Section 2

2.1 Introduction

Within this section three visualisation's will be presented and discussed to identify the patterns and trends present within the paper citations. The questions of what trends are present, and why these trends are present will be addressed. It is important to note, many of the papers reported Na for the citations for each year. These will not be treated as missing values but as papers with no citations and as such their values were changed to '0'.

2.2 Visualisation 1

The first visualisation (**Figure 5** below) displays a line chart of the total citations for each year from 2017-2020 for each research paper topic. A line chart is used to visualise how quantitative values have changed over time for a specific categorical item (Kirk 2016). This allows the user to quickly understand the general trend present in the total citation gains per topic.

To obtain the total citations for each year, the last recorded observation of each year within the data (i.e. 14.12.2017, for 2017) were extracted for each year as the citations are accumulative. Presenting this information by year decreases the complexity of the plot, relative to a plot of all data points, which allows greater clarity and understanding of the general trend for the user but may hide patterns present at the monthly scale.

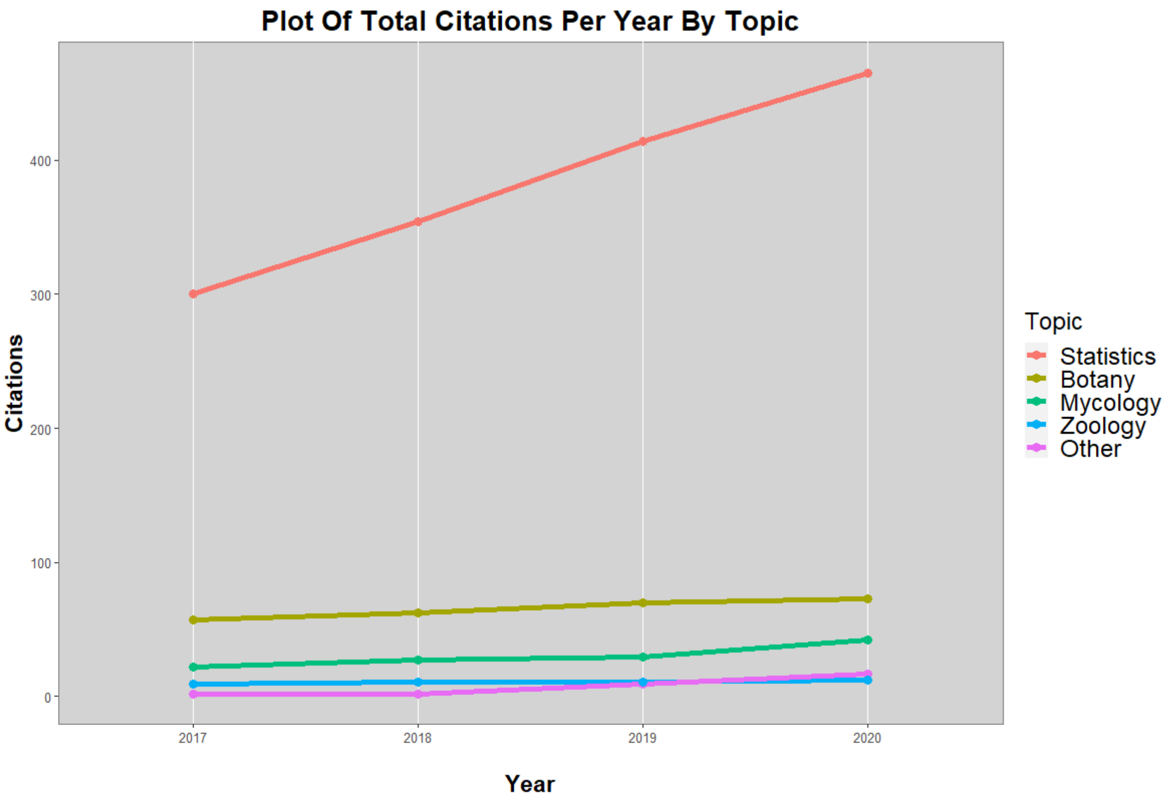
Additionally, each topic is identified by an individual colour which allows the user to make comparisons across the topics. A light grey background was chosen to emphasise these colours and a simplistic theme to remove visual noise (Tufte 1983).

From this plot it may be reasonable to suggest that papers of the "Statistics" topic showed the greatest growth from 2017-2020 and the highest total citations for each year. This suggests that this topic is the most popular in terms of citations. It can also be seen that the other topics present showed only very small growth in comparison.

However, this visualisation does not display the change in a form that allows for fair comparison across the topics. As can be seen the Statistics topic appeared to have the highest growth, however this topic also had significantly more citations at the start of the time period which and contains the most papers at 63, which will obscure the patterns present in the other topics.

In terms of principals of visualisation, the issues outlined above may indicate this plot has a high 'lie factor' as outlined by (Tufte 1983), as the size of the effect in the graphic may over emphasises the size of the effect shown in the data.

Figure 5. Line Chart



2.3 Visualisation 2

Visualisation 2 (**Figure 6**, below) attempts to overcome the limitations and issues present within the first visualisation. This visualisation presents the data in a new dimension and presents more levels of information to the user.

The new dimension presented is a way to standardise the citations to make meaningful comparisons across the different categories. This standardisation, which will be referred to as rate, was calculated by taking the difference between the last available citation and the first available citation and dividing this difference by the number of papers for that given category (01.12.2020 – 02.11.2017 / papers). This provides a number that represents the amount of citation gains per paper. This was to account for differences in the amount of papers produced between the different topics. In doing this the visualisation better represents the actual trends in the data as outlined by Tufte and allows for better understanding and addressing of the task, identifying ways to increase citation gains. by identifying the areas that have high rates we can identify areas to focus.

This visualisation also presents more levels of information about the data to the user. The previous visualisation was separated by topic only. However, this visualisation is separated by the topic of the research and the type of research, utilising a facet wrap. These are also referred to as ‘small multiples’ and are effective in inducing the user to make comparisons across categories (Tufte 1983). This presents information to the user which would otherwise be hidden, for example, it can be seen that for the types of research ‘Book Chapter’, ‘Conference Paper’, and ‘Working Paper’ the only topics of research were Statistics and

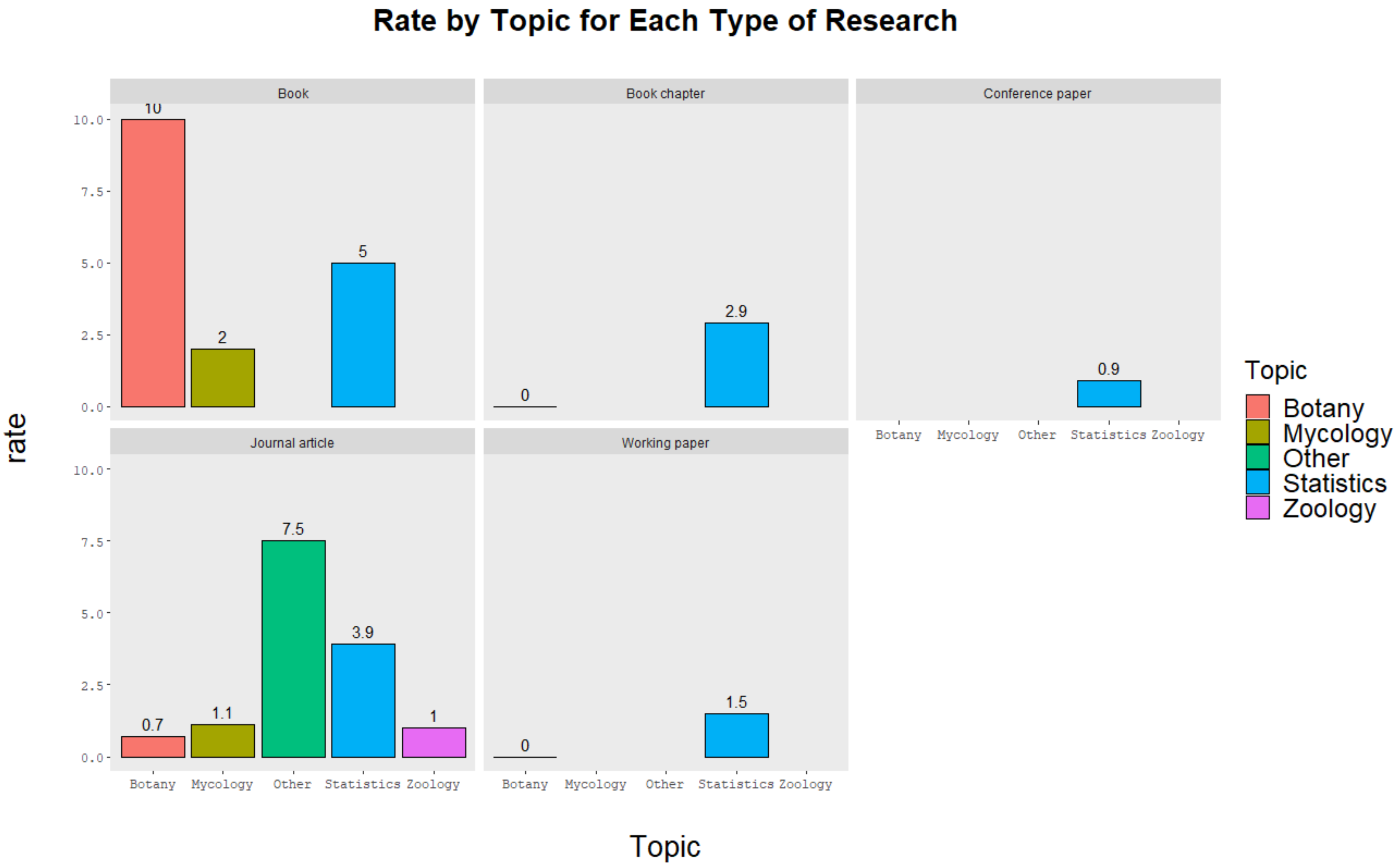
Botany (Book Chapter and Working Paper only). Additionally, within Book Chapter and Working Paper Botany showed zero citation gains. Whereas in the research type ‘Book’ the topic of botany showed the greatest rate.

The style of the visualisation is a simple bar chart, these types of charts are useful in comparisons and therefore fitting in comparisons of differences in rate across multiple categories (Kirk 2016). This was chosen over potential alternatives, such as the clustered bar chart, to maintain simplicity and clarity of the visualisation to aid user interpretation (Evergreen and Metzner 2013).

The conclusions that can be drawn from this visualisation are different from those of the first visualisation and emphasises the importance of analysing and exploring many dimensions and levels of data. These results suggest that for research of type Journal Article the topic of “other” showed the greatest rate of citation gain per paper followed by Statistics, the topic of Botany showed the lowest rate.

However, it is reasonable to suggest that one important question has not been answered by either of these visualisations, why do these patterns exist?

Figure 2. Bar Char of Citation Gain Per Paper for Each Topic and Research Type



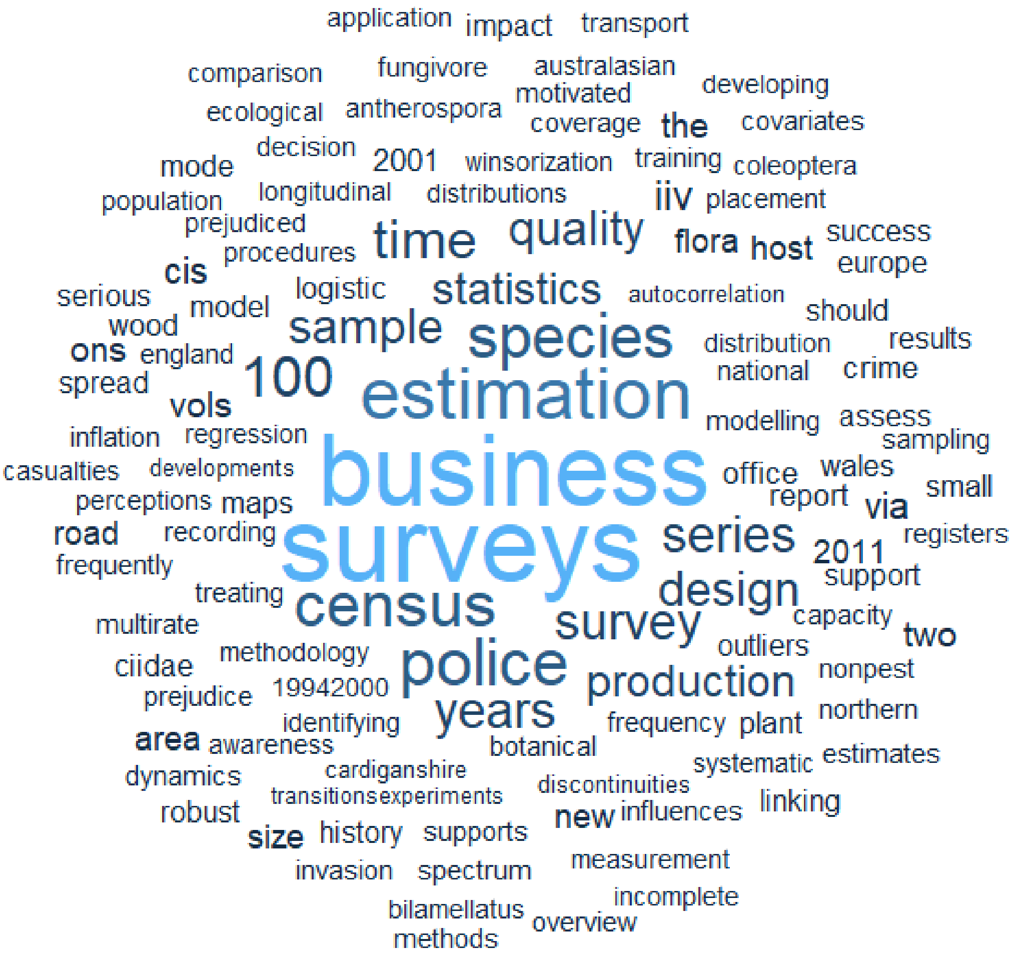
2.4 Visualisation 3

To answer the question of why these trends exist a word cloud was generated for the titles of the 20 highest cited papers. This was done to identify the themes of these papers to perhaps provide insight into why they performed well.

word clouds are visualisations for comparisons, showing the frequency of the individual words within a specified data set (Kirk 2016). Within this visualisation the size of the word indicates the frequency of the word's occurrence. This allows the user to consume a large amount of information in a relatively short period of time. This was done in reference to Tuftes elegant statement "Complex ideas communicated with clarity, precision and efficiency".

From this plot it is clear the most occurring words are business, surveys, estimation, and species and may indicate potential future research themes and areas to generate more citations.

Figure 3.


application
impact
transport
comparison
fungivore
australasian
motivated
developing
ecological
antherospora
coverage
the
covariates
mode
decision
winsorization
training
coleoptera
population
longitudinal
distributions
iiv
placement
prejudiced
procedures
time
quality
flora
host
success
europe
cis
serious
logistic
statistics
autocorrelation
should
wood
model
sample
species
distribution
results
ons
england
spread
estimation
national
crime
vols
modelling
assess
inflation
regression
sampling
casualties
developments
business
office
wales
perceptions
maps
report
small
via
road
recording
surveys
series
registers
frequently
support
treating
census
design
multirate
survey
capacity
two
outliers
ciidae
methodology
police
nonpest
prejudice
production
identifying
years
frequency
plant
northern
area
awareness
botanical
systematic
estimates
dynamics
cardiganshire
transitionsexperiments
discontinuities
robust
new
influences
linking
size
history
supports
invasion
spectrum
measurement
bilamellatus
incomplete
overview
methods