. doi:10.5281/zenodo.1004683, R package version 0.99.22, http://quanteda.io.

Bird, S., Klein, E. and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. " O'Reilly Media, Inc."

Bilisoly, R. 2008. Text Patterns. In *Practical Text Mining with Perl*. R. Bilisoly (Ed.). doi:10.1002/9780470382868.ch2.

Bitam, S. and Mellouk, A. 2008. "The Term Vocabulary and Postings Lists." In *Introduction to Information Retrieval*, 19-48. Cambridge University Press.

Blei, D.M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4). ACM: 77-84.

Catriona, J. and MacLeod, H.M 2006. Intensification and diversification of New Zealand agriculture since 1960: An evaluation of current indicators of land use change. *Agriculture, Ecosystems & Environment* 115: 201-218, https://doi.org/10.1016/j.agee.2006.01.003.

Feinerer, I. and Hornik, K. 2017. tm: Text Mining Package. R package version 0.7-3. https://CRAN.R-project.org/package=tm

Francis, L. and Flynn, M. 2010. "Text Mining Handbook." In *Casualty Actuarial Society E-Forum*, 1.

Günther, E. and Quandt, T. 2016. "Word Counts and Topic Models." *Digital Journalism* 4 (1) Routledge: 75-88. doi:10.1080/21670811.2015.1093270.

Huang, A. 2008. "Similarity Measures for Text Document Clustering."

Lynch, P.B. 1971. “Foreword.” *Proceedings Agronomy Society of New Zealand* 1: a.

Meyer, D., Hornik, K. and Feinerer, I. 2008. “Text Mining Infrastructure in R.” *Journal of Statistical Software*. American Statistical Association.

Munzert, S., Rubba, C., Meißner, P. and Nyhuis, D. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons.

Ooms, J. 2017. “pdftools: Text Extraction, Rendering and Converting of PDF Documents.” R package version 1.5. https://CRAN.R-project.org/package=pdftools.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/

Rani, J., Ramachandran, S. and Rauf Shah, A. 2014. *An R Package for Text Mining of Pubmed Abstracts.*

Salloum, S.A., Al-Emran, M., Monem, A.A. and Shaalan, K. 2017. “A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives.” *Adv. Sci. Technol. Eng. Syst. J* 2 (1): 127-33.

Silge, J. and Robinson, D. 2016. “Tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” *The Journal of Open Source Software* 1 (3).

Slowikowski, K. 2017. *ggrepel: Repulsive Text and Label Geoms for 'ggplot2'*. R package version 0.7.0.https://CRAN.R-project.org/package=ggrepel.

Strehl, A., Ghosh, J. and Mooney, R. 2000. “Impact of Similarity Measures on Web-Page Clustering.” In *Workshop on Artificial Intelligence for Web Search (Aaai 2000)*. Pp. 58:64.

Welbers, K., Van Atteveldt, W. and Benoit, K. 2017. “Text Analysis in R.” *Communication Methods and Measures* 11 (4). Taylor & Francis: 245-65.

Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wilke, C.O. 2017. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 0.9.2. https://CRAN.R-project.org/package=cowplot.