Master's thesis
Master's Programme in Data Science

Using Metadata to Analyze Trajectories of Finnish Newspapers

Zafar Hussain

May 18, 2020

Supervisor(s): Dr. Eetu Mäkelä

Examiner(s): Dr. Mark Granroth-Wilding

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Koulutusohjelma — Utbildningsprogram — Degree programme Master’s Programme in Data Science

Tekijä — Författare — Author

Zafar Hussain

Työn nimi — Arbetets titel — Title

Using Metadata to Analyze Trajectories of Finnish Newspapers

Työn laji — Arbetets art — Level Master’s thesis	Aika — Datum — Month and year May 18, 2020	Sivumäärä — Sidantal — Number of pages 99

Tiivistelmä — Referat — Abstract

The National Library of Finland has digitized newspapers starting from late eighteenth century. Digitized data of Finnish newspapers is a heterogeneous data set, which contains the content and metadata of historical newspapers. This research work is focused to study this rich materiality data to find the data-driven categorization of newspapers. Since the data is not known beforehand, the objective is to understand the development of newspapers and use statistical methods to analyze the fluctuations in the attributes of this metadata. An important aspect of this research work is to study the computational and statistical methods which can better express the complexity of Finnish historical newspaper metadata. Exploratory analyses are performed to get an understanding of the attributes and extract the patterns among them. To explicate the attributes’ dependencies on each other, Ordinary Least Squares and Linear Regression methods are applied. The results of these regression methods confirm the significant correlation between the attributes. To categorize the data, spectral and hierarchical clustering methods are studied for grouping the newspapers with similar attributes. The clustered data further helps in dividing and understanding the data over time and place. Decision trees are constructed to split the newspapers after attributes’ logical divisions. The results of Random Forest decision trees show the paths of development of the attributes. The goal of applying various methods is to get a comprehensive interpretation of the attributes’ development based on language, time, and place and evaluate the usefulness of these methods on the newspaper data. From the features’ perspective, area appears as the most imperative feature and from language based comparison Swedish newspapers are ahead of Finnish newspapers in adapting popular trends of the time. Dividing the newspaper publishing places into regions, small towns show more fluctuations in publishing trends, while from the perspective of time the second half of twentieth century has seen a large increase in newspapers and publishing trends. This research work coordinates information on regions, language, page size, density, and area of newspapers and offers robust statistical analysis of newspapers published in Finland.

ACM Computing Classification System (CCS):
General and reference → Document types → Surveys and overviews
Applied computing → Document management and text processing → Document management → Text editing

Avainsanat — Nyckelord — Keywords

Heterogeneous Data, Metadata Analyses, Exploratory Analyses, Newspaper Data, Finnish Newspapers

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1. Introduction

Recent technological advances in retrieval and utilization of information have enabled the research communities to develop the methods to get a better understanding of the complex data. The utilization of data is helping us in every aspect of our lives. All the technological changes occurring around us are primarily dependent on the data we are generating currently but there is a lot more historical data, which is being studied and still not fully rendered.

The primary source of the historical data is libraries. The historical data varies, from novels to legislative documents, letters to autobiographies, from artifacts to paintings. All these data sources are extremely beneficial when studying the sociocultural evolution of societies. One important aspect of the historical data is newspapers, as it also comes under the umbrella of big data research, because of the many efforts of libraries to digitize them. As a whole the newspapers data contains many facets, which can be studied thoroughly to paint a picture of the development of a society. One segment of the newspapers data is metadata. Generally, metadata serves various purposes for example, discovering and managing resources, facilitating interoperability and integrating resources, and making resources available for future use. In particular, newspapers' metadata is important as it keeps track of entire life-cycle of a newspaper, which can be useful to study the changes in newspaper publication against time. Newspaper life-cycle is a good example to understand the complexity of newspapers publication.

The National Library of Finland has digitized the newspapers from 1771 until 1929, with over 10 million pages. The newspapers until 1929 are publicly available. The digitized newspapers contain page images, which are accessible through any search engine. The page content data is available in Analyzed Layout and Text Object (ALTO) format. Since this data consists of content and metadata, the metadata is extracted separately [23] to study the material development of newspapers in Finland, which is the scope of this research work.[23][26]

Since the data is so enrich and expanded over a century, the questions to be answered can have a widespread range. To make this work compact and more comprehensive, some abstract questions are focused: 1) study the categorization of newspapers based on the metadata and observe how these categories fluctuate through time, place and language? 2)

How do newspapers in Finland developed? 3) What are the common paths of development for long running newspapers? 4) Study the dynamics of processes and find out different kinds of lives of newspapers, patterns in the development of newspapers? 5) Study and apply statistical methods to interpret the newspaper metadata and find the similarities among newspapers based on features?

As the data is challenging because it reflects complex, heterogeneous and dynamic phenomena, from a computational perspective it is not easy to weight features against other. Therefore, a sufficient part of this research work focused on reading the computational approaches to cluster and model dynamic and heterogeneous data in an interpretable manner, and on applying and evaluating the usefulness of these methods on the newspaper data.

This thesis is structured as follows. Chapter 2 gives a detailed description of data sets. The chapter also includes the explanation of additional feature extraction and data manipulation. Chapter 3 focuses on the literature review. In Chapter 4, the exploratory data analysis and regression models are discussed in detail. These models are applied on the dataset and the performance of these models is evaluated. Chapter 5 gives a brief explanation of clustering methods, application of clustering methods on this data and the results evaluation. Chapter 6 provides a comprehensive discussion of these results and Chapter 7 concludes this research work.

2. Data Sets

The National Library of Finland made the data publicly available [†] in ALTO file format. Various features such as page size, printed area, words and character count for each page have extracted by Eetu et al [23]. The measurement unit of quantitative features, such as page size, in the ALTO files is 10th of a millimeter. The data is from 15-01-1771 to 31-12-1917, containing over 2.5 million of records. Each newspaper is represented by an ISSN, which is a unique entity. Total number of newspapers are 424. With each ISSN, are attached issueIds, which represent newspaper published on a specific day. There are 592285 issueIds. Date column represents the day of publication and page indicates the page number. Against each page number are given width and height. Each page's total number of words and characters are also provided. Following is a snippet of data.

Table 2.1: Data set.

issueId	ISSN	date	page	width	height	words	chars
483775	1457-4756	1771-01-05	1	7.91	15.06	256	1195
483775	1457-4756	1771-01-05	2	7.91	15.06	181	861

With each ISSN, paper name is provided, but for this research work, only ISSN is considered throughout the analysis. Against each issueId and page number, total number of columns are also given as below. For an ISSN, language and place of publication are also known.

Table 2.2: Data Sets

issueId	page	wmodecols
100029	1	4
100029	2	4

ISSN	Kieli
fk100065	swe
fk10037	fin

ISSN	Kaupunki
0352-7502	Helsinki
0786-5511	Ilmajoki

Though there are newspapers published in different languages, e.g. Finnish, Swedish, German, Russian and English, but this research work focused on just two of them, Finnish

[†] https://digi.kansalliskirjasto.fi/search?formats=NEWSPAPER

and Swedish. Since the publication starting dates of newspapers are given, life of each newspapers is extracted by calculating the number of years a newspaper published. Minimum life is less than a year, where a newspaper published few issues and then stopped. Maximum life is 98 years, keeping in mind that this life is relevant to the data. The newspaper with life of 98 years, kept publishing but since it started in 1820 and our data ends at 1917, so its life is extracted as 98 years.

Out of 424 newspapers, there are 124 newspapers which published less than 3 years. Since most of our questions are directly or indirectly related to time analysis, so only those newspapers which lived for at least 3 years are selected. This gives us 300 newspapers, which are published in 54 cities.

Though the features are very straightforward, but there were some anomalies in the data, for which it has been wrangled for further analysis. As already mentioned, one of the feature is number of pages, there are instances when number of published pages were in hundreds, for example, on 22-12-1881, ISSN number 1458-8765 published 503 pages. Manually checking on the National Library of Finland's website [*], a book published/distributed by the newspaper along the issue has been found. To elude these kind of anomalies, only the issues with 16 or less number of pages are selected.

To increase the dimensionality of the data, new features are generated from the given ones. By multiplying height and width, a new feature 'area' is calculated. Since the number of characters per page are known, dividing them by area, a new feature 'density' is added to the data set. As each ISSN has assigned a date to each published issue, one possible feature is to calculate the publication period. A publication period represents the number of days an ISSN takes to publish a new issue. A value of 4 suggests that the newspaper is publishing every 4th day.

Since the data contains number of columns, height, width, area, number of words and characters of each page for every issue, any calculation on this data is a cumbersome, as the number of issues are 592285 and each issue can have a maximum of 16 pages. To make the data understandable and manageable, yearly averages of all the quantitative features are calculated which gives the following data.

Table 2.3: Yearly Averages Derived

ISSN	City	Life	Year	Pages	Area	Density	Char	Pub_Period	Cols	Lang
1457-4756	Turku	12	1771	8	118.81	9.10	1080	16	1	swe
1457-4756	Turku	12	1772	8	118.34	9.35	1104	7	1	swe

The features height and width are dropped from the data set, since the area feature is

[*] https://digi.kansalliskirjasto.fi/search?formats=NEWSPAPER

calculated by using them. Now the data shows yearly average values of all the quantitative features. City and Lang represents the place and language of publication respectively, whereas Life represents the total number of years a newspaper published. The yearly average data gives us 300 newspapers and 3906 number of records. Since the yearly averages of all ISSNs are calculated, so all the analysis will be based upon ISSN, and the attribute 'issueId' is not considered for analysis anymore.

Though the underlying structure of the newspapers may not be characterized well with the yearly averages, especially in case a newspaper changed some of its features for a short period of time in a year, but still these yearly averages keep the overall essence of the data and make it applicable for further analysis. This was a brief description of the data. Even though some more manipulation has been performed but they will be discussed in the relevant sections.

3. Literature Review

As discussed in some detail in the above section, the data is collected and mapped from different sources. One important aspect of this research work is to study computational methods and practices while working with heterogeneous data. Heterogeneity arises when the data is being generated or collected from multiple sources, with high variation in the type and format of data. If the data is coming from multiple geographic locations, there will always be correlation between time and space. Any historical or current data, will always suffer with these correlations and complexities. To understand the issues related with heterogeneity, a paper titled as *Heterogeneous Data and Big Data Analytics* [35] has been studied as a starting point.

To better understand heterogeneity, a starting point is to study the type of heterogeneity, which includes semantic, syntactic, terminological, and semiotic heterogeneity. Of these types, semantic and semiotic are vital as semantic represents the differences in modelling the same domain of interest and semiotic denotes different interpretation of the same domain by different users. The problem of semantic heterogeneity is crucial in handling a data set in any case but it exacerbates when dealing with the semi-structured data. Since the semi-structured data is collected from multiple sources, semantic heterogeneity exists in this data from the beginning and with the flexible schemas in semi-structured data, there can be more variations in the data. Generally semi-structured data comes up with the possibility of adding more features or deriving additional features from the existing ones. Number of additional attributes can increase to the level of impossible interpretation. Since semiotic heterogeneity is caused by the different interpretation of the data elements by different users, it is not easy for the computer to detect. [12] [35][17]

Another key element related to heterogeneous data is the data representation. The data can be in raw form collected with different data types, different measurement units and from different sources. Representation of the data also depends on the size of the data, and possibly generating more information from the given data. As discussed in the previous chapter, various additional attributes are generated from a single attribute in our data set. With the heterogeneity, comes many challenges to handle the data in a meaningful way. To solve these challenges, possible steps can be taken into account, such as data pre-processing, data mining, increasing or decreasing the dimensionality depending on the requirement and normalization of the data before modelling. These methods are applied on the data set (as explained in detail in chapter 2), so are discussed briefly below.

Data cleaning is the preliminary stage of unraveling heterogeneity. It is a process of identifying any incomplete or inaccurate data and modifying or removing the elements. In the data set discussed above there were some 'issueIds' with blank pages, so the number of words/characters for these pages are replaced by zero words/characters. Data preprocessing includes an important step of data integration in which data sets are mapped and merged based on shared attributes. As discussed in the earlier chapter, language, cities and columns are merged in the main data set based on the ISSN. For the geographical analysis, cities' coordinates are gathered separately and then mapped with the cities for various calculations such as measuring distances among the cities to locate metropolitan cities near small towns. Often integrating heterogeneous data is challenging in the absence of unique identifier, but in our data set, ISSN proves crucial to map and merge various data sets.

Generally when dealing with the heterogeneous data, dimensionality reduction is a useful practice, but depending on the domain, sometimes deriving extra features from the existing ones can add value to the data and make its interpretation simpler. Relating this argument to the data set under study, a new feature 'area' is derived from two features 'width' and 'height', which were not considered for any analysis afterwards. This new feature makes the data more meaningful for further analysis. Data integration is a vast research field, discussing tools and techniques of it is beyond the scope of this research work. These methods discussed above are the starting points while working with heterogeneous data but the computational approaches and machine learning methods differ as per domain requirements. Following are some approaches to deal with the heterogeneous data.

While studying the problems of heterogeneous data, clustering is one of the most likely solutions. In the paper *Clustering Heterogeneous Data Sets* [1], the authors have discussed various clustering techniques such as multi-view, ensemble, and collaborative clustering. Multi-view clustering is a supervised learning method, which divides the features into multiple views (subsets). Multi-view algorithms train two independent hypotheses with bootstrapping by providing each other with labels for the unlabeled data [3]. The objective of training algorithms is to maximize the agreement between the two independent hypotheses and optimally combine the multiple views. Utilizing the method that the clustering from one view should agree with clustering from another view, spectral clustering is exploited with multiple views [37]. This method assumes that the true underlying clustering would assign corresponding points in each view to the same cluster. First spectral clustering is performed on individual views to extract eigenvectors and then iteratively find the similarity matrix projection of one view against the eigenvector of second view and vice-versa. These projections are used to compute the Laplacian matrix and find the updated values in eigenvectors of the views. Once the discriminative values of these eigenvectors are obtained, the most informative views are selected and using k-means these eigenvectors are clustered. [1] [3] [37]

Another important approach for clustering heterogeneous data is Ensemble-based clustering. The basic concept of ensemble-based clustering is to combine multiple partitions of the data set into a single clustering solution [1]. Analyzing heterogeneous data, ensemble-clustering dedicates different clustering process to each domain and aggregates the results within an ensemble framework, which emphasized an agreement between the different domains. This is very useful approach when the format and type of data varies too much [37]. This method partly applied in the clustering of the data under study. Using the idea of ensemble-clustering, different clustering processes, such as Spectral and Hierarchical, are applied on a subset of features, and based on these clustering; results are interpreted in a more meaningful way.

Hierarchical clustering are a tree-like clustering algorithms, in which the data points are grouped using a top-down (divisive) or bottom-up (agglomerative) approaches [36]. Each data points in the bottom-up approach is a single cluster in itself in the start, and then by grouping the similar data points bigger clusters are generated. On the other hand the divisive approach works other way around by starting from one cluster which consists of all the data points and then dividing them into more clusters. An extension of ensemble-clustering, a Hierarchical Ensemble Clustering method is proposed in [36]. This approach takes both hierarchical clustering and partitional clustering and returns a hierarchical clustering based on consensus. Partitional clustering just decomposes the data into various disjoint clusters that represent a local optimum of some predefined objective function. This proposed method has three main variations. If the input is partitional clustering, aggregate consensus distance is constructed first and then a consensus clustering is generated. Using the consensus distance a structure hierarchy is further generated on top of the consensus clustering. In the case of hierarchical clustering as input, a Dendogram [28] is used to represent a hierarchical decomposition of the data set. A dendogram is a graphical representation of partitioned data. To characterize a corresponding dendogram, distance function describing the relative position of a pair of leaves is used, which can be view by dendogram descriptors. various dendogram descriptors are studied in [36], but the relevant to this research work is Cophenetic distance, which is explained in some detail in hierarchical clustering section. The third scenario, if the input is partitional clustering and hierarchical clustering, first a consensus distance from the partitional clustering and dendogram distance using hierarchical clustering is constructed. A hierarchical clustering is generated by combining these two distances. The generalizations of these approaches are discussed here, but in the coming sections these approaches completely or partly are implemented on our data set. [36]

While working with hierarchical clustering, the goal of combination schemes is to find a new dendogram, which is a representation of the whole dendogram set. For this purpose a metric in dendogram space is needed for a proper representation. Since working directly with a dendogram is difficult, a new algorithm called Min-trAnsiTive Combination of Hierarchical clusterings (MATCH) is proposed in [24]. The basic idea behind this algorithm is to use an intermediate matrix representation of dendogram to generate the consensus dendrogram. The MATCH algorithm use similarity matrices of hierarchical clustering as their descriptors in the first step and then the combination method aggregates the input matrices into a consensus matrix [24]. Next this consensus matrix is made transitive. Finally, by drawing α-cut of transitive consensus matrix hierarchical clustering is created. To evaluate this algorithm, experiments were conducted in two different sets. In the first experiment, the proposed algorithm is compared with hierarchical clustering without ensemble methods and then with ensemble methods. The results show ensemble methods are better than non-ensemble methods, whereas the algorithm proposed in [24] achieves highest accuracy when combined with Cophenetic distance method. [24]

One key aspect of this research work is clustering the heterogeneous data, which is the reason various clustering algorithms and approaches have been studied. Besides clustering other computational approaches have also been studied which are discussed next.

In the field of computational intelligence, current trends show that understanding of heterogeneous data is extremely important, since the explanation and interpretation of data depends on the understanding of the data. One possible way is to use logical rules to describe the complexities of data, for that purpose decision trees are the most appropriate approach. Since decision trees provide several set of rules, this makes the data explanation simpler and increases the accuracy of analysis. Decision trees can be augmented based on different criterion, one such criteria proposed in [11] is, Separability of Split Value(SSV). With this criteria, forests of heterogeneous tress can be built instead of a single tree. This method can be applied on both discrete and continuous data, defining split-off value differently for both cases. A real number is used as a cut-off value in continuous features, whereas for discrete feature a set of alternative values can be used. Any value which separates the largest number of objects from different classes is the best split value. This method has a basic assumption that each feature which has at least two different values, there exists a split. Decision trees are constructed by searching best splits. At each step after finding the split, if the data is not pure (belonging to more than one classes), the same procedure is applied to that subset until data is separated in distinct classes. This gives the maximal possible accuracy [11]. Since this method is creating logical rules at each step, there is a possibility of over-fitting and to avoid that, cross validation is performed to find the optimal pruning parameters for the tree. These separability measures can be applied on individual features and combinations of the features, though it will not increase the computational complexity of the analysis but it will make the problem more complex with high number of possible features [11] . This was the basic idea of SSV approach, which can be applied on subset of the data to generate forests of trees. This approach is tested on various data sets from health care domain. The results suggest that this approach has found many set of rules with very high accuracy and the highest results are obtained by cutting the space perpendicularly to the axes [11]. This method has the capability to discover simple, accurate and very sensitive description of the complex data. [11]

Another interesting hypothesis, 'Small Heterogeneous Is Better Than Large Homogeneous' [9] is presented in a Decision Tree Ensemble method. The authors proposed a new algorithm, Mean Margins Decision Tree (MMDT), which builds oblique decision boundaries. It used linear combination of attributes to define these boundaries. This approach is designed to be parameter-less and simple. At an abstract level, this method constructs the tree as any other decision tree, with the only difference that it chooses the decision boundaries in the form of linear combination of inputs. This decision boundary maximizes the margins between a subset of class True and a subset of class False for a binary classification problem. Since not all classification problems are binary, this approach consider each value as an orthogonal dimension in the case of nominal attributes. MMDT converts class labels to real vectors every time before choosing decision boundary. Mean and first principal component of this class vector is computed first and then each value is assigned either a true class or false class depending on the difference with this mean value and projection on the first principal component [9]. If the difference is negative, the set of values belong to true class otherwise to the positive class. This method is applied for different subsets of data to generate ensemble decision trees and compared with Random Decision Trees (RDT) and Entropy-Reducing Decision Trees (ERDT). Though the results show the effectiveness of random decision trees on 17 different data sets while the ERDT and MMDT have effectiveness on 13 and 12 different data sets respectively, but the intuitiveness and simplicity of MMDT makes it well-suited algorithm for ensemble methods. [9]

The algorithm Homogeneous data In Similar Size (HISS) [16] allows the users to select number of subsets, similar to bootstrapping, with a difference that the number of data points in all the subsets will be equal. This approach creates model for these homogeneous subsets and ensembles them for better accuracy [16]. To study heterogeneous data, merging data from different domains is one of the initial steps. For this purpose, a novel algorithm Multiple Kernel Preserving Embedding is proposed in [10].

This algorithm maps the objects from different domains into a unified embedding space by preserving within-domain similarities and cross-domain interactions. These similarities are approximated with Gaussian kernels. The experimental results of this model show wide applicability. [10]

Various methods have been studied for this research work but the focus is on clustering and decision trees approaches, as they are well suited to the data under study. Of these models, Spectral clustering, Hierarchical clustering and Random Forest Decision Trees are studied thoroughly and implemented with various sets of features later on.

4. Exploratory Analysis and Modelling

As the data is prepared for the analysis, exploring basic features such as publication period, pages, languages etc. is the first goal. The objective of the data exploration is to get a better understanding of the features and how these features developed over time.

4.1 Feature Exploration

Since the language is an important feature, which can divide the data into two categories, Finnish and Swedish, so this gives a good starting point. The number of newspapers for both these languages were increasing and decreasing (as the newspapers stopped publishing), analyzing this feature against time gives a better explanation. The following result shows that in the beginning, the number of newspapers publishing in Swedish language were high as compared to Finnish. Starting from 1771, the number of Swedish papers were increasing slowly, for next almost 100 years. The data shows that around 1880, number of Finnish language newspapers starts increasing significantly, while the number of Swedish newspapers keeps increasing gradually.

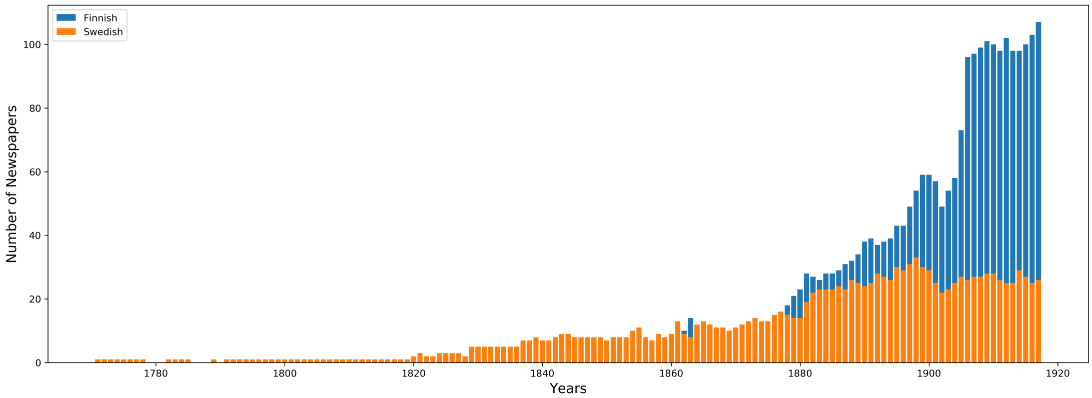


Figure 4.1: Distinct newspapers published in Swedish and in Finnish, 1771-1917

The above result explains the data through language feature, but to understand the development of quantitative features, publication period and number of pages are studied over time. As the data contains yearly averages, the unique publication period and percentage of newspapers publishing in that period are calculated, while dividing the data into six periods. The figure below shows the publication period over time. In the last quarter of eighteenth century, a publication period of seven days is dominant. The value ‘others’ represent the publication period of six, eight etc. In the first quarter of the nineteenth century, two and three days publication periods show significant increase while the ratio of seven days publication period drops 60%. It’s interesting to see that in the next quarter, the ratio of publication period of two days drops, while publication period four’s ratio increased, though three days publication period still dominates others. In this period, the ratio of one day publication period starts increasing, which is the popular trend for next 50 years. In the early years of twentieth century, one and two days publication periods are the popular trend.

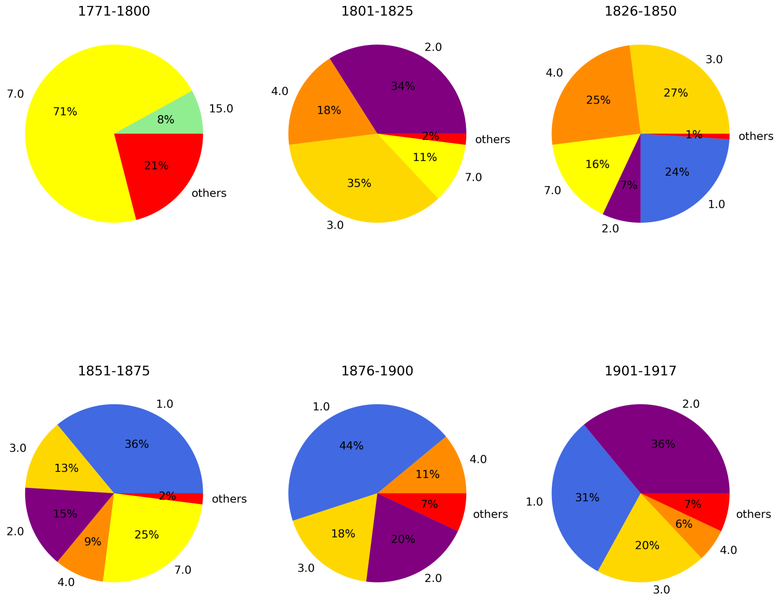


Figure 4.2: Common publication periods and percentage of newspapers following them over time

Another important feature to study against time is the number of pages. As discussed above, the yearly averages of newspapers are already computed, so percentage of newspapers publishing unique number of pages is calculated. The following figure shows that in the last quarter of eighteenth century newspapers were publishing four or eight pages. The first quarter of nineteenth century indicates a drastic decrease in the ratio of eight pages while four pages are still a popular trend. This four pages trend continues for the whole century, and gets a slight decrease in the early years of twentieth century.

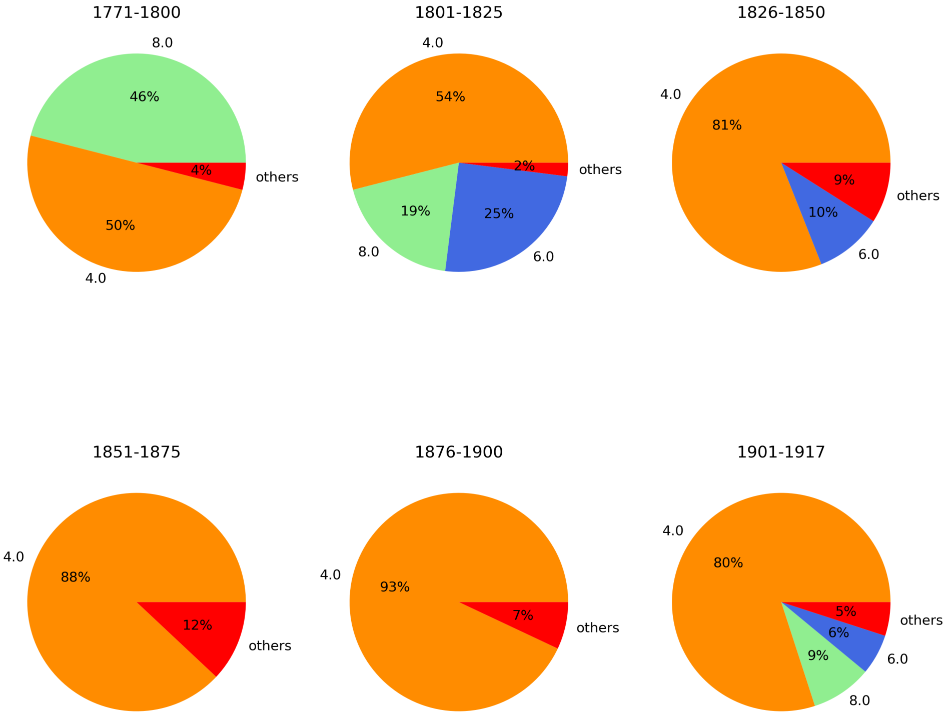


Figure 4.3: Popular trends of number of pages and percentage of newspapers following them over time

These were just the analysis of development of individual features, but to study the changes in different features, a comparison of all of them is the next step. Even though, the quantitative features are publication period, pages, columns, area and density, but any comparative analysis on area and density is not feasible since the variation in area is so high, and density is dependent on area in any case, so a new feature, 'page size', is generated. The area is in the unit of 10$^{\text{th}}$ of a millimeter, so using this area in millimeter square, page size is extracted. These dimensions are not exactly as the current paper sizes, A5, A4 etc. but these are mapped to the nearest possible ones. As per the current measurements, a paper of size 148 x 210mm^2, is A5. Since the measurements are in 10$^{\text{th}}$ of a millimeter, so an ISSN with area 71*151mm^2 is close to A5. The smallest area in the data set is 115 mapped as A5 and the largest is 5553 mapped as A2. There are four paper sizes in the data, A2, A3, A4, and A5.

To compare which features are changing first and which are the ones with least variation, yearly average data is utilized for four features, publication period, page size, pages, and columns. For a better representation, the numerical and categorical variables are plotted together in the (Fig. 4.4). The result shows that publication period is changing early on, then number of pages are changing. Around 1800 page size is changing but the interesting factor is that the changes in page size and number of pages, do not affect number of columns early on. Columns started to change after 1830.

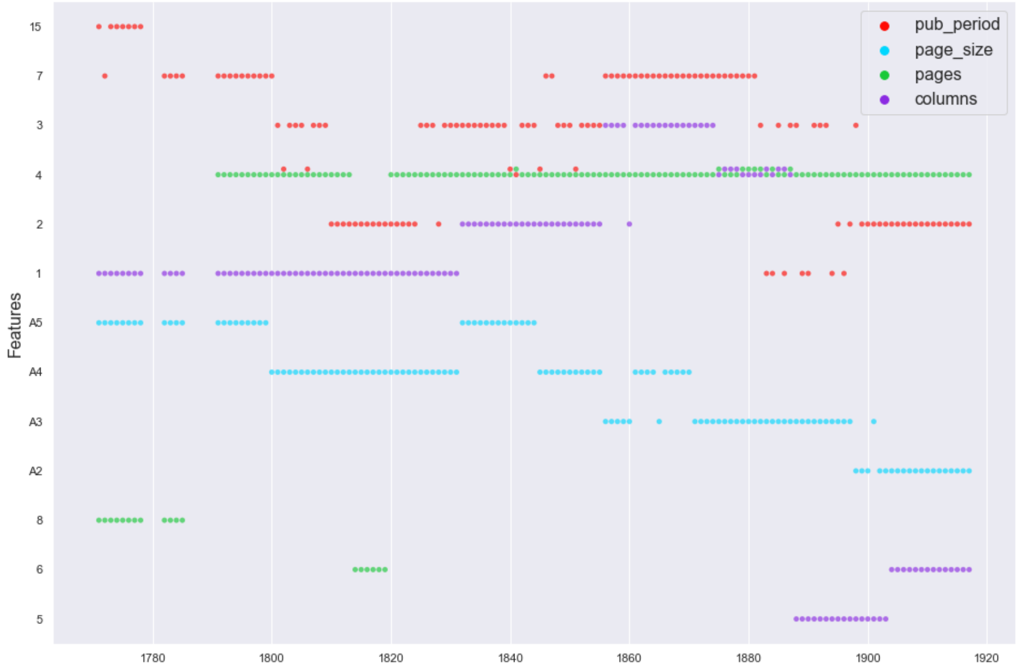


Figure 4.4: Comparison of feature changes over time

The above figure shows which features are changing earlier than other. The results are based on the popular category of a feature in a specific year. The feature is changing when another category is getting more popular. For example page size ’A5’ is popular in the initial years and then ’A4’ gets more popularity. This popularity is calculated by the percentage of newspapers in a specific year following that category. Though this result answers the question “which feature changes earlier than other”, but it does not show the ratio of newspapers following that category change. To understand it in more detail, percentage of newspapers falling under the popular category is calculated. The result below shows the percentage of newspapers following a category when the feature changes. This result is explained in the context of (Fig. 4.4) below. The results of each feature show that except the initial years, there is no category of any feature with a 100% popularity. So when in (Fig. 4.4), a feature changes, it represents the majority class. On average, popular categories have a ratio of more than 40% over the years for each feature. Number of pages always had a ratio of more than 50%, whereas page size also shows the same ratio after 1880. (Fig. 4.4) shows page size ‘A3’ and ‘A2’ after 1880, which indicates that a page size of ‘A3’ had more than 50% popularity and when ‘A2’ overtakes ‘A3’, page size feature changes. (Fig. 4.4) indicates 4 as a popular category for number of pages throughout the years after 1820 and the result below shows that after 1850, this category had a popularity of more than 80%. These results indicate that when a feature changes, it represents a majority category, though there can be still variations in the newspapers following less popular categories.

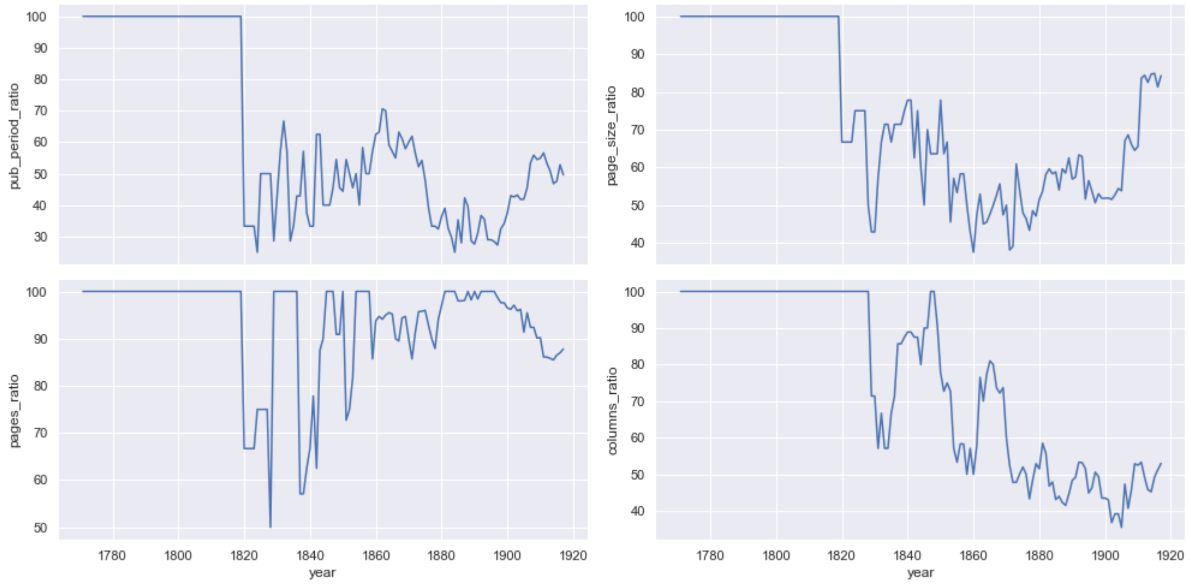


Figure 4.5: Percentage of newspapers following the most popular category per year

Now that the changes and development of features is described, analyzing the combination of these features and comparing them against time and place will give a clear understanding. Since the data is based on yearly averages of these features, so a new feature 'pattern' is generated by just combining the above four features. If a newspapers' yearly average shows a publication period of 7 days, page size A4, number of pages 4 and number of columns 1, this new feature will be of the shape 7-A4-4-1. As this pattern represents an ISSN for a specific year, so the mode value of this feature is computed for a given year. This mode value is considered as the popular trend of that year.

To understand the importance of these patterns, percentage of newspapers following a popular trend of that year is also computed. This gives the ratio of newspapers following a popular trend over the years. The figure below (Fig. 4.6) shows the popular trends over the years against total number of papers.

In the data, there are one or two newspapers in total for the first fifty years, that shows the ratio of 1, but when the number of newspapers starts increasing around 1820, ratio of newspapers following a popular trend decreases drastically. This suggests that there is a large variation in publishing patterns. Between the years 1835 and 1865, ratio of newspapers following a popular trend is almost 0.30, but after that exponential increase in number of papers causes a substantial decrease in the ratio. The following figure shows the wide range of publishing trends in the late nineteenth and early twentieth centuries.

These patterns are discussed against time, but to analyze them against publication places, the 'city' feature is divided into three categories, metropolitan cities, coastal cities, and small towns. The division of metropolitan and small towns is based on obvious indices, such as population, official status etc. This gives eight metropolitan cities,

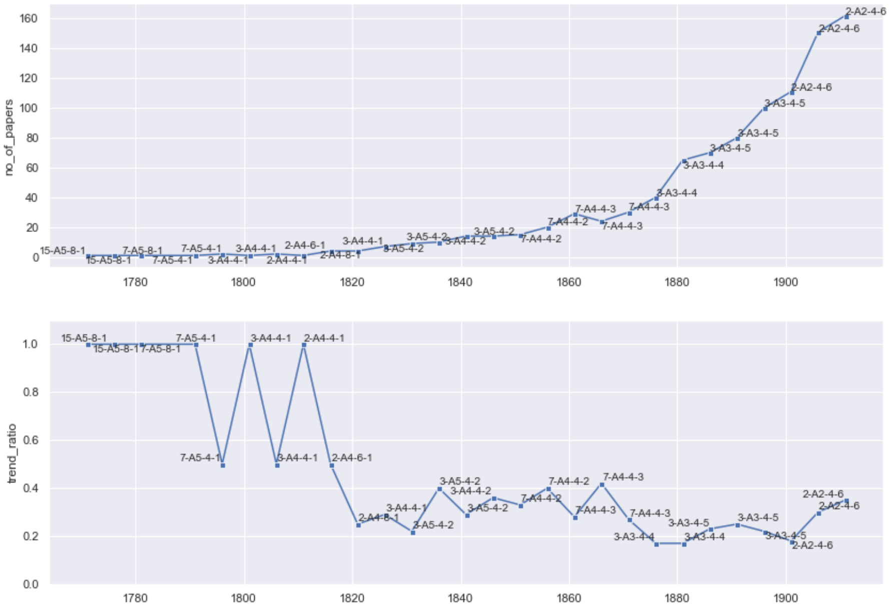


Figure 4.6: Number of newspapers and ratio of newspapers following popular trends

eleven coastal cities, and thirty-two small towns. Though the separate popular trends of metropolitan, coastal and small towns are calculated, but instead of showing the trends and ratio individually, (Fig. 4.7) shows the number of newspapers and ratio of following a trend for the three different categories. The blue line shows the number of newspapers published in metropolitan cities over the time. Similarly, green line represents coastal cities and red line denotes small towns. The coastal cities and small towns start publishing newspapers after 1825 and the number of newspapers were almost the same for both the categories until 1880.

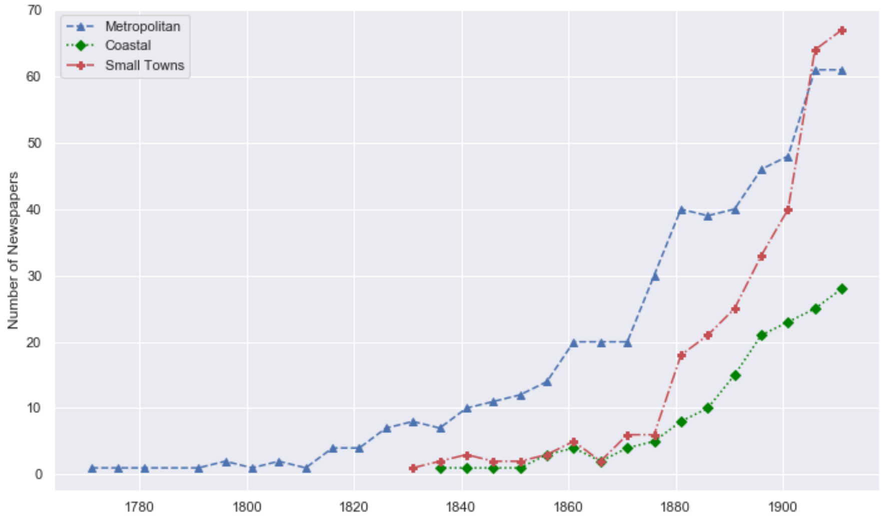


Figure 4.7: Total newspapers in different regions, 1771-1917

The result in (Fig. 4.7) shows an immense increase in the number of papers in small towns, surpassing metropolitan cities in the end, while the total number for coastal cities increases gradually. Though these results explain the increase in number of newspapers in different regions over time, but they do not uncover the underlying variations in the number of newspapers for each city in these regions. Total number of small towns publishing newspapers is 32, whereas total coastal cities and metropolitan cities are 8 and 11 respectively. Any comparison on total number of newspapers will most probably show small towns ahead of other regions which does not provide any concrete results of region comparison. To better understand the increase in number of newspapers in each region, average number of newspapers per city is calculated. The result in (Fig. 4.7) shows continuous increase in the number of newspapers for each region, but the result below shows some fluctuations for each region, though metropolitan cities' are publishing more and more newspapers on average over time. In small towns and coastal cities, on average each city published between one and two newspapers whereas metropolitan cities' published on average less than three newspapers between 1820-1880 and after that the average number increased up to six. By comparing this result with (Fig. 4.7) it can be seen that the total number of newspapers in small towns surpasses metropolitan cities after 1900 which indicates more and more small towns start publishing around that time but on average the metropolitan cities' are publishing more newspapers than two other regions.

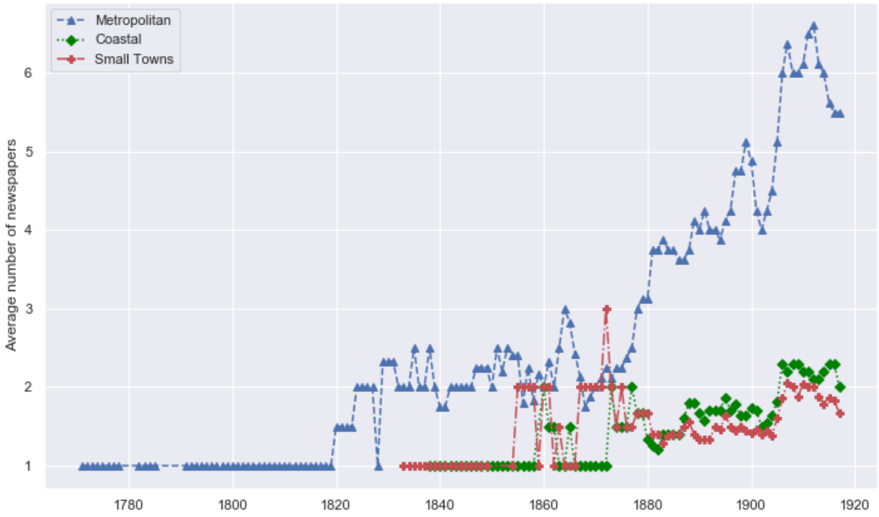


Figure 4.8: Average newspapers per city in different regions, 1771-1917

The ratio of newspapers following a popular trend in three regions is calculated as shown in (Fig. 4.9). The marker size indicates total number of newspapers in that region. As discussed above in (Fig. 4.6), with the increase in total papers, the percentage of newspapers following a popular trend decreases. The interesting factor here is that the total newspapers in small towns are increasing exponentially but the ratio of following a trend is decreasing, while there is a steady increase in total newspapers in coastal cities, but the ratio of following a popular trend shows more fluctuations. Around 1880, the ratio of newspapers in coastal cities following a trend is almost 0.20, which doubles for next decade, and then fell back to around 0.20 before increasing to 0.50 in the early twentieth century. With the sudden increase of total newspapers in small towns, a more fluctuation was expected in the ratio of following a trend, but interestingly coastal cities are showing more fluctuation as compared to small towns. To sum up (Fig. 4.7), (Fig. 4.8) and (Fig. 4.9), overall the metropolitan cities published more newspapers before small towns surpasses this number after 1900. On average metropolitan cities are ahead of two other regions in publishing newspapers but coastal cities are leading in following popular trends.

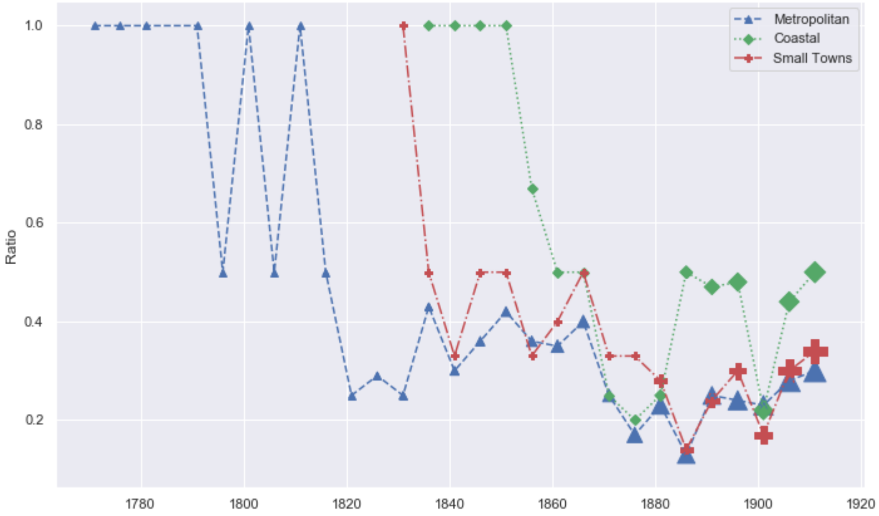


Figure 4.9: Ratio of newspapers following the popular trends in different regions

Though the above results give some understanding of the popular trends and percentage of newspapers following popular trends in different places, but they do not explain the relation between metropolitan cities, coastal cities and small towns in terms of following popular trends. For this purpose, coordinates of all the cities are used to find the distance among them, and based on the distance, metropolitan cities close to coastal cities and small towns are grouped. For example, distance of the town Mikkeli is calculated with all other cities in metropolitan cities list, and the closest main city to Mikkei is, Kuopio. Similarly, distance for coastal cities to metropolitan cities is calculated and the closest metropolitan city is assigned to the coastal cities. For the comparison purpose, percentage of coastal cities and small towns, which are closely located around Helsinki region, is calculated. The objective is to find out either small towns and coastal cities follow a popular publishing trend of their nearby main cities or the trend of capital region, Helsinki.

The following results are for the coastal cities. As mentioned already, there are

11 coastal cities, and 38% of them are located around Helsinki whereas 62% have other metropolitan cities nearby. The second pie chart shows that only 19% of the coastal cities are following popular trends of Helsinki and 55% are following trends of their own nearby main cities. There are 10% coastal cities which are following common pattern,which means these trends are same in Helsinki and their nearby main cities while 16% of them are not following any trend.

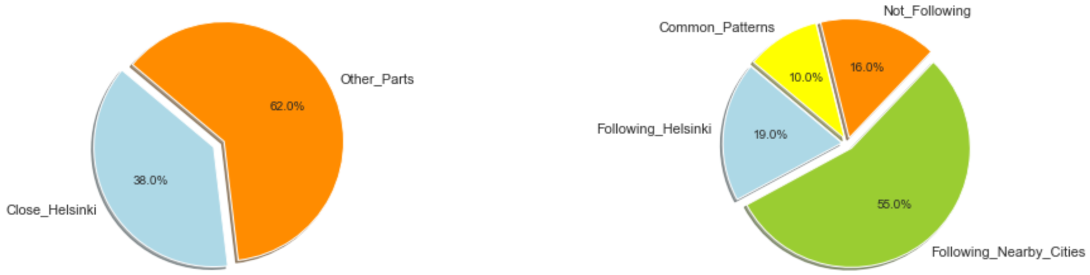


Figure 4.10: Coastal cities and comparison to capital region: Left) Ratio of coastal cities near Helsinki. Right) Ratio of coastal cities following popular trends of Helsinki or their nearby main cities

The same procedure is applied on small towns and the results are shown in (Fig. 4.11). Of all the small towns, 27% are located near Helsinki and remaining are located nearby other metropolitan cities. The trend following ratios are almost equal for small towns, as 33% following popular trends of Helsinki while the same percentage of small towns following popular trends of their nearby main cities. Almost one third of the small towns do not follow Helsinki or their nearby main city's trends, which is interesting in a way that the small towns are publishing with a trend of their own.

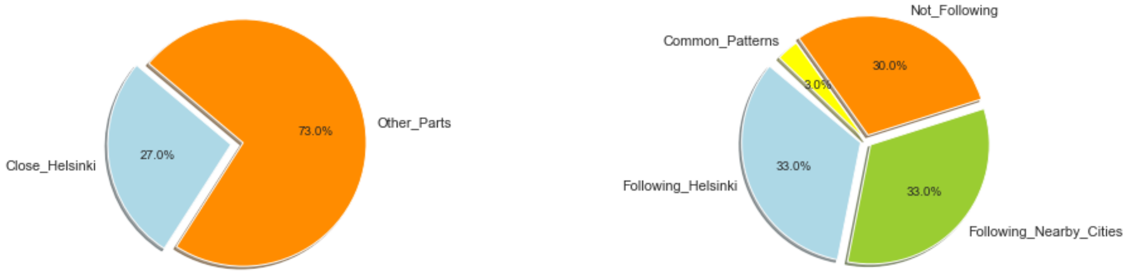


Figure 4.11: Small towns and comparison to capital region: Left) Ratio of small towns near Helsinki. Right) Ratio of small towns following popular trends of Helsinki or their nearby main cities

The focus of interest in above analysis was cities, their ratio of following popular trends and comparison of small towns and coastal cities in terms of following a popular trend of capital region. Another aspect of analyzing the popular trends is to study when a trend starts getting popular and over time which of the trend disappears. Since the above results show a lot of variation in the trends and no single trend has a strong following ratio, which produces many trends in a given time.

To go further with the analysis, a threshold value is required to get only the trends with a ratio equal or greater than the threshold. The distribution of the ratio is plotted in the figure below, and it shows the variation in the trends over the years. As the number of newspapers increases with time the ratio of following a specific pattern decreases. The graph shows that most of the trends had the newspapers' following of 10-20%.

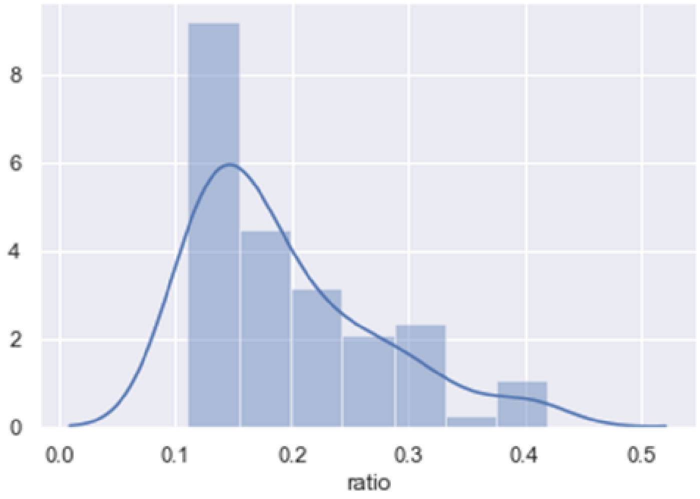


Figure 4.12: Distribution of trends following ratio of total newspapers

To study when a trend gets popular and when it disappears, only the trends with the following ratio of at least 0.20 are selected. The result below shows the trends against time. The vertical bubbles show more than one popular trends at a given time, while the bubbles on horizontal axis show the popularity of the trend for a longer period. The bubble size indicates the ratio of newspapers following the trend. Since there are maximum two newspapers before 1820, so that data is excluded from this analysis. Between 1820 and 1825, several newspapers start publishing all at once, each with its own publishing trend, which can be seen in the figure below. Around 1830 a trend '3-A5-4-2' gained popularity, which increased around 1835, and it disappeared after 1840. Around 1850, four publishing trends are popular but all disappeared except one '7-A4-4-2', which is dominating between 1850 and 1860. In the next decade, the trend '7-A4-4-3' is the only trend with prominent following ratio, but after that, for more than 10 years, there is no trend with a following ratio of 0.20. This shows the variation in publishing trends at that time. Considering the number of newspapers during this time, it is obvious, with the exponential increase in number of papers, there is a less possibility of any trend getting large following ratio.

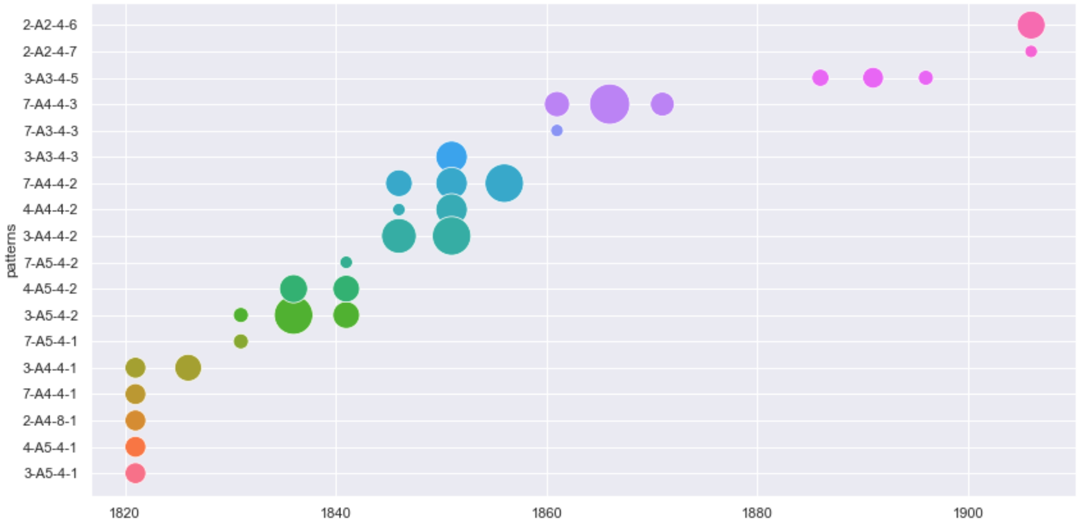


Figure 4.13: Trends popularity between the time period of 1820-1917

The above is a general depiction of trends getting popular and disappearing over time, but to understand it in more depth, language based analysis are performed. The result below shows trends popularity for Swedish newspapers. By comparing this result with (Fig. 4.13), it shows that Swedish language newspapers were influencing the popular trends mostly over the time. The reason for this is very less number of newspapers in Finnish language early on. The general popular trends until 1850 are almost the same as in Swedish language. After 1860, there are not many popular trends in Swedish language with a following ratio of at least 20%.

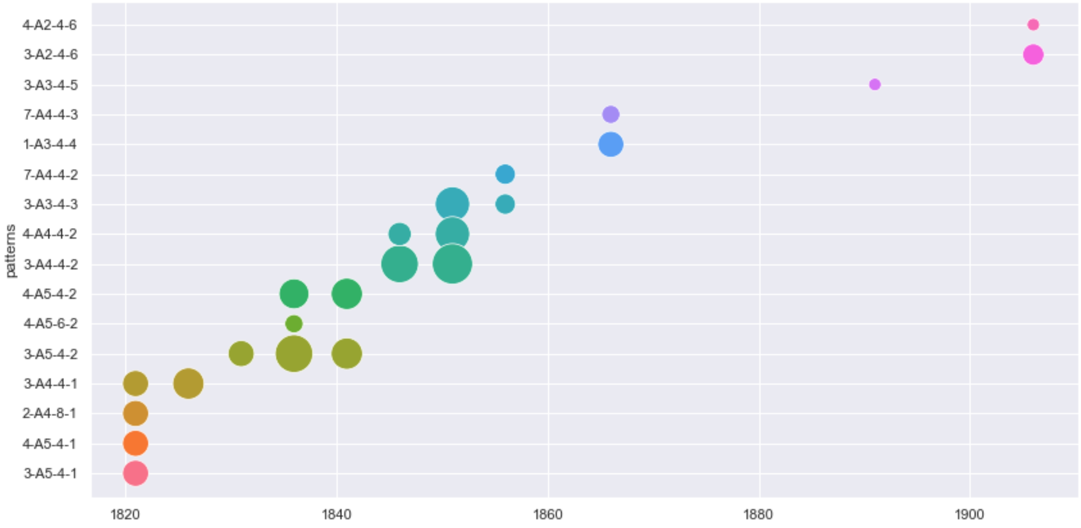


Figure 4.14: Swedish newspapers trends popularity between the time period of 1820-1917

The figure below shows the increase and decrease in a trend's popularity of Finnish language newspapers. The popular trends in Finnish newspapers are different than general trends in the early years, but after 1880 the general trends are almost the same as of Finnish newspapers. If the above result indicates the impact of Swedish language on setting the trends in the early years, Finnish newspapers are effecting the popular trends in the later part of nineteenth century and early twentieth century. These results are completely synchronised with (Fig. 4.1), as they are effected by the total number of newspapers for each language. Until the number of Swedish language newspapers were high, they are having an impact on setting the popular trends and when the number of Finnish language newspapers starts increasing, they started effecting the popular trends. Interesting factor here is the last quarter of the nineteenth century when there are not many popular trends in general and of Swedish language also, but Finnish newspapers show several trends with a popularity of at least 20%.

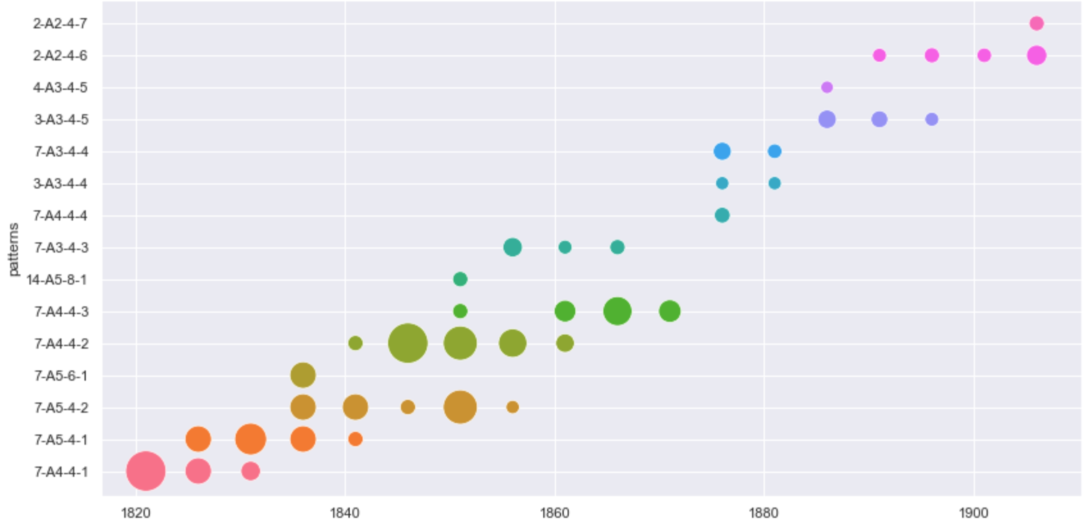


Figure 4.15: Finnish newspapers trends popularity between the time period of 1820-1917

Since the individual features and their combination, is studied so far in various capacities, but the effects of following a popular trend is not touched at all. It is important to know the impact on newspapers lives, following a popular trend partly or completely for all the publishing years. For this, two main scenarios are discussed in the following sections.

As there are two language categories in the data, the impact of following a popular trend is studied for both of them. The newspapers which published for more than 10 years are selected, and based on the publishing language, the ratio of newspapers following a popular trend is calculated. Following figure shows that the newspapers published in Finnish language have a higher ratio of following popular trends as compared to Swedish. On average, Finnish newspapers are following popular trends with a ratio of 0.20, while Swedish language newspapers have average ratio of around 0.05.

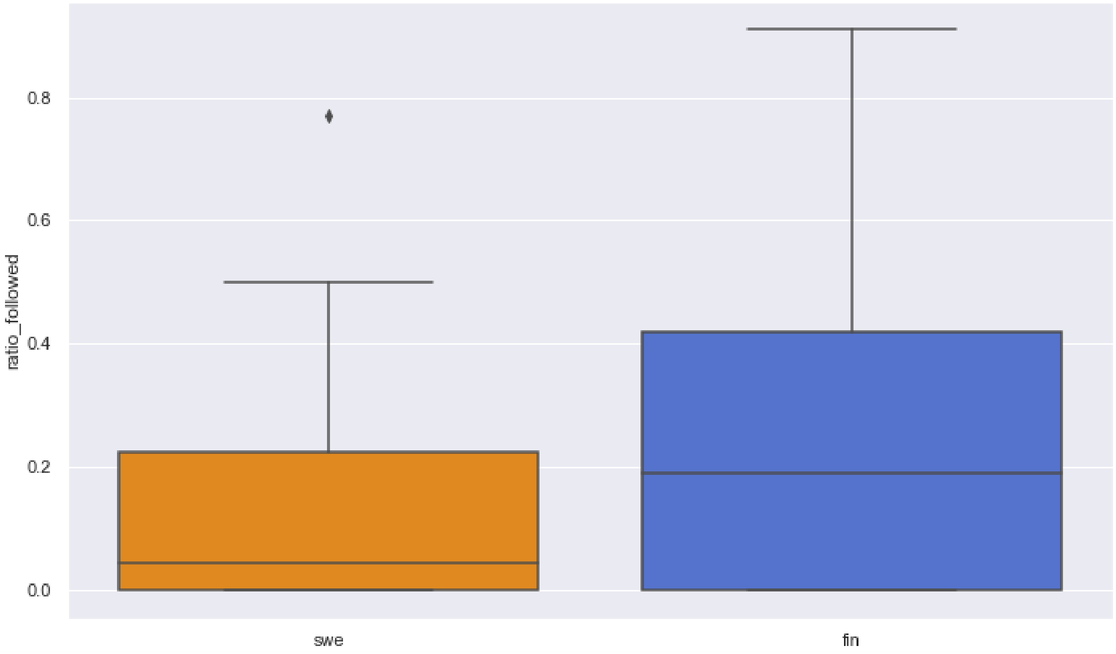


Figure 4.16: Ratio of Swedish and Finnish newspapers following popular trends

This was a case study of long lived newspapers of different languages and their ratio of following popular trends, but another scenario to discuss is the effect of not following a popular trend in the last years of newspapers. This case study is to find the association between popular trends and newspapers last years' publishing trends. The newspapers and their last two years publishing trends are considered for this case study. The variable 'fully follow' in the figure below indicates that the newspaper followed popular trend of that time in the last two years of its life. The variable 'partly follow' shows that the newspaper in its last two years followed one year a popular trend of that time, while the variable 'not follow' represents the newspapers which did not follow popular trend in last two years at all. The result below is dependent on the interpretation. One argument is that it shows a strong impact of popular trends on the last years of newspapers, as 70% of the newspapers died when they did not follow a popular trend. The other way to interpret it is that there were many popular trends and none of them had a strong following ratio, so this result does not strongly suggest that the reason for the end of newspapers is because it was not following a popular trend. This is the explanation of a coarse-grained system, but talking about fine-grained system, this result makes sense when just compared the three values of the figure below. The percentage of dead newspapers which fully or partly followed a popular trend is 13 and 12 respectively, while the percentage of dead newspapers which did not follow popular trend at all is 75. Just focusing on the result below, it puts emphasis on the impact of following a popular trend.

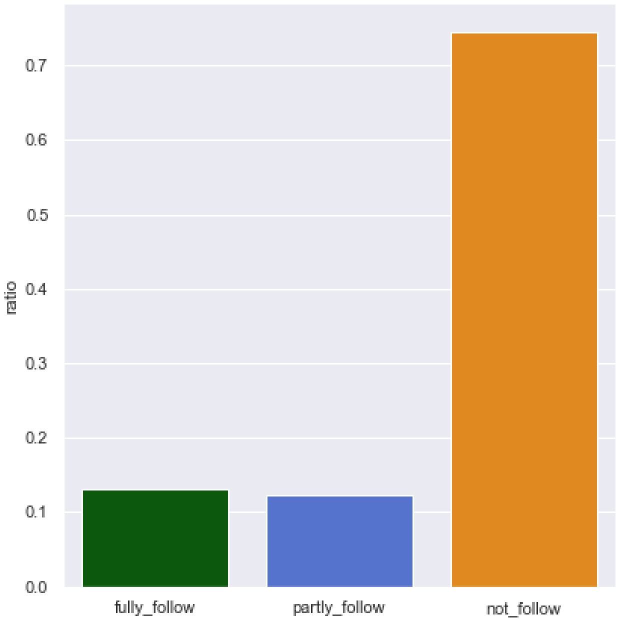


Figure 4.17: Ratio of newspapers following popular trends in the last two years of their lives

The last case study is the comparison of long running newspapers against time and place. Though in the above sections, newspapers' lives are discussed one way or the other, but no direct case study is discussed. Here the newspapers and their lives are compared not only in different regions but also in different times. The result below shows the life of newspapers in metropolitan cities, coastal cities and small towns. Average life of a newspaper in metropolitan cities is higher than small towns but equal with coastal cities. On average newspapers in coastal and metropolitan cities lived more than 20 years whereas in small town the average life is just above 10.

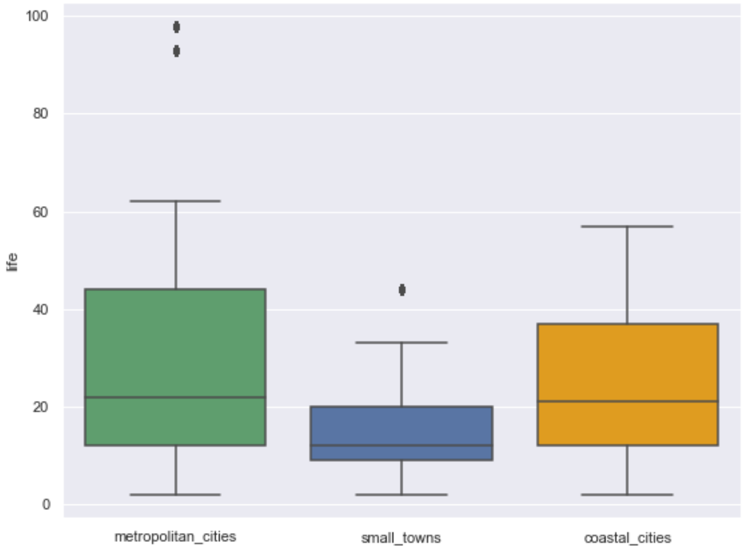


Figure 4.18: Newspapers' life in different regions

The above result shows lives of newspapers in different regions, where long living newspapers appear and the average lives of newspapers. Following analysis is performed to find the significance of time when the long running newspapers start publishing. Since a single value(year) in time cannot give any concrete result, so the time is divided in three periods, Before 1850, Between 1850 and 1875, and After 1875. The result below shows the newspapers and their lives in three different time periods. Two newspapers which published for more than 90 years started before 1850 but overall newspapers from this time period lived less than 20 years with a couple of exceptions. The newspapers which started after 1850 show longer lives than the first half of the century. The newspapers which started after 1850 mostly lived less than 20 years. An important factor to remember here is the time period of the data. Since the data is until 1917, the newspapers which started after 1875 can not show lives of more than 40 years. To understand the importance of different time periods in association with long lived newspapers, next analysis is performed with a fixed threshold life of newspapers.

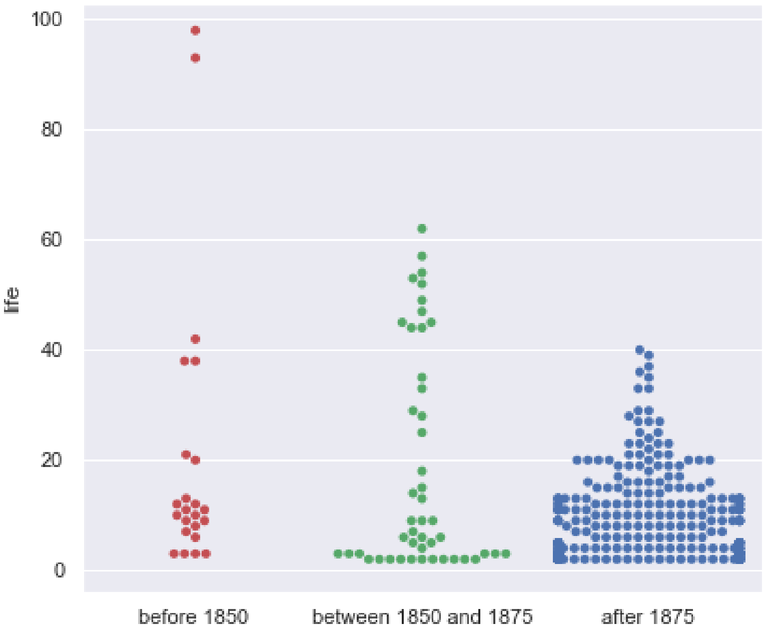


Figure 4.19: Newspapers' life in different time periods

Total numbers of newspapers which started publishing in the given time periods are extracted and the ratio of newspapers which lived for more than 20 years is computed. The result shows that 26% of the newspapers from the time period Before 1850 lived 20 years or longer. Third quarter of the century shows the highest ratio, as 34% of the newspapers which started publishing in this time lived for more than 20 years. The interesting fact here is to see the result of third time period, which shows only 9% of the newspapers from this era living at least 20 years.

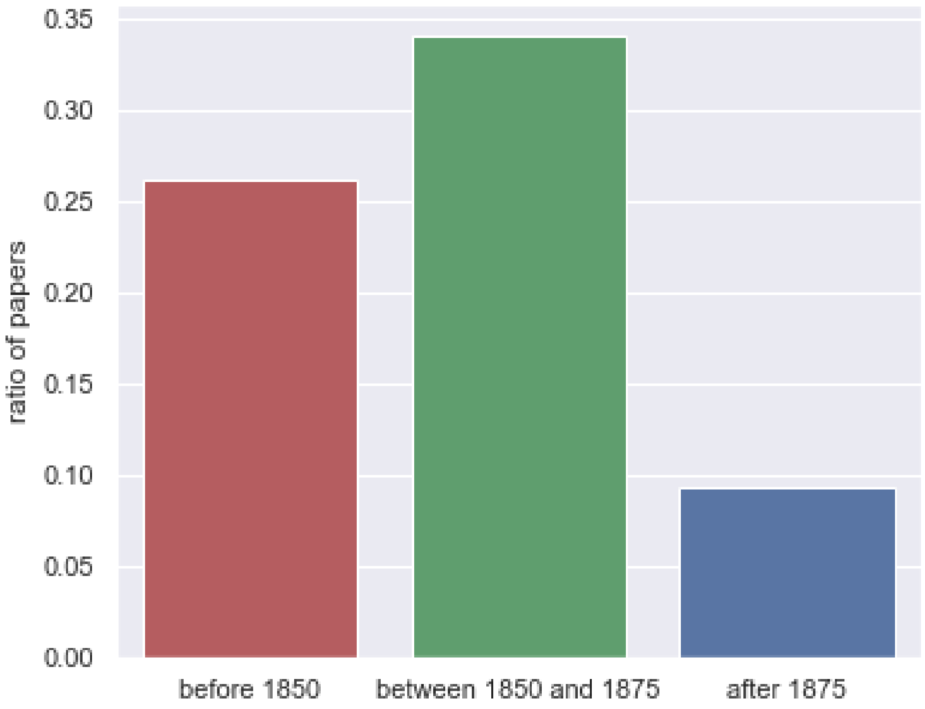


Figure 4.20: Comparison of three time periods when the long lived newspapers started publishing

These results are dependent on the life threshold used here. If the threshold is changed to 15 or 25 the results completely change, but this threshold is a better representation of the data. Though in the above figure the average life is said to be 43, the reason for not applying a threshold of 43 is that the data is until 1917, and dividing the time in three periods, cannot possibly give better results for the time period After 1875 with a threshold of 43.

4.2 Correlation Analysis

Different features and their impacts have been studied until now, but an important aspect of this research work is to find the dependencies of features on each other while studying the development of newspapers in Finland. In the above sections, it has been observed that features are changing over time continuously, but the combination of feature change is not discussed. In the next sections, first couple of correlation methods are discussed in general and then their results on the features and features' combinations are evaluated thoroughly.

The aim of this case study is to find the correlation among the features. At its core, correlation is a statistical method to find a linear relationship between continuous features. The correlation can be negative with the increase in one variable decreases the other, or it can be positive where the value of one variable increases so does the value of second variable. These changes in the same or opposite direction are related with the changes in the magnitude of variable.[21][33]

A numerical measure, correlation coefficient, is used to calculate correlation, with the range between -1 and +1. A value of -1 represents a perfect negative correlation, suggesting that the variables are changing in the magnitude in opposite direction. A +1 suggests perfect positive correlation where variables are changing in the same direction. Correlation coefficient zero indicates no correlation among the variables at all. There are multiple correlation coefficients methods, such as Pearson, Spearman, Kendall etc, but for this study, the focus will be on Pearson, and Spearman correlation coefficient. The reason for studying two methods is to validate the variables' relationship suggested by one method with the other. Pearson correlation coefficient method holds some assumptions such as linearity among variables and normal distribution of the the variables whereas Spearman does not hold any such assumption.[21][33]

4.2.1 Pearson Correlation Coefficient

Pearson correlation coefficient assumes that the variables are normally distributed [21], so the normality test is performed on the features. With the increase in sheer volume of data, normality test will most probably be rejected for any data set even with the smallest deviation from the perfect normality. Since there will always be some degree of randomness, it is not possible for a data set to be perfectly normally distributed. In this case study, the importance is not whether a data set holds perfect normality, but holds enough normality for the assumption to be true.

As the features area and density are continuous, a Quantile-Quantile (QQ) plot [5] is a better representation to check if the data is normally distributed. Normal distribution of the area is tested first. The following result shows that the sample points are almost fitting the expected diagonal points of a Gaussian distribution. The values at extremes are deviating from the expected values, showing few number of samples actually exist at those limits. This is not a perfect normal distribution but it follows the theoretical line of normal distribution, so this feature is good to be considered for Pearson correlation coefficient measurement.

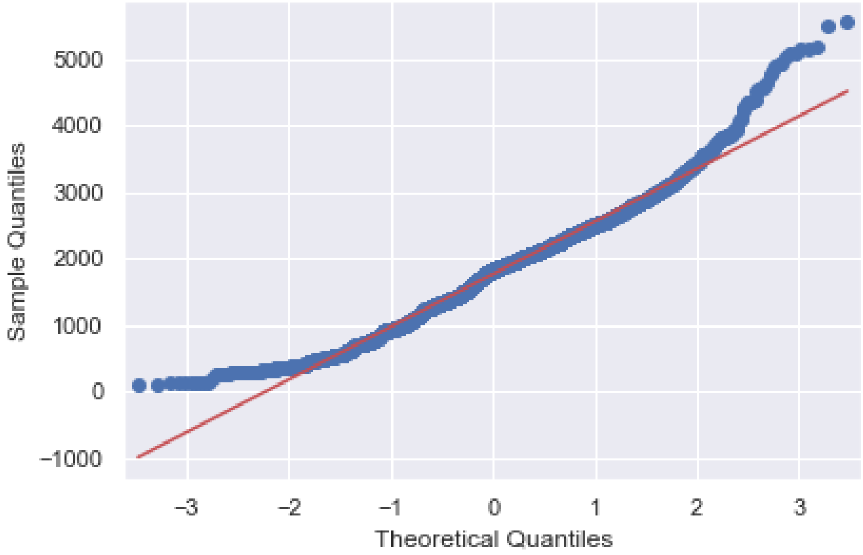


Figure 4.21: Normal distribution test of Area

Following the same method, normal distribution of density is tested. The below result shows the same patterns as above. The sample data points fit well with the expected Gaussian distribution, but at one extreme, there is a slight deviation, which is negligible. Density also shows the behavior of normal distribution, so it is also a good candidate for the Pearson correlation coefficient measurement.

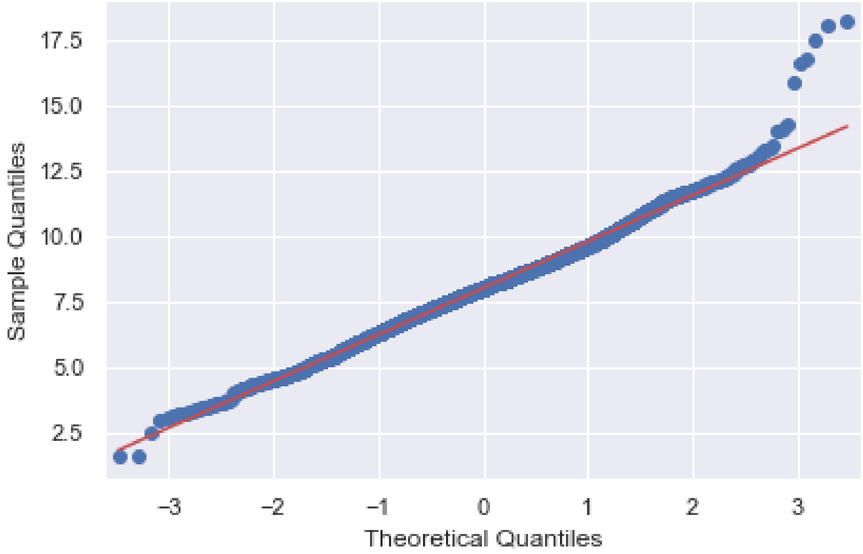


Figure 4.22: Normal distribution test of Density

Since the remaining three quantitative features, columns, pages, and publication period, are discrete variables, so instead of applying qqplot to check either they come from a normal distribution or not, Poisson distribution is used. This does not assume that these features come from a normal distribution, but looking at the plots, it shows a good approximation of normal distribution. The figure shows the Poisson distribution of columns. The Poisson distribution is computed with the mean value of columns. The result shows a bell shape representation of columns, which is an approximation of normal distribution.

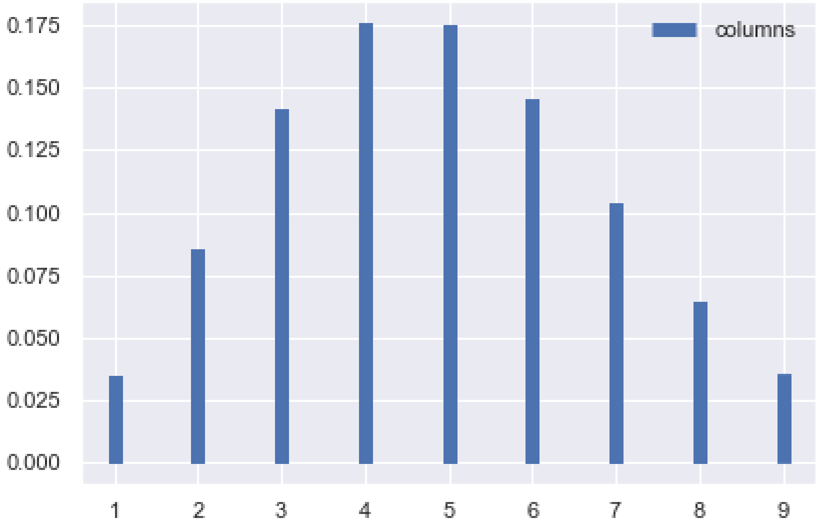


Figure 4.23: Normal distribution test of Columns

Similarly, Poisson distributions of pages and publication period are calculated. Following figure shows the approximation of normal distribution for feature columns.

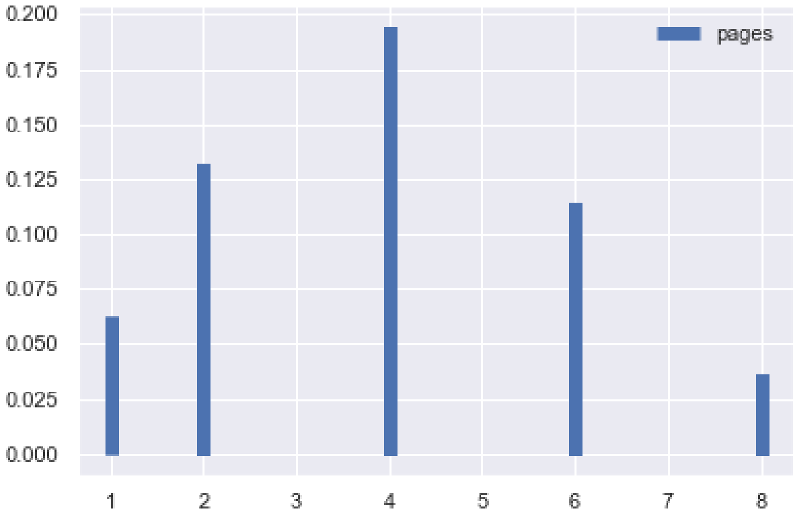


Figure 4.24: Normal distribution test of Pages

The mean values of both these features are considered for the calculation and plotting purpose. The figures below shows the approximation of normal distribution of publication period, though statistically they are not generated through a normal distribution.

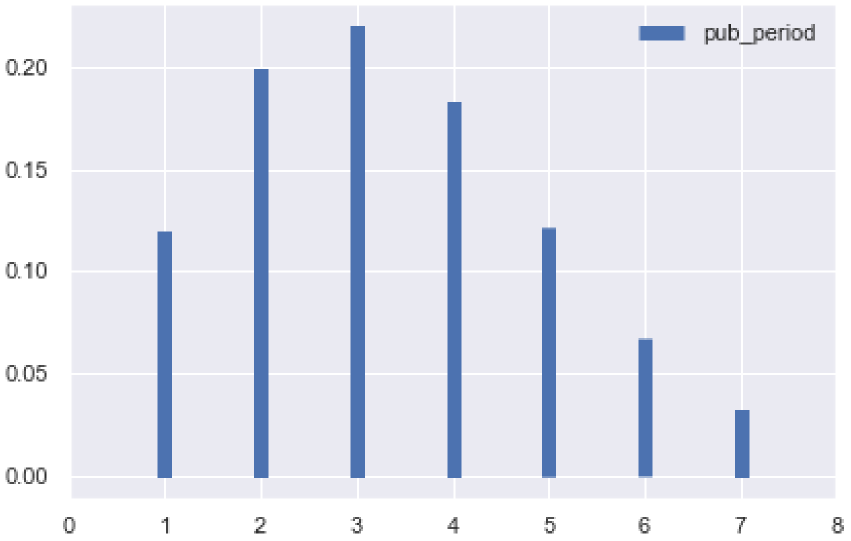


Figure 4.25: Normal distribution test of Publication Period

The formula to calculate Pearson correlation coefficient is

$$pearson\_corr = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4.1}$$

where $\mathrm{x_i}$ is the $\mathrm{i^{th}}$ value of vector x, $\overline{x}$ is the mean value of vector x, $\mathrm{y_i}$ is the $\mathrm{i^{th}}$ value of vector y and $\overline{y}$ is the mean value of vector y [21]. By applying this formula on the features, the correlation quantities are measured as shown in (Table 4.1). The table shows negative correlation between publication period, columns and area while a positive correlation between area and columns. There also exists a negative correlation between density and area. Year is also added in this analysis to understand it's correlation with other features. The result below shows that year is positively correlated with columns and area and has a negative correlation with other features.

4.2.2 Spearman Correlation Coefficient

The above results are a good approximation of the features' correlation in general but as the primary assumption of Pearson method, variables should come from a normal distribution, is not fully verified in three features, so these results are less reliable. Another method Spearman correlation coefficient is applied on these features.

Detail study of both these methods is out of scope of this research work, but with the basic assumption that Spearman method can be used for random or discrete variables

Table 4.1: Pearson Correlation

	year	Publication_Period	Pages	Columns	Area	Density
year	1.00	-0.269	-0.034	0.816	0.639	-0.113
Publication_Period	-0.269	1.00	0.025	-0.388	-0.390	0.057
Pages	-0.034	0.025	1.00	-0.023	-0.030	0.054
Columns	0.816	-0.388	-0.023	1.00	0.825	-0.030
Area	0.639	-0.390	-0.030	0.825	1.00	-0.367
Density	-0.113	0.057	0.054	-0.030	-0.367	1.00

along with continuous variables makes it useful in this case study. The formula for the Spearman correlation coefficient is

$$spearman\_corr = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)} \tag{4.2}$$

where n is the number of observation, x and y are the vectors[21]. By applying this formula, the correlation among the features is measured, as shown in (Table 4.2). This method has the same relations among variables as Pearson, such as publication period, pages, and area are negatively correlated while columns is positively correlated with area. Density in this method is also negatively correlated with area. Year is negatively correlated with publication period, pages and density while has a positive correlation with other features. This method shows strong correlations, among the features, as compared to Pearson method.

Table 4.2: Spearman Correlation

	year	Publication_Period	Pages	Columns	Area	Density
year	1.00	-0.295	-0.073	0.725	0.662	-0.270
Publication_Period	-0.295	1.00	-0.130	-0.472	-0.505	0.054
Pages	-0.073	-0.130	1.00	0.023	0.009	0.065
Columns	0.725	-0.472	0.023	1.00	0.855	-0.082
Area	0.662	-0.505	0.009	0.855	1.00	-0.369
Density	-0.270	0.054	0.065	-0.082	-0.369	1.00

To get a better understanding of the correlations among features, figure(4.26) is a good representation. The Spearman method's results are plotted here. The positive correlations are represented by red color while negative with blue color. As discussed above, Publication Period, Columns and Area have negative correlations. Area has positive correlation with Columns and negative correlation with Density. Interestngly pages and publication period show negative correlation in Spearman method, but their correlation coefficient is not strong enough to be considered for further analysis.

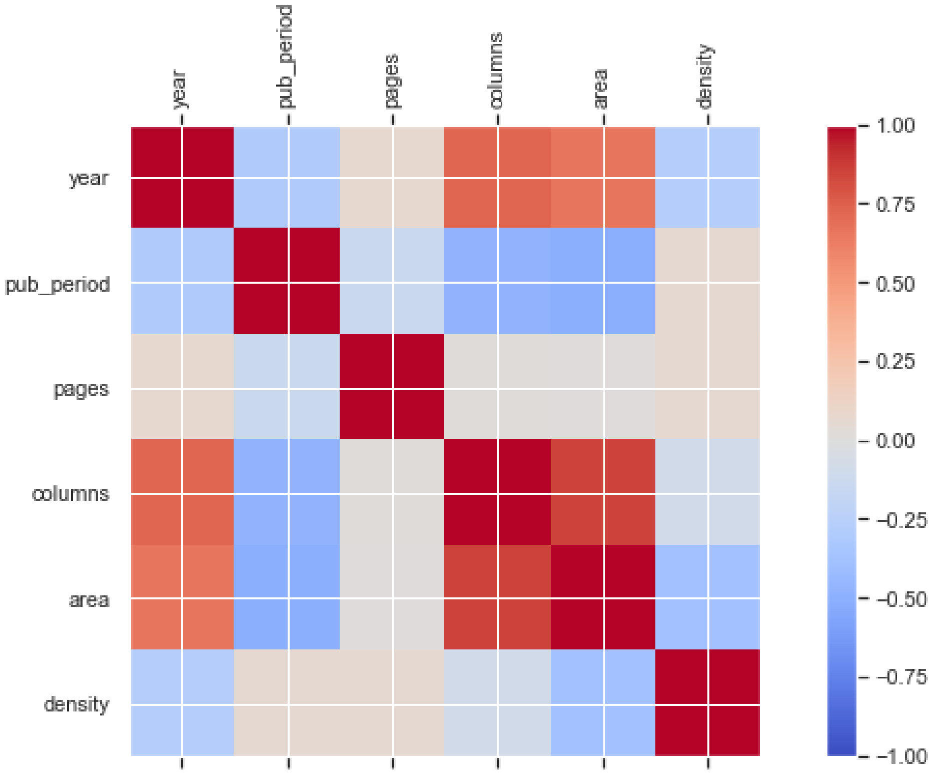


Figure 4.26: Correlation of different features: negative correlation (blue), positive correlation (red), no correlation (gray)

4.3 Ordinary Least Squares method

Measuring correlation among various features is part of analyzing the data, but for predictive modelling, statistical methods need to be applied. As the features in correlation table indicate linear relationships, linear regression models are studied and implemented.

A simple linear regression model is just an approach of studying relationship among a dependent variable and an independent variable. A good example of this model could be the feature density which is dependent on the area. This linear regression model is studied later, but first the features year, publication period, area and columns are studied as they all are showing significant correlation. Since pages does not show strong enough correlation with any feature, so it is not considered for the further analysis. Columns have strong correlation with year, area and publication period, this feature is suitable to explore further. Since this involves multiple features, Ordinary Least Squares (OLS) regression model [15] is best suited for this scenario. OLS is the most widely used method for multiple linear regression. It minimizes the error such that sum of all squared residuals is minimized. The formula to derive OLS is given as

$$y_i = B_0 + B_i X_i + \epsilon \tag{4.3}$$

where y_i is the dependent variable or estimated value, B_0 is the intercept, B_i is the slope of the explanatory variable X_i and ϵ is the error term [15]. The slope B_i is calculated by summing the differences of the dependent and independent variables from their mean values and dividing them with difference of independent variables' sum of squares.[15]

$$B_i = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4.4}$$

and

$$B_0 = \overline{y} - B_i\overline{x} \tag{4.5}$$

The objective of OLS is to find the values of B_0 and B_i which minimize the errors between true value of y and expected value of $\overline{y}$.[15]

$$sum\_squared\_error = \sum(y - \overline{y})^2 \tag{4.6}$$

As the number of variables increase, number of slopes increase, for example in the case under study the formula of OLS would be

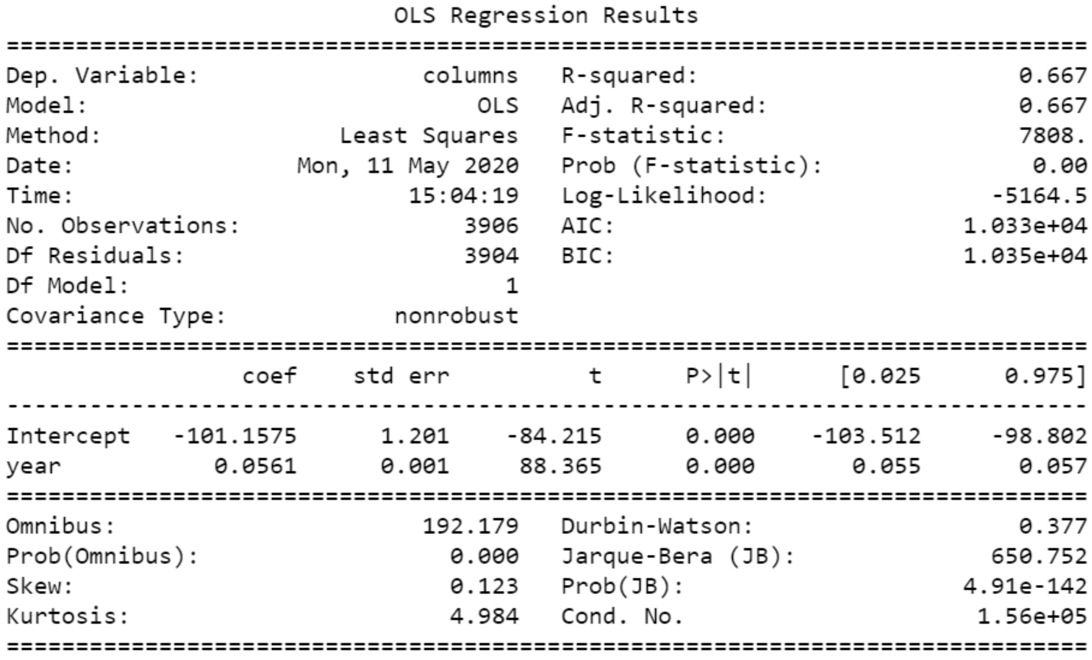
$$columns = B_0 + B_1 * Publication\_Period + B_2 * Area + B_3 * Year + error\_term \tag{4.7}$$

The above equation is to estimate columns through three mentioned features. Since a driving force in the result of correlation analysis is the feature year ans Columns shows strong positive correlation with year so first OLS method is applied to estimate columns by year. The equation is given below.

$$columns = B_0 + B_1 * Year + error\_term \tag{4.8}$$

OLS method built by the statsmodels* is implemented for this case study. The model's summary is extracted below

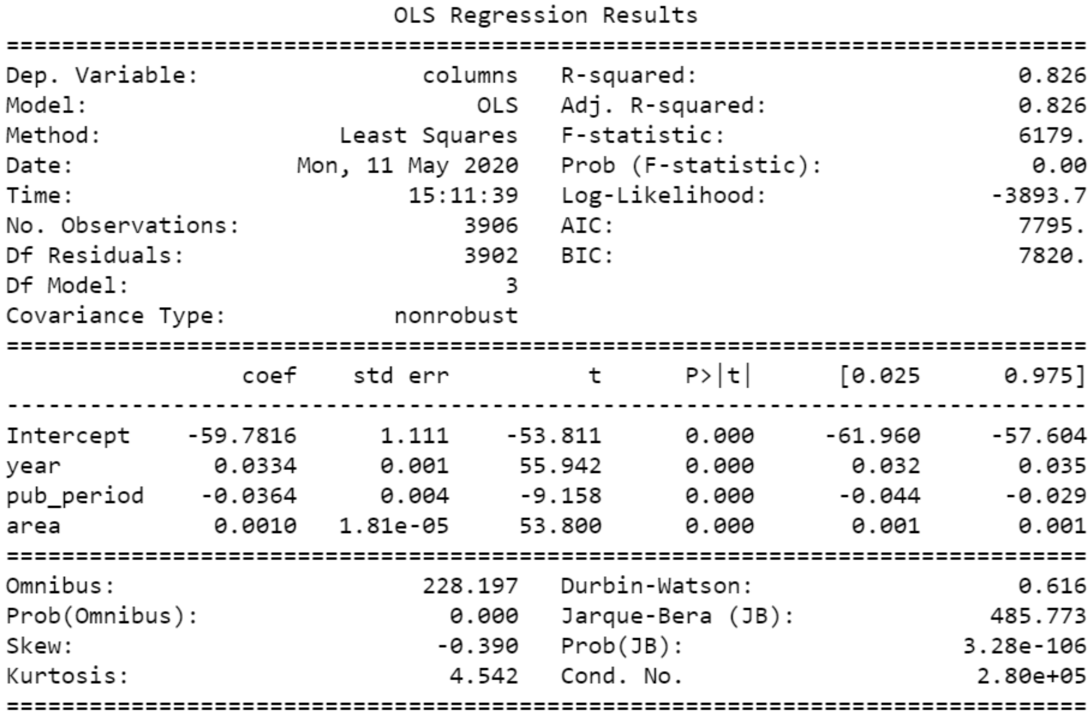
*https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                columns   R-squared:                       0.667
Model:                            OLS   Adj. R-squared:                  0.667
Method:                 Least Squares   F-statistic:                     7808.
Date:                Mon, 11 May 2020   Prob (F-statistic):               0.00
Time:                        15:04:19   Log-Likelihood:                -5164.5
No. Observations:                3906   AIC:                         1.033e+04
Df Residuals:                    3904   BIC:                         1.035e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -101.1575      1.201    -84.215      0.000    -103.512     -98.802
year           0.0561      0.001     88.365      0.000       0.055       0.057
==============================================================================
Omnibus:                      192.179   Durbin-Watson:                   0.377
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              650.752
Skew:                           0.123   Prob(JB):                    4.91e-142
Kurtosis:                       4.984   Cond. No.                     1.56e+05
==============================================================================
```

Figure 4.27: Summary of OLS regression analysis for predicting Columns with Year

The result shows p value less than 0.05, which indicates a strong evidence to reject null hypothesis. The model has an adjusted R-square value of 0.667, which suggests that this model has explained 66.7% of the variance in the data. Since columns have strong enough correlation with two more features, which are added in this model as shown below.



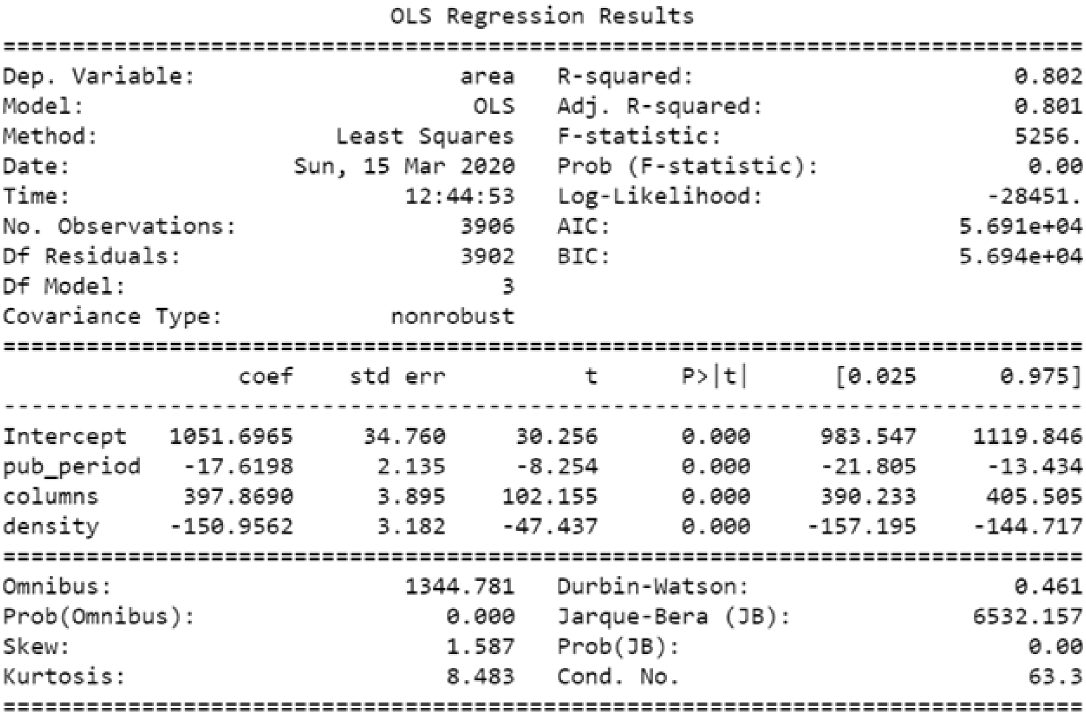
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                columns   R-squared:                       0.826
Model:                            OLS   Adj. R-squared:                  0.826
Method:                 Least Squares   F-statistic:                     6179.
Date:                Mon, 11 May 2020   Prob (F-statistic):               0.00
Time:                        15:11:39   Log-Likelihood:                -3893.7
No. Observations:                3906   AIC:                             7795.
Df Residuals:                    3902   BIC:                             7820.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -59.7816      1.111    -53.811      0.000     -61.960     -57.604
year           0.0334      0.001     55.942      0.000       0.032       0.035
pub_period    -0.0364      0.004     -9.158      0.000      -0.044      -0.029
area           0.0010   1.81e-05     53.800      0.000       0.001       0.001
==============================================================================
Omnibus:                      228.197   Durbin-Watson:                   0.616
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              485.773
Skew:                          -0.390   Prob(JB):                    3.28e-106
Kurtosis:                       4.542   Cond. No.                     2.80e+05
==============================================================================
```

Figure 4.28: Summary of OLS regression analysis for predicting Columns with three features

The intercept is -59.7816, where as slopes for year, publication period and area are 0.0334, -0.0364 and 0.0010 respectively. Since the p-value is really small, it suggests that there is a relationship between independent variables and dependent variable. The Adjusted R square gives a value of 0.826, which suggests that 82.6% of the variance in the data is explained with this model. Comparing this model with the previous one, the results improved significantly. By using the intercept and slopes, Columns can be estimated as

$$columns = -59.7816 + 0.0334 * Year - 0.0364 * Publication\_Period + 0.0010 * Area \tag{4.9}$$

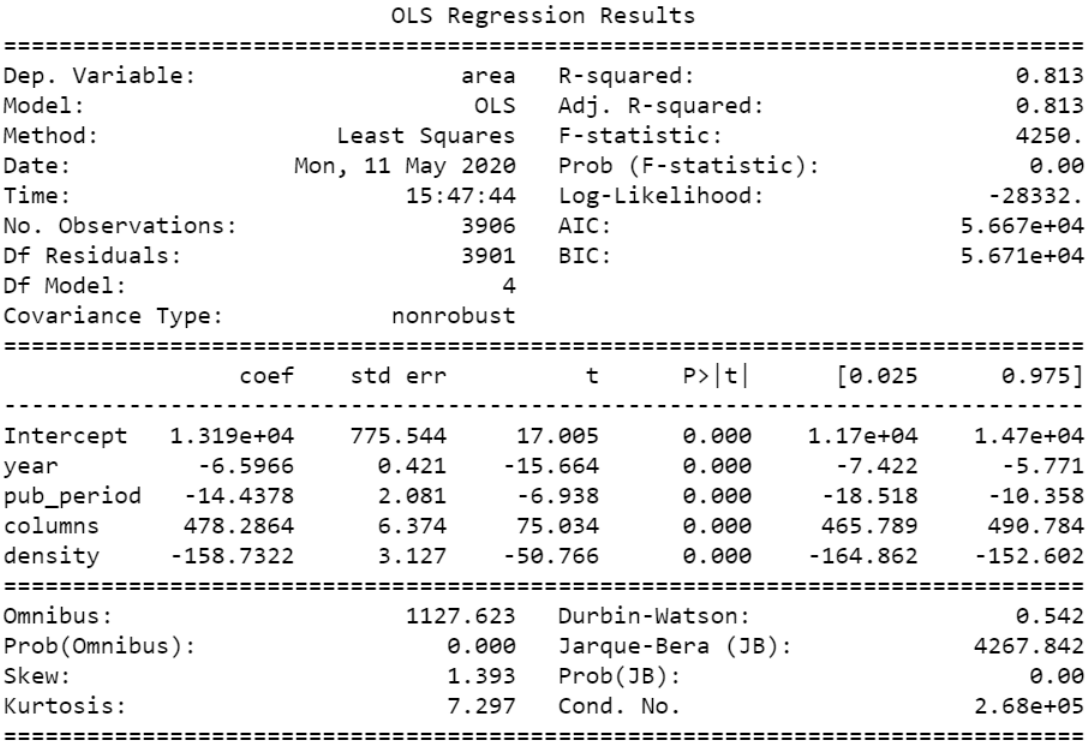
Similarly, year also has a strong positive correlation with area. But in this scenario other features such as publication period, columns and density are fist used as explanatory variables to estimate area. The result below shows that 80.1% of the variance can be explained with this model. The model shows really small p values rejecting null hypothesis.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.802
Model:                            OLS   Adj. R-squared:                  0.801
Method:                 Least Squares   F-statistic:                     5256.
Date:                Sun, 15 Mar 2020   Prob (F-statistic):               0.00
Time:                        12:44:53   Log-Likelihood:                -28451.
No. Observations:                3906   AIC:                         5.691e+04
Df Residuals:                    3902   BIC:                         5.694e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1051.6965     34.760     30.256      0.000     983.547    1119.846
pub_period   -17.6198      2.135     -8.254      0.000     -21.805     -13.434
columns      397.8690      3.895    102.155      0.000     390.233     405.505
density     -150.9562      3.182    -47.437      0.000    -157.195    -144.717
==============================================================================
Omnibus:                     1344.781   Durbin-Watson:                   0.461
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6532.157
Skew:                           1.587   Prob(JB):                         0.00
Kurtosis:                       8.483   Cond. No.                         63.3
==============================================================================
```

Figure 4.29: Summary of OLS regression analysis for predicting Area with three features

Using the same model as above, area is estimated with addition of year as an explanatory variable. As in the case of columns, the objective here is also to study the impact of year in estimating area as it showed strong positive correlation. The result below shows slight improvement as compared to previous model. The OLS model has an adjusted R-square value of 0.813, which is 1% higher as compared to the model not using year as explanatory variable.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   area   R-squared:                       0.813
Model:                            OLS   Adj. R-squared:                  0.813
Method:                 Least Squares   F-statistic:                     4250.
Date:                Mon, 11 May 2020   Prob (F-statistic):               0.00
Time:                        15:47:44   Log-Likelihood:                -28332.
No. Observations:                3906   AIC:                         5.667e+04
Df Residuals:                    3901   BIC:                         5.671e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1.319e+04    775.544     17.005      0.000    1.17e+04    1.47e+04
year          -6.5966      0.421    -15.664      0.000      -7.422      -5.771
pub_period   -14.4378      2.081     -6.938      0.000     -18.518     -10.358
columns      478.2864      6.374     75.034      0.000     465.789     490.784
density     -158.7322      3.127    -50.766      0.000    -164.862    -152.602
==============================================================================
Omnibus:                     1127.623   Durbin-Watson:                   0.542
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4267.842
Skew:                           1.393   Prob(JB):                         0.00
Kurtosis:                       7.297   Cond. No.                     2.68e+05
==============================================================================
```

Figure 4.30: Summary of OLS regression analysis for predicting Area with four features

The equation to estimate area is given as

$$\begin{aligned} Area = 1.319e+04 - 6.5966*Year - 14.4378*Publication\_Period \\ +478.2864*Columns - 158.7322*Density \end{aligned} \tag{4.10}$$

The above models explain the correlation results in more depth. The results show that year can be used to estimate columns and when more features are added, the dependent feature columns can be estimated with more accuracy. On the other hand to estimate area, a combination of features prove to be good instead of a single feature. Area can be used alone to estimate density as they have a negative correlation, which is discussed in more detail in the next section.

4.4 Linear Regression

The correlation table above shows a negative correlation between density and area. To study the relationship among these two, a simple linear regression model [15] is applied after scaling the variables in which density is considered as a dependent variable and area as an independent. The intercept and coefficient of the model are given as

$$Density = -0.01775159 - 0.34731312*Area \tag{4.11}$$

A line is fit to the data, which captures the behavior of the data but still it does not prove to be a good fit as it can be seen in the figure below.

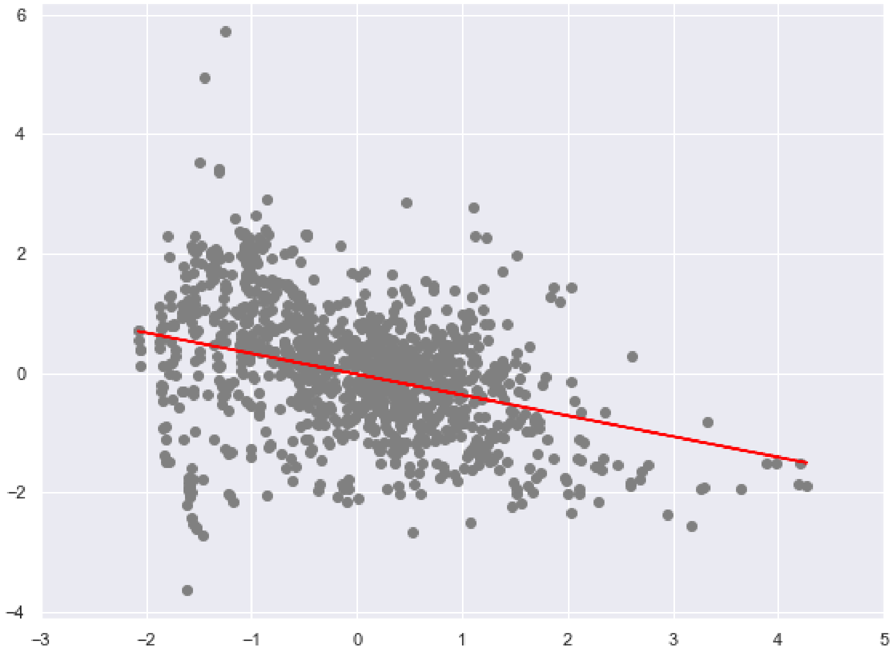


Figure 4.31: Area and Density linear fit

Three different metrics are used to calculate the accuracy of this model [15]. The result shows root mean square error of 0.98, which proves that this algorithm is not very accurate but can still make reasonably good predictions. The scores of these metrics are given in the following table.

Table 4.3: Linear Regression Results

Metrics	Scores
Mean Absolute Error	0.7345
Mean Squared Error	0.9620
Root Mean Squared Error	0.9809

4.5 Random Forest

The regression algorithms discussed above are good estimates to find the coefficient values of independent variables for predicting the dependent variables, but an important aspect of this research work is to study the development of features.One important feature throughout this research is proven to be the area of newspaper. Since it has linear relationships with columns, density, and publication period, a case study emerges to analyze the changes in these features which led to change the area sizes. This scenario fits well with the concept of decision tree.

A decision tree maps the possible outcomes of a series of related choices. Its representation is a flow-chart diagram where nodes represent conditions and based on their outcome, branches are derived. The leaf nodes represent the labels or classes of a series of decisions. For this case study, Random Forest algorithm has been applied as it forms multiple decision trees and merges them for a stable and accurate prediction.

Random forest is a supervised learning method, which is an ensemble of multiple decision trees. The decision trees are generated with Bagging method, which is a technique to improve the accuracy and stability of the machine learning methods by reducing variance. Bagging works by taking random samples with replacement from the data and training the model on each sampled data. Since this method is taking samples with replacement, there is always the possibilities of finding the similar structures in the trees, which can affect the outcome. Random forest is an extension of this method, which generates de-correlated trees. The method increases the diversity as it has selected random data points with a random subset of features to create trees and predict a class. In the end, the class with majority vote is predicted, as shown below.[18]

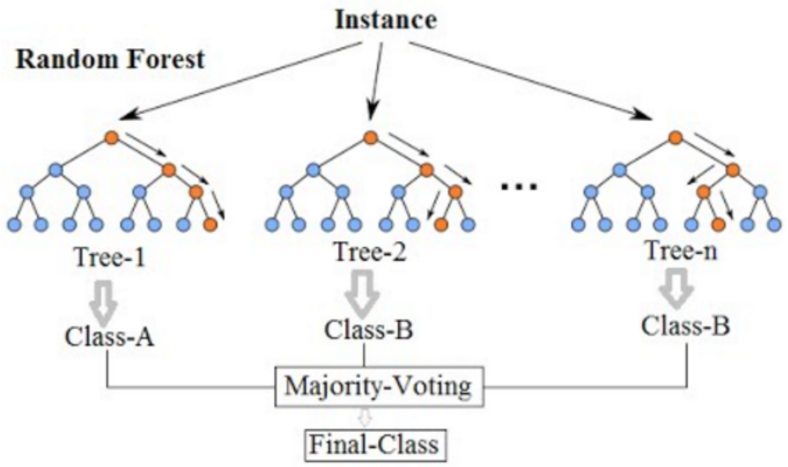


Figure 4.32: Example of Random Forest Classifier [18]

Using this algorithm a random forest is created. The algorithm divides each node in exactly two more nodes. The nodes are divided based on the best numerical or categorical feature, which is calculated by impurity criterion. Gini impurity has been applied for this case study which indicates the probability p_i of classifying a randomly selected data point

Algorithm 1 Random Forest Algorithm [13]

```
for b = 1 to B do
    Draw a bootstrap sample n of size N from the data
    until the minimum node size $\mathrm{n_{min}}$ is reached, recursively repeat the following
      steps to generate a random-forest tree $\mathrm{T_b}$
      i. Select m variables at random from the p variables
      ii. Pick the best variable/split-point among the m
      iii. Split the node into two child nodes
    Output the ensemble of trees $\{T_\mathrm{b}\}_1^B$
    To predict a new x:
      i. Regression: $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\ \sum_{b=1}^{B} T_b(x)$
      ii. Classification: Let $\hat{C}_b(x)$ be the class prediction of the $b_{th}$ random-forest
      tree. Then $\hat{C}_{rf}^{B}(x)$ = majority vote $\{\hat{C}_b(x)\}_1^B$
```

with the wrong class. The formula to calculate gini impurity is [13]

$$G(j) = \sum_{i=1}^{k} p_i(1 - p_i) \tag{4.12}$$

The objective of applying Random Forest is to study the changes in various features leading a change in area size. Since the area ranges from 119 to 5553 tenth of a mm^2, applying any decision tree algorithm on this feature will not help in understanding and interpreting the results as area feature holds a wide range of values. In the beginning, it has been discussed that using area, a new feature ‘Page Size’ is generated which represents the area in the form of paper size with four different categorical values A5, A4, A3 and A2. So the random forest algorithm is applied on the features year, publication period, pages, density, and columns to classify the data into page size(s). Data is divided into training and testing sets with an 80-20 split.

Random forest built by scikit-learn python library* is utilized here. 100 total trees are generated to build a random forest. The results below show decisions on the basis of features and based on these decisions a page size is predicted. For the representational purpose, only three trees out of hundred are shown below. Below is the first tree generated by the algorithm. This tree indicates that a newspaper is supposed to publish on an A5 page size with a probability of 0.54, when year is less than 1846 and columns are less than 4 but if the year is greater than 1845, newspaper published on an A4 page with a probability of 0.56. If columns are 4 or 5, newspaper is supposed to publish on an A3 page whereas with columns more than 5, an A2 page size will be used. From this tree it is obvious that a decision to publish on an A2 or A3 is purely based on number of columns.

*https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

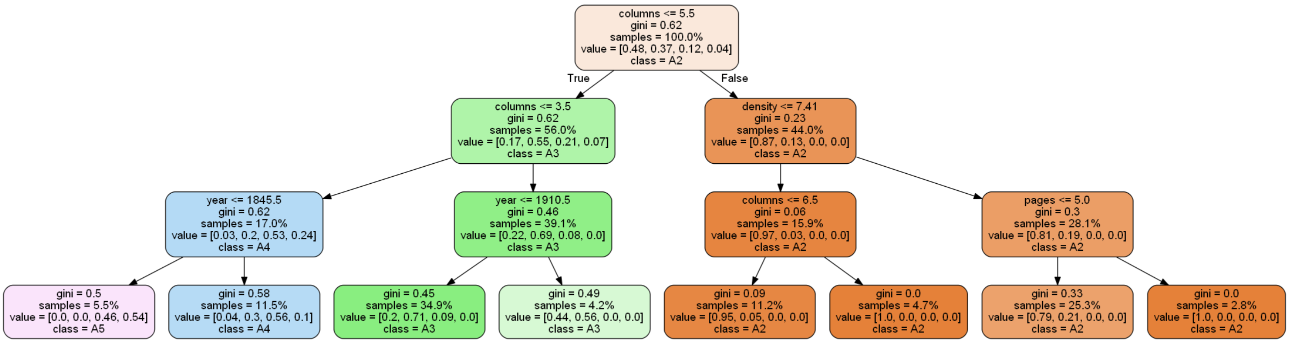


Figure 4.33: First Decision Tree to classify Page Size

Below is shown tree number fifty, which has different paths as compared to the above one to specify a page size. It shows if publication period is greater than 2, pages are more than 7 and columns are less than 3, with a probability of 0.84, the newspaper will use A5 page but if pages are less than 8 with 3 or less columns newspaper will use A4 otherwise A3. Similarly, if columns are less than 6, A3 page will be used otherwise A2.

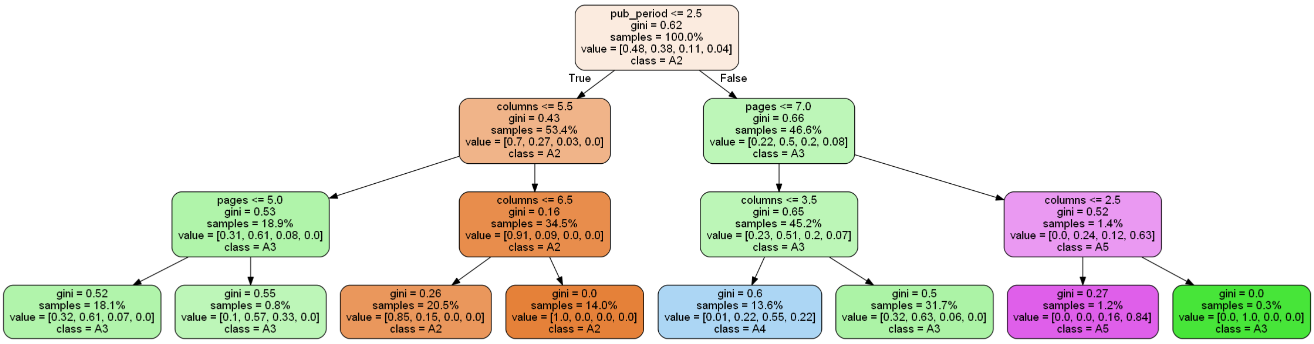


Figure 4.34: Fiftieth Decision Tree to classify Page Size

The last shown tree is tree number 100. This tree has different paths as compare to previous ones to specify a page size. The result of this tree can also be interpreted the same way as above. The objective of showing three trees was to understand the concept behind random forests algorithm.

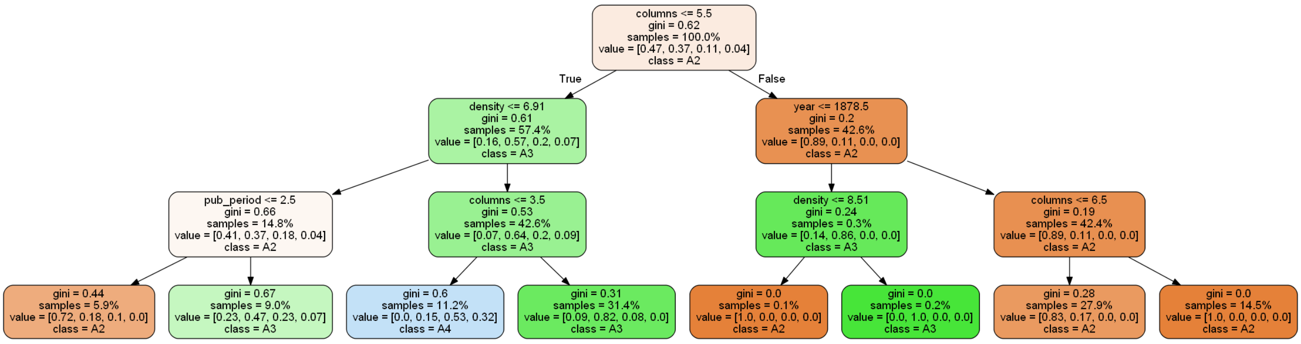


Figure 4.35: Hundredth Decision Tree to classify Page Size

The above shown trees are generated with a depth of three for the representational purpose. To analyze the impact of tree depth on the accuracy, the model is trained with tree depth from 2 to 10. The result below shows that with the depth of two, the model's accuracy is 0.74, but as the depth increases, model's accuracy also increase. When the tree is generated with a depth of 7, the model's accuracy is 0.81 and after that with the increase in depth, model's accuracy remains at 0.81. This depth is the optimal level to achieve maximum accuracy.

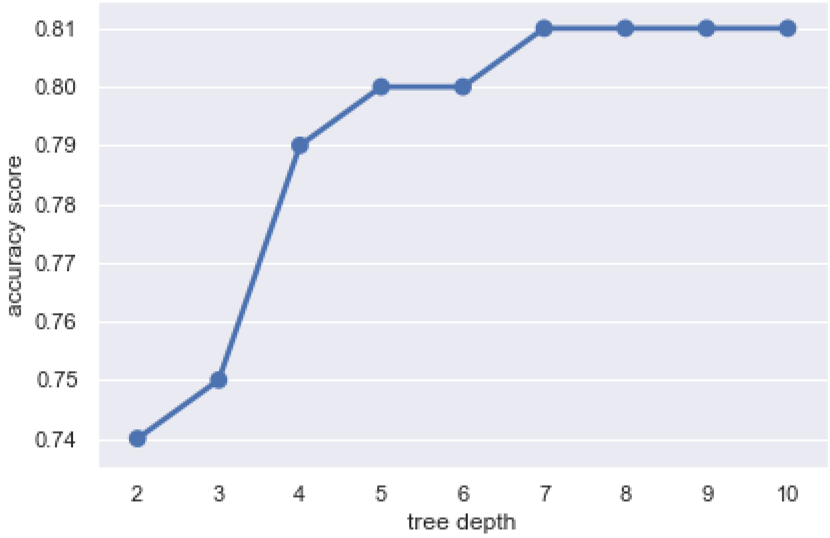


Figure 4.36: Accuracy score with different depth levels

Using the result from the above analysis, random forest is generated with a depth of 7 and the results of the model are given below for each page size. The precision and recall values of A2 are higher than other page sizes. The model is good with predicting A2 overall. Page size A4 has the lowest precision among all, with a ratio of 0.68, though it is still acceptable in general.

Table 4.4: Random Forests Results

	Precision	Recall	F1-Score
A2	0.88	0.86	0.87
A3	0.75	0.74	0.75
A4	0.68	0.75	0.72
A5	0.76	0.96	0.85

The last and an important aspect of these analysis is to find out the important features when classifying page size. For that purpose, using the built-in method of random forests algorithm, features' importance is shown below. The result shows that columns is most important feature to classify page size. After that are year, density and publication period. Number of pages is the least important feature to classify the page size. Hypothetically, it should have been an important feature, but as we have observed in the feature analysis earlier in this chapter that the dominating number of pages were four throughout the time, so this feature does not affect the page size.

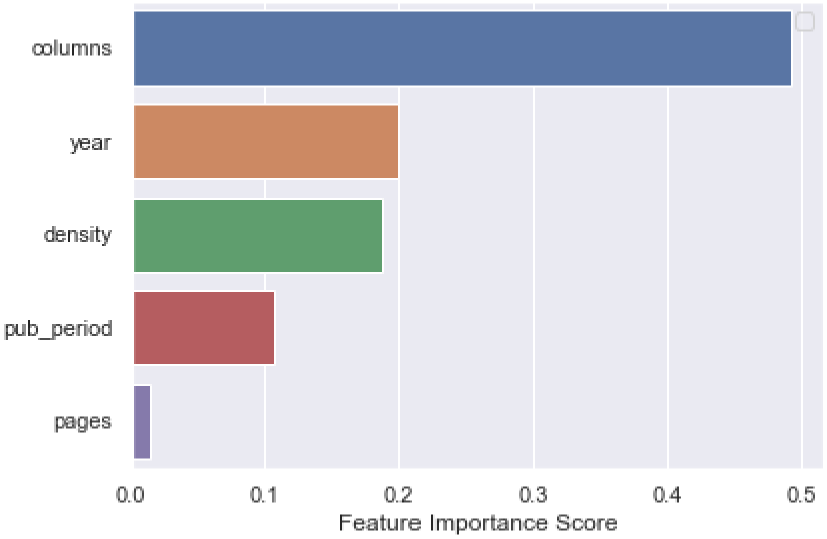


Figure 4.37: Feature importance in classifying page size

5. Feature Clustering

The features have been studied individually and jointly, regression methods have been applied on them, and decision tree has also been created to see the divisions of the features but the case study either there exist some groups within the data or not is still unresolved. To create the groups of data points, which share similarity, clustering is the best approach.

The data points in one cluster are close to other data points in the same cluster but different from the data points in another cluster. Since there exist various methods to find the similarities among the data points before clustering them, so it depends on the data, which clustering algorithm suits better in a situation. For this case study, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), K-means and Spectral clustering algorithms are considered, and based on the initial analysis, spectral clustering algorithm seems a good choice. Before discussing spectral clustering in detail, first DBSCAN and K-means are discussed briefly and their results are evaluated.

5.1 K-means Clustering

K-means is the simplest and a widely used clustering algorithm. It is well suited for large scale datasets because of its computational speed. It partitions n observations into k clusters, such that each observation belongs to the cluster based on the observation's proximity to the mean of the cluster. After the observations is associated with a cluster, the cluster's mean is recomputed and the process continues. The algorithm has four basic steps: 1) arbitrarily select k points as the initial cluster centers. 2) Observations in the dataset are assigned to the closest cluster by using Euclidean distance among the observations and cluster centers. 3) Cluster centers are recomputed. 4) Steps 2 and 3 continue until no cluster changes anymore.[27] [22] [34]

A key point in the implementation of K-means algorithm is to provide the number of clusters. Generally, while solving unsupervised problems, knowing the number of clusters in advance is not possible though there exist several methods to choose optimal number of clusters. One such method is elbow method. K-means algorithm is run for a range of values of k, and Sum of Squared Errors (SSE) are calculated for each k [22]. The result of

SSE is plotted against k and the optimal k is chosen based on the SSE. The objective of this method is to choose k for which SSE are small enough. Though if we keep increasing k, in the end each data point will be a separate cluster with the lowest SSE, but generally when the SSE starts decreasing linearly, that value of k is chosen as optimal.

Publication period, area, pages, columns and density features are selected to apply k-means algorithm* with k ranging from 1 to 15. The data is scaled using minmaxscaler† before applying this method. Below is the result indicating SSE values against different k values. The result shows that the SSE decreased exponentially in the beginning but when k is 4, it slows down.

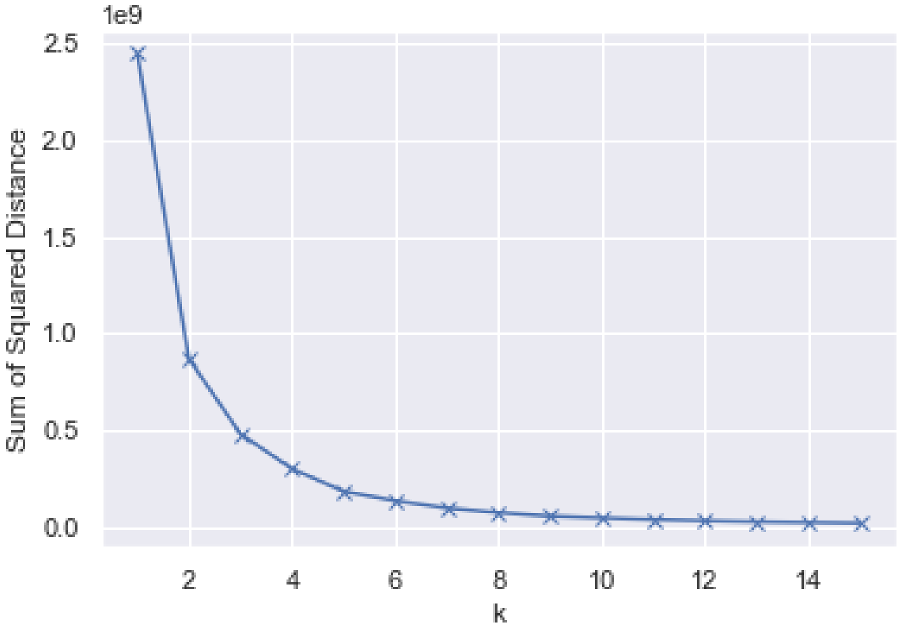


Figure 5.1: Optimal number of clusters using Elbow method

Using k as 4, K-means algorithm is applied on the scaled dataset. The result below shows the number of observations in each cluster. The algorithm puts more than 3000 records out of 3906 in two clusters, and 173 observations in fourth cluster. Though mathematically there is nothing wrong with this division, but with the domain knowledge and early exploratory analysis, this clustering does not seem the correct division of records based on the widespread data.

*https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.htmlsklearn.cluster.KMeans

†https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

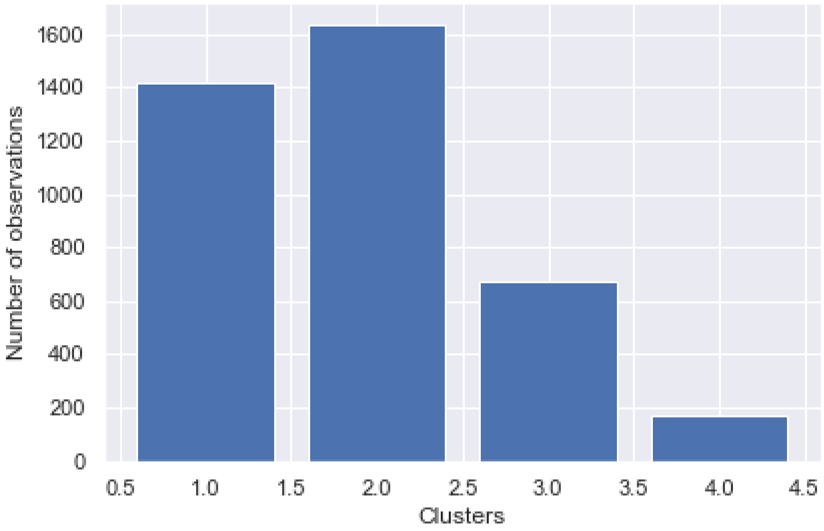


Figure 5.2: Number of observations in each cluster

Reading through the literature, various limitations of K-means appear to be the probable reasons for this division. K-means does not yield the same results in each run because the resulting clusters are dependent on the initial arbitrary assignments of centroids. If the clusters are of a spherical shape (radius of the cluster is equal to the distance between the centroid and the farthest data point), this method has the capability to capture structure of the data but soon as clusters have a complex geometric shape, k-means' performance start to decrease. K-means is not robust in clustering data where clusters are of varying sizes and density. An important limitation of k-means is the sensitivity to outliers. When outliers are present, the resulting cluster centroids may not represent the data correctly.[34]

By comparing the data against the above mentioned limitations, it can be seen why K-means algorithm is not being able to cluster the data effectively. The result in (Fig. 5.3) confirms that the data is not of spherical shape, and when k-means applied, it could not find optimal centroids for clustering. The red circles are representing centroids of the clusters. Though the algorithm is applied on five features but to understand the limitations of this method, the result is visualized in two dimensions. For the first visualization, area and number of pages are selected and the result indicates that the centroids of all four clusters are not optimal. As already mentioned, one reason for this result could be the varying sizes of clusters and density. A very few number of observations have number of pages two and area under a specific range. Most newspapers published four pages and the area for these newspapers vary from 1000mm^2 to 5000mm^2. Visualizing k-means clustering for columns and publication period in (Fig. 5.4) also proves the above discussed limitations of this method.

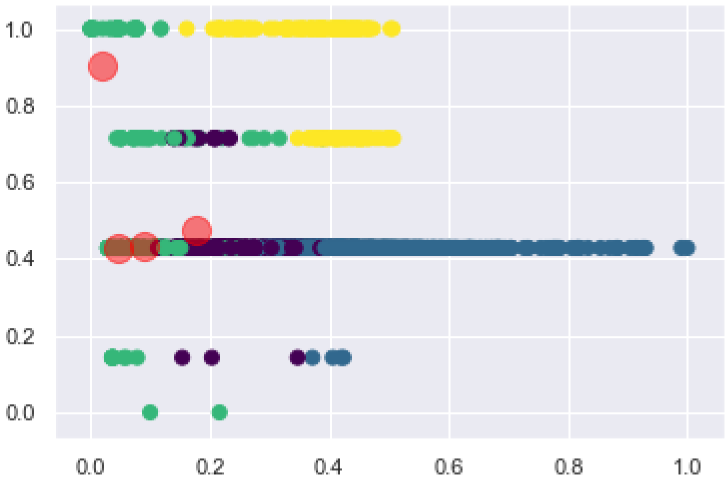


Figure 5.3: Cluster results of Area and Pages

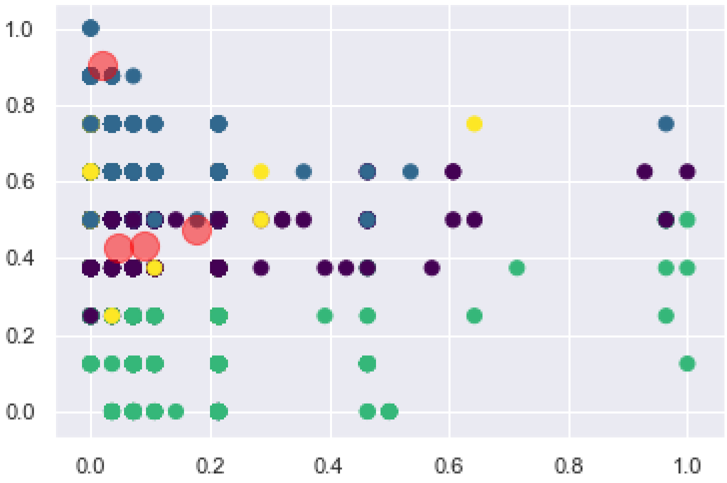


Figure 5.4: Cluster results of Columns and Publication Period

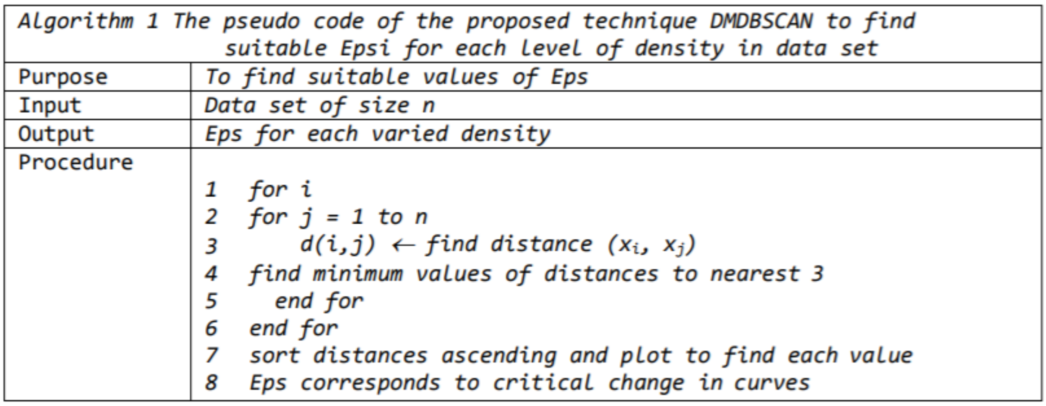
Another reason for this method's failure on the dataset is the presence of outliers in the data. For example a newspaper publishing four pages every day on an A3 page size is the normal trend for a cluster. A newspaper in the data with the same values for all features except publication period, publishing once a week, clustering this data for k-means is not possible, as publication period 7 will be considered as an outlier. The algorithm is applied with k=3 and k=5 also but the structure of this data makes k-means not suitable for the clustering analysis on it.

5.2 DBSCAN Clustering

Since k-means is a centroid-based method, which did not perform well with the given dataset, another approach which is density-based is considered for comparative analysis. DBSCAN [8] is the most used density-based clustering algorithm. It has two basic parameters, ϵ and minimum number of data points, *minPts*, in the neighborhood. This ϵ value specifies how close two points need to be to be considered part of a cluster. If any two points have a distance less than or equal to this parameter, they are considered as neighbors. The second parameter is used to form a dense region. If the value of second parameter is 25, then at least 25 points are needed to form a dense region. DBSCAN categorizes the observations into three categories. Core points, which have more than *minPts* within epsilon distance. Border points, which have fewer than *minPts* within epsilon distance but still are in the reach of core points and Noise points which are not reachable from any other point. DBSCAN starts by selecting a random point which is not assigned to a cluster or has been identified as an outlier. Then the algorithm computes selected point's neighborhood to determine if it's a core point. If yes, it starts a cluster around this point otherwise it designates the selected point as an outlier. After finding a core point and a cluster, the algorithm expands the cluster by adding all border points. If it adds an outlier, it changes that point's status from outlier to border point. These steps are repeated until each point is either assigned to a cluster or designated as an outlier. [8][6] [31] [4]

As the name suggests, DBSCAN performs well by finding areas in the data that have a high density of observations, as compared to the areas of the data that are not very dense. The key advantage of this method over other clustering methods is that it does not require a-priori specification of number of clusters. DBSCAN infers the number of clusters based on the data, and it can discover clusters of arbitrary shape. Since this algorithm is resistant to noise it can handle clusters of arbitrary sizes and shapes. By looking at the characteristics of this algorithm, it solves almost all the issues arise when using K-means algorithm in the previous section. [4][8][6]

Though DBSCAN algorithm does not need to know the number of clusters in advance, but it requires ϵ to find the data points within the distance of this value for clustering. To find the optimal value of [7] has proposed a method. The algorithm as defined below is very simple. It calculates the distances between data points on each pair of latitude and longitude data for three nearest neighbors. In the next step the distances are sorted in ascending order and the result is plotted. The plot is supposed to show the sharp change in the graph just like elbow method as discussed above. This sharp change at the value of k-distance corresponds with an optimal value of ϵ. [29][7]



Algorithm 1 The pseudo code of the proposed technique DMDBSCAN to find suitable Epsi for each Level of density in data set

Purpose	*To find suitable values of Eps*
Input	*Data set of size n*
Output	*Eps for each varied density*
Procedure	

```
for i
for j = 1 to n
    d(i,j) ← find distance ($x_i$, $x_j$)
find minimum values of distances to nearest 3
  end for
end for
sort distances ascending and plot to find each value
Eps corresponds to critical change in curves
```

Figure 5.5: Algorithm to find optimal ϵ [7][29]

The authors have provided the result of this algorithm as shown below. The result suggests the optimal value of ϵ is 0.4194.

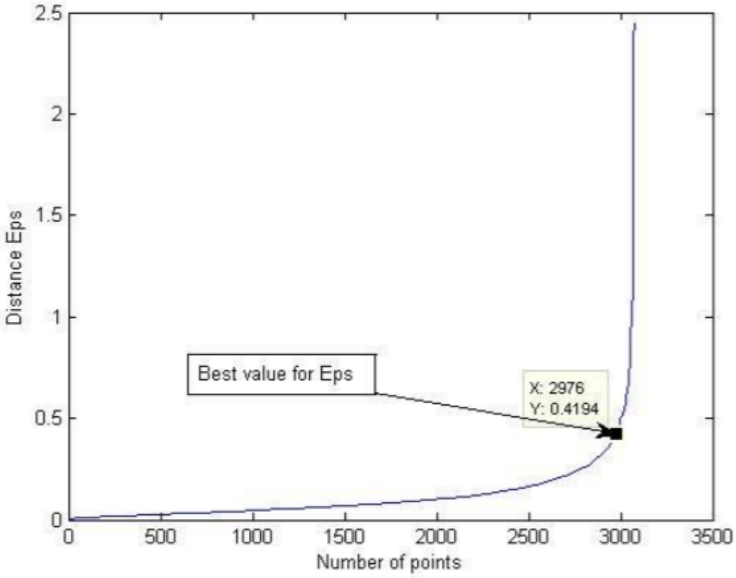


Figure 5.6: Result of the above algorithm to find optimal ϵ. [7]

Above discussed algorithm has the same behavior as k-nearest neighbor method. NearestNeighbor* method implemented in scikit-learn is an unsupervised learning algorithm which acts as a uniform interface to k different nearest neighbors. This algorithm is applied to find Euclidean distance on the scaled dataset for k ranging from 3 to 100. Then the results are sorted and plotted as shown below for k=100. This result suggests that the optimal value of ϵ is between 0.5 and 0.7. After that there is a sharp change in the graph.

*https://scikit-learn.org/stable/modules/neighbors.html

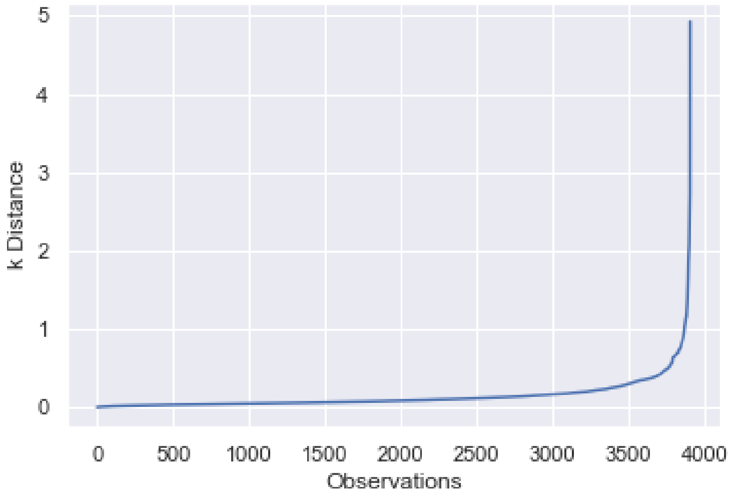


Figure 5.7: Above algorithm applied on the dataset to find optimal ϵ.

Since the optimal value of ϵ is computed, DBSCAN* algorithm is applied with multiple values of ϵ ranging from 0.5 to 0.7 with 100 *minPts*. The result below shows when ϵ is 0.5 more than 2000 records are classified as outliers and total four clusters are identified. When the values of ϵ are 0.55 and 0.6, total five clusters are identified but still the number of outliers is very large. With ϵ changing to 0.65 and 0.7, the number of outliers decreased slightly, but in this case total two clusters are identified.

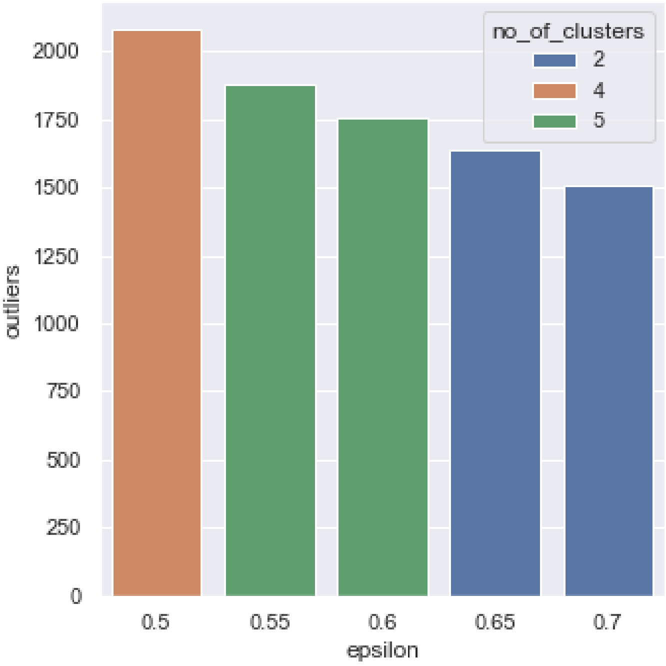


Figure 5.8: Number of outliers and number of generated clusters for selected ϵ.

*https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Since the above studied algorithm could not perform well with this dataset for the selected values of ϵ, exploratory analysis are performed with a wide range of ϵ values from 0.01 to 5. The *minPts* are given as 50 for the first experiment and 100 for the second experiment. The objective is to find an ϵ value for which number of outliers are very low and the number of identified clusters are satisfactory. The term satisfactory here is relevant to the number of outliers. If the algorithm identifies two clusters with only 100 observations as outliers and remaining 3806 in one clusters, the results are not helping in understanding the underlying behavior of the data. Similarly if the number of clusters are seven with 1500 observations identified as outliers and remaining 2100 observations are divided in six clusters, this is also not beneficial. Considering these observations the algorithm is applied with above defined configurations. The result below shows with the increase in ϵ from 0.01 to 1, the number of outliers dropped sharply, but the number of identified clusters are 2 or 3 after that. With the ϵ value of more than 2.5 there are only two clusters identified. The only acceptable values of ϵ seem to be between 1.5 and 2.5 as these values are identifying three clusters with the number of outliers around 250.

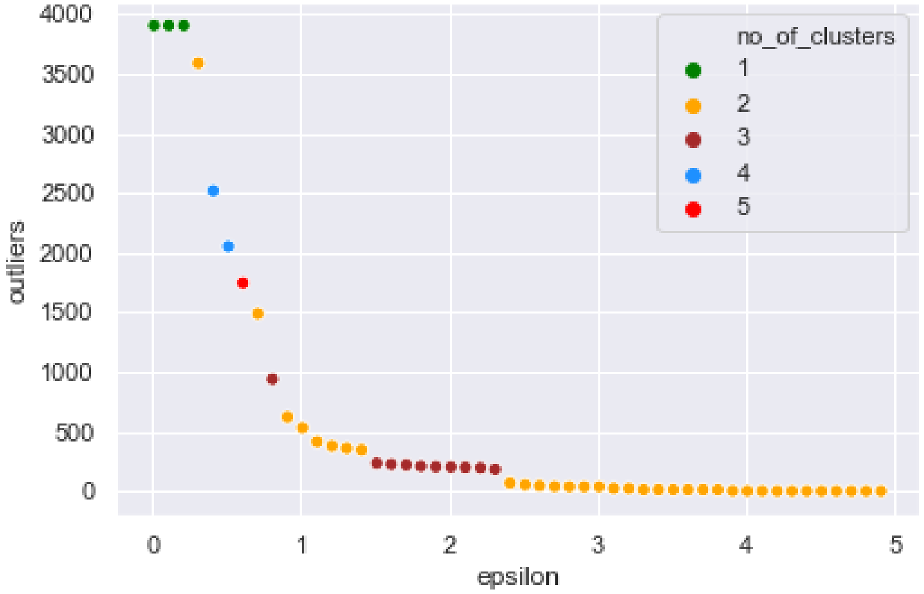


Figure 5.9: Number of outliers and number of generated clusters for selected ϵ.

The above result is for the *minPts* equals to 100. The experiment conducted with *minPts* of 50, did not show any better performance either. Utilizing the output of the above result, the ϵ values between 1.5 and 2.5 and *minPts* equals to 100 is used for further analysis. The figure below shows that the selected values of ϵ identified three clusters and the number of outliers are also less than 200, but the clusters are not useful as more than 3000 records are identified as one cluster. With the increase in ϵ value, the number of observations in one clusters increases, which does not help in our analysis of the data.

While going through the literature to understand the reasons DBSCAN is not performing well with this dataset, it appears that this method faces troubles when data has varying densities [6]. As the two main parameters of this method are ϵ and *minPts*, which cannot be chosen aptly for all the clusters in case of varying densities. If the data is too sparse and densities vary, this algorithm fails to identify clusters, which seem the plausible reasons of the failure of this method on the given dataset. The algorithm is run for *minPts* 10 to 100, and with a wide range of ϵ values, but the output of this algorithm is not satisfactory in any set of configurations. With the increase in *minPts*, the observations fall in one clusters and as the *minPts* decreases, observations start making their own clusters. With any value of ϵ, algorithm either identifies more than half observations as outliers or it designates more than half observations in a single cluster.

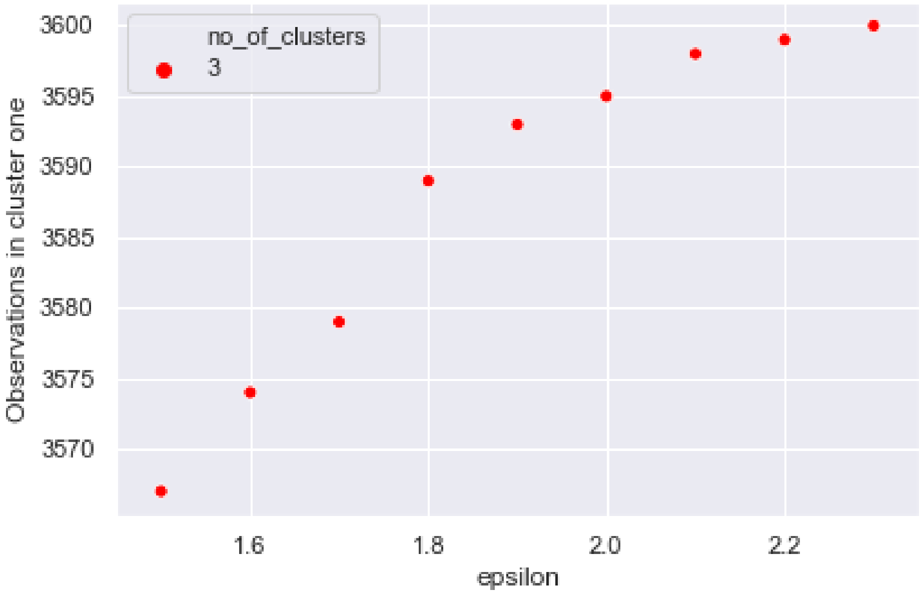


Figure 5.10: Number of observations in one clusters for selected ϵ.

The results of the centroid-based and density-based algorithms helped in shifting the approach to solve clustering problem on this dataset. As an alternate approach to the above algorithms, Spectral clustering is studied and based on initial exploratory analysis, it seems to give good results. This method is studied in detail in the next section.

5.3 Spectral Clustering

Spectral clustering is a popular algorithm for clustering because of its implementation's simplicity. It is a useful approach when the structure of clusters are non-convex. Spectral Clustering originates from graph theory where the communities are built based on the edges of a node, but it is not limited to the graph data [20]. As spectral clustering is a similarity-based algorithm, it suits the problem in hand where the objective is to find the similarities among the newspapers published in different languages, in different times, and from different regions. Since spectral clustering does not make strong assumptions about the shape of data, and inherently it works good for sparse data, this looks the most suitable method for clustering the data under study. The algorithm has three basic steps:

Build a Similarity Graph

An adjacency matrix A, is built representing a similarity graph using one of the two possible options. The epsilon-neighborhood graph or K-nearest neighbor [20]. In the epsilon-neighborhood method, a parameter epsilon is fixed first and then distance of each data point from the epsilon value is calculated. A data point is connected to those which exist in it's epsilon radius. In the K-nearest neighbor, a parameter k is fixed beforehand, then for any two vertices u and v, an edge is directed from u to v if v is among the k-nearest neighbors of u. These were just simple explanation of the methods. [20]

Data Projection on Lower Dimensional Space

To project the data onto a lower dimensional space, laplacian matrix is needed. The rationale for a lower dimensional space is that there are possibilities that the data points of the same cluster can be far away from each other in a large dimensional space. With the lower dimensional space, those data points can be closer and will be clustered together [20]. To compute laplacian matrix, degree of a node is required, which can be computed using this equation

$$d_i = \sum_{j=1}^{n} W_{ij} \tag{5.1}$$

where w_{ij} is the edge between node i and node j [20]. Using this formula, a degree matrix D, can be derived as follows

$$Dij = \begin{cases} d_i, i = j \\ 0, i \neq j \end{cases} \tag{5.2}$$

To compute laplacian matrix, diagonal matrix and adjacency matrix are used as

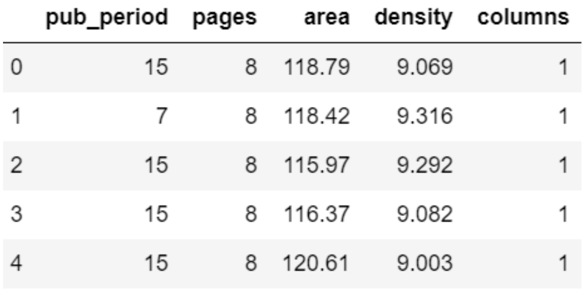
$$L = D - A \tag{5.3}$$

where A is the Affinity matrix calculated in the first step [20].

Using this matrix, eigenvalues and eigenvectors are calculated and depending on the k number of clusters, first k eigenvalues and respective eigenvectors are arranged in a matrix. The matrix has eigenvectors as columns. The details of calculating eigenvectors and eigenvalues are beyond the scope of this research work but are followed here [2].

Clustering the Data

To start clustering, five features publication period, pages, area, density and columns are selected. Below is a glimpse of how the data looks.



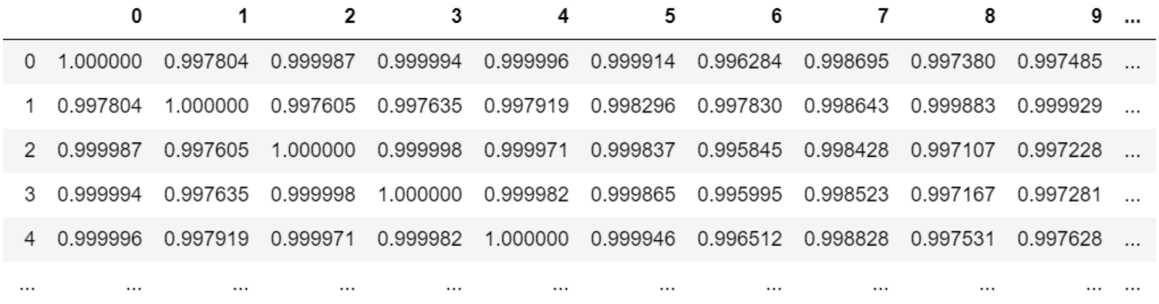
	pub_period	pages	area	density	columns
0	15	8	118.79	9.069	1
1	7	8	118.42	9.316	1
2	15	8	115.97	9.292	1
3	15	8	116.37	9.082	1
4	15	8	120.61	9.003	1

Figure 5.11: Preview of data for selected features

Since the spectral clustering algorithm needs a similarity graph, cosine similarity is used. The goal of using cosine similarity is to compute similarities between observations. It calculates the cosine of the angle between two n-dimensional vectors in an n-dimensional space. Cosine similarity determines whether two data points are pointing in the same direction. The formula to compute cosine similarity is given as

$$Cosine\_Similarity = \frac{A.B}{|A||B|} \tag{5.4}$$

where A and B are two vectors. The cosine similarity[*] method implemented in scikit-learn python is used here. This method resulted in a 3906 x 3906 matrix as shown below.



	0	1	2	3	4	5	6	7	8	9	...
0	1.000000	0.997804	0.999987	0.999994	0.999996	0.999914	0.996284	0.998695	0.997380	0.997485	...
1	0.997804	1.000000	0.997605	0.997635	0.997919	0.998296	0.997830	0.998643	0.999883	0.999929	...
2	0.999987	0.997605	1.000000	0.999998	0.999971	0.999837	0.995845	0.998428	0.997107	0.997228	...
3	0.999994	0.997635	0.999998	1.000000	0.999982	0.999865	0.995995	0.998523	0.997167	0.997281	...
4	0.999996	0.997919	0.999971	0.999982	1.000000	0.999946	0.996512	0.998828	0.997531	0.997628	...
...	...	...	...	...	...	...	...	...	...	...	...

Figure 5.12: Cosine similarity

Once the data is prepared, Spectral clustering provided by scikit-learn[†] is applied on this data set and it applies clustering to a projection of the normalized laplacian.

[*] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine$_{s}imilarity.html$

[†] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html

Algorithm 2 Spectral Clustering Algorithm [20] [25]

```
Construct a similarity graph by one of the methods, epsilon-neighborhood or K-nearest neighbor
Let A be its weighted adjacency matrix
Compute the normalized Laplacian L
Compute the first k eigenvectors $v_1, . . . ,v_k$ of L
Let V $\in \mathrm{R}^{n\times k}$ be the matrix containing the vectors $v_1, . . . , v_k$ as columns.
Form the matrix U $\in \mathrm{R}^{n\times k}$ from V by normalizing the row sums to have norm 1, that is $u_{ij} = v_{ij}/(\sum_k v_{ik}^2)^{1/2}$
For i = 1, . . . , n, let $y_i \in \mathrm{R}^k$ be the vector corresponding to the i$^{\mathrm{th}}$ row of U.
Cluster the points $(y_i)$ with the k-means algorithm into clusters $\mathrm{C}_1, ..., C_k$.
```

Spectral Clustering is performed on the data set with various number of clusters. The reason to go with a number of clusters is to choose which cluster size gives a good accuracy. To compare the results of different cluster sizes, Silhouette Score [14] is calculated. The silhouette score is a measure of similarity of a data point within its cluster and with the data points of other clusters. Silhouette score is calculated as

$$sil\_score = (a - b)/max(a, b) \tag{5.5}$$

where a is the distance between a data point and its nearest cluster, whereas b is the average inter-cluster distance. [14]

The range of silhouette score is between -1 and +1. A score close to -1 indicates that a data point is poorly matched to its own cluster and a value close to +1 shows that a data point well matched to its own cluster. Spectral clustering is performed for cluster size 3 to 10 and a silhouette score is calculated for each cluster size. The result below shows a maximum score of 0.52 on this dataset when the cluster size is 4.

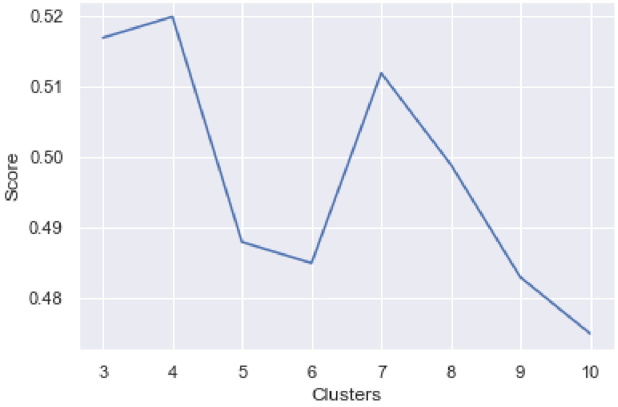


Figure 5.13: Silhouette score to find optimal number of clusters

Following the output of silhouette score, Spectral clustering is performed with 4 clusters. The kernel method is set to be 'Nearest Neighbor' and to create an affinity matrix, number of neighbors set to be 100. The data is grouped in four clusters and the number of data points in each cluster are given in (Table. 5.1)

A visual representation of spectral clustering is given below. Since five features were selected for clustering, and there exist a huge variation within these features, so there is no clear division in the clusters. Result of this clustering can be interpreted by looking at the following plot. Cluster 1 consists of newspapers publishing more frequently with 4 pages and 6 columns. This cluster has a density between 8 and 9 and an area of 2000 which represents A2 paper size. Cluster 2 has papers publishing once or twice a week, with 4 pages and columns less than 5. This cluster has an area around 1000 and density ranging from 7-10. Cluster 3 contains mostly those newspapers which publish frequently on 4 pages, with columns 6 or more,and published on A2 with moderate density. Cluster 4 has papers with wide range of publication period, with 4 pages, 4-5 columns, a high density and published on A3 paper.

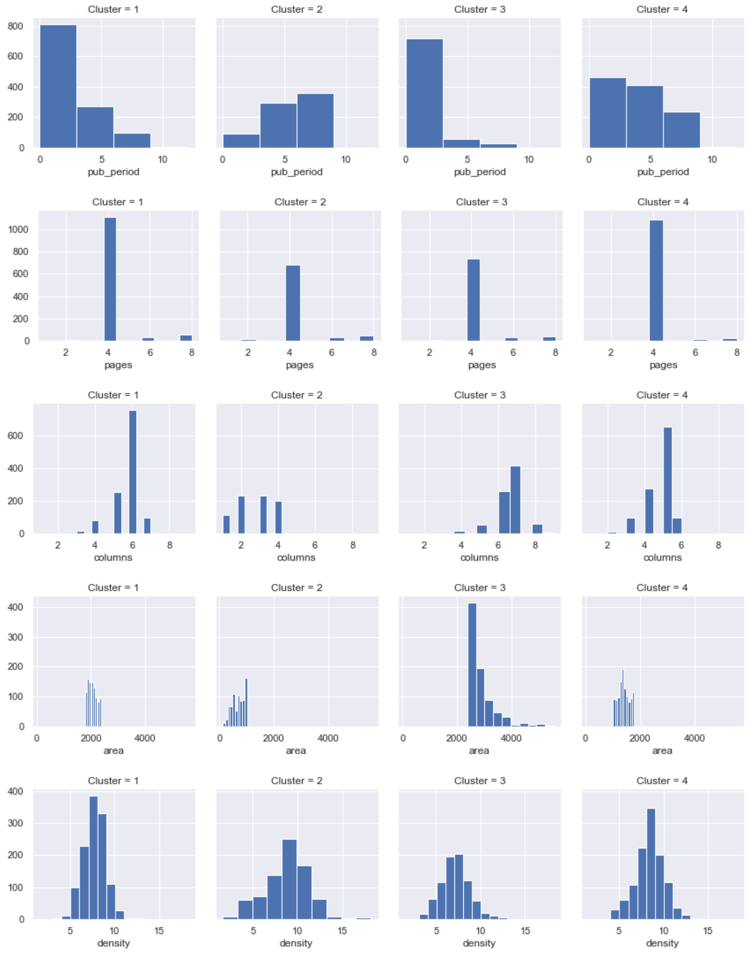


Figure 5.14: Features' comparison for four clusters

Table 5.1: Cluster sizes

Cluster Number	Size
1	1200
2	769
3	805
4	1132

To study the distribution of single variable and its relationship with other variables, pair plot is shown below. Feature Life is also added in this pair plot to understand the relation of variables and newspaper's life. Publication period and pages have clustered data all over the place, but area and density are easily interpretable. With the increase in area, newspapers are publishing more frequently and mostly newspapers fall in the cluster 3. The increase in area also increases number of columns. The newspapers which have moderate density have longer lives as compared to more dense papers. Newspapers which published on 4 pages are showing more density also.

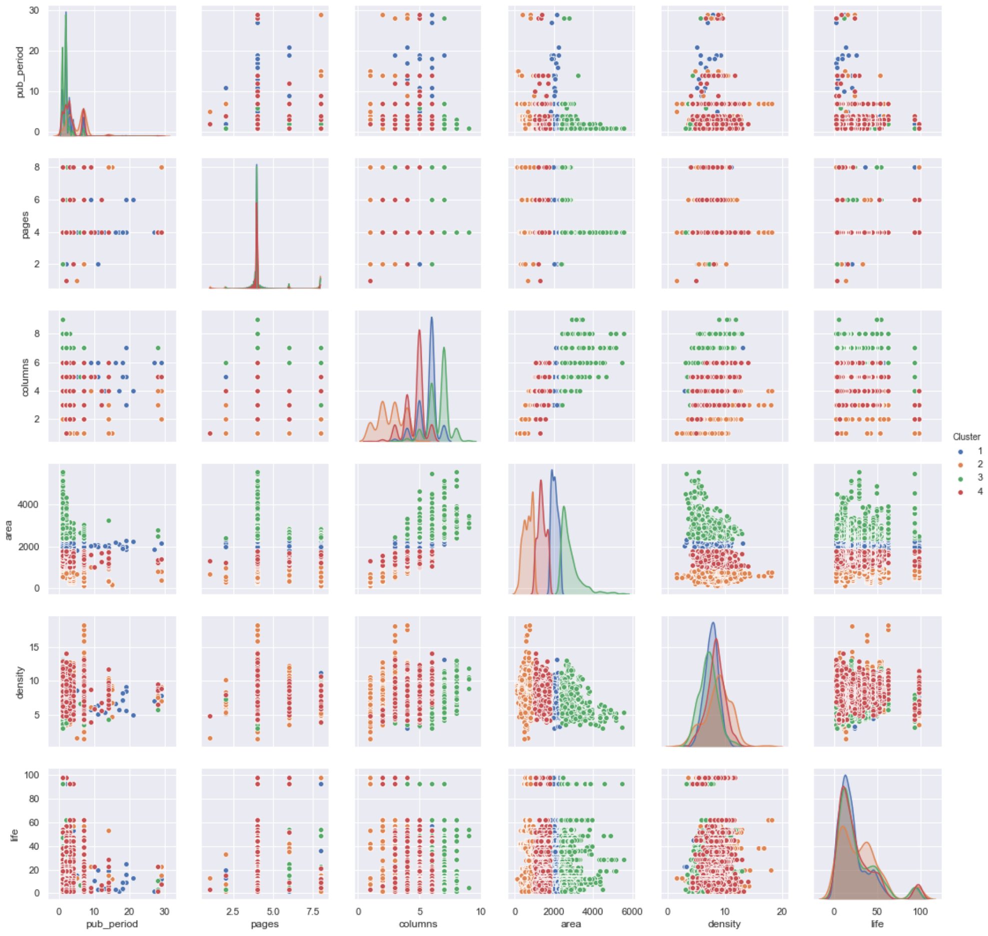


Figure 5.15: Pair plot for feature comparison in four clusters

Since the size of clusters are not evenly divided, so increase or decrease in the size of a cluster, is of focus of interest here. As the clusters' divisions are interpreted above, the increase and decrease in size of clusters will help us understand the changes in features against time. Following figure shows the clusters' size against time. All the newspapers from start are in cluster 2 until 1830, at that time cluster 4 starts emerging. As discussed earlier cluster 2 has newspapers publishing once or twice a week, with 4 pages and less than 5 columns and area less than 1000mm^2. Cluster 4 starts emerging around 1820, indicating an increase in columns and area. Near 1880, cluster 4 starts increasing but cluster 1, and cluster 3 were already emerged at that time. Around 1900, cluster 2 is already vanished with less than 8 newspapers, while cluster 1 and 4 are equally dominating in the trends. Cluster 1 and 4 both started increasing around 1880 and at 1900, they have the same size. Cluster 1 has the papers with almost same features as cluster 4 but the newspapers are publishing more frequently. After 1900 cluster 4 starts decreasing but cluster 1 and 3 are increasing. At the end of first decade of twentieth century cluster 1 and 3 have almost the same size, but soon cluster 1 decreased and cluster 3 is the dominant cluster until the end. Cluster 3 has papers with more columns, bigger area and publication that is more frequent.

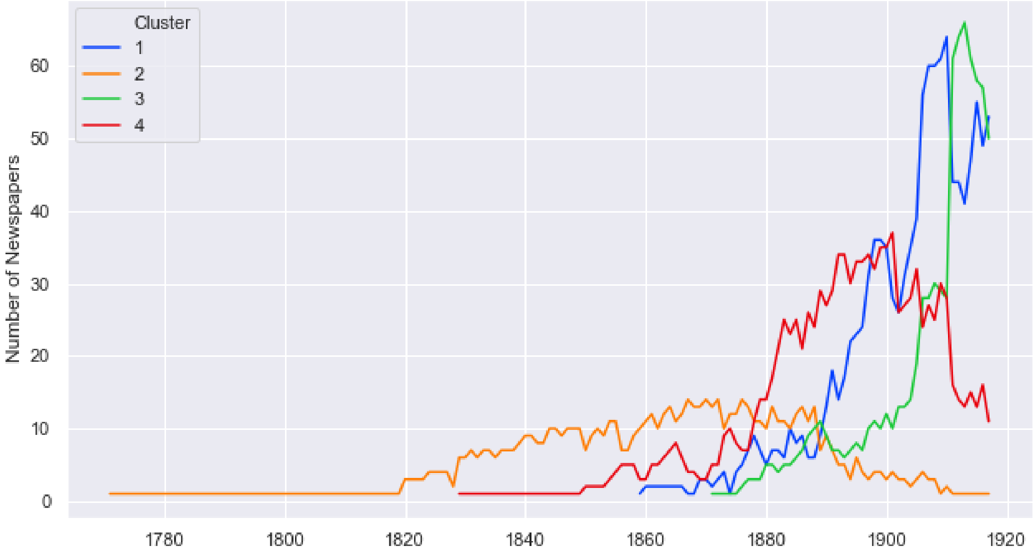


Figure 5.16: Cluster size increase and decrease, 1771-1917

In the (Table. 5.1) clusters and the number of data points in each cluster is given, but that represents the overall data. If a newspaper lived for 20 years, that newspapers has 20 data points, representing each year's features' values. There are possibilities, with change in features, a newspaper may have moved to another cluster. Following is a figure showing the number of newspapers staying in one cluster constantly or moving to more than one cluster over the life time. The figure shows that 45.4% of the newspapers lived within one cluster throughout their lives. 38.6% of the newspapers moved from one cluster to another, while 12% of the newspapers lived in 3 different clusters over the years. A mere 4% of the newspapers were changing their features more rapidly and moved within all the clusters over the years.

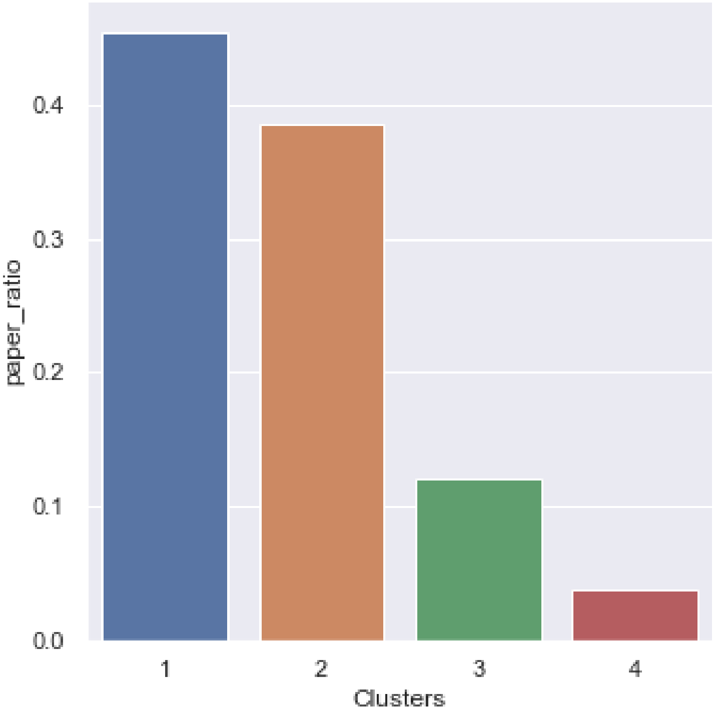


Figure 5.17: Ratio of newspapers living in one cluster or more

Since each record is assigned a cluster number, so every newspaper's record is extracted for further analysis. If a newspapers lived for 25 years, so there are 25 records of that newspaper's. The attributes' values for first year are compared with the next year if the cluster number is changed. The threshold is 0 for columns, pages, and publication period. If with the cluster change, any of these features' difference with the previous cluster's features is greater than this threshold, a change is recorded. Similarly, the threshold for density is 1, and for area it is 100mm^2. If a newspaper moves from one cluster to another, these features are compared with the previous record and the change is logged. Then the ratio of feature change is calculated by dividing feature changes with number of times cluster change. For example, a newspaper moved from cluster 2 to cluster 4, all other features remain in the same range except area, which changes 250mm^2. A new feature 'area change' is created and incremented with 1. Once again newspaper moves from cluster 4 to cluster 3, and this time it changes area more than 100mm^2 and changes columns from 5 to 6, so variables 'area change' and 'columns change' are incremented with 1. In the end both these features are divided by the number of time newspaper changed the cluster. In this case newspaper changed it's original cluster two times, so 'area change' and 'columns change' are divided by 2. This gives ratio of feature changes.

Following is a figure showing the features which changed considerably over the times, causing a change in cluster of the newspaper. To study the effect of individual features on cluster change, a threshold value of 0.75 is set. Any feature which changes more than 75% while all other features are not changing substantially, gives the following result. This indicates that 26% of the times a newspaper moved from one cluster to another because of a major change in area. None of the other features are solely dominating in this study.

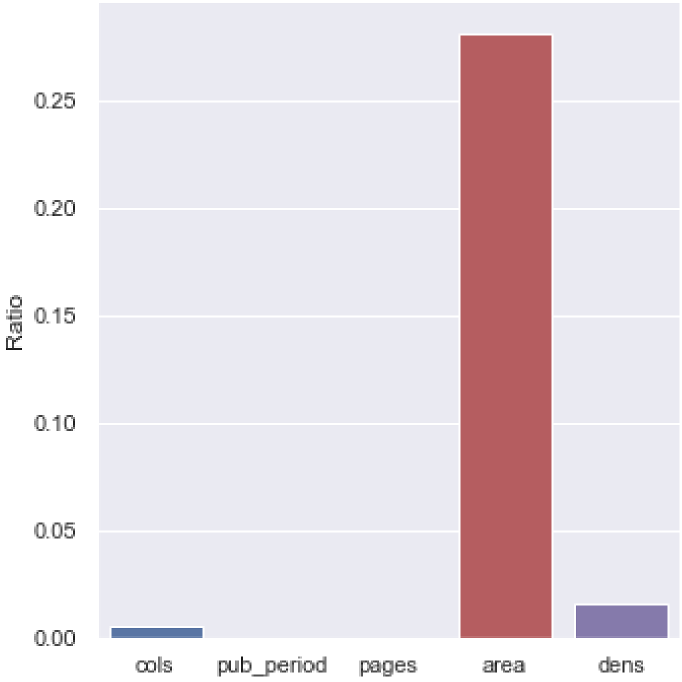


Figure 5.18: Feature causing newspaper change a cluster

This case study does not give a comprehensive understanding of the cluster change, since most of the newspaper moved from one cluster to another with changes in various features at the same time, but it shows the importance of the feature area. To understand the joint changes in various features, combination of features are tried the same way. Following result shows that 27.6% of the times when a newspaper moved from one cluster to another, the reason was changes in area and density. No two other features combined give any noteworthy results. Then another feature publication period is combined with area and density. This combination let a newspaper move from one cluster to another 8% of the times. Adding another feature columns to the combination helped a newspaper move from one cluster to another 7.6% of the times.

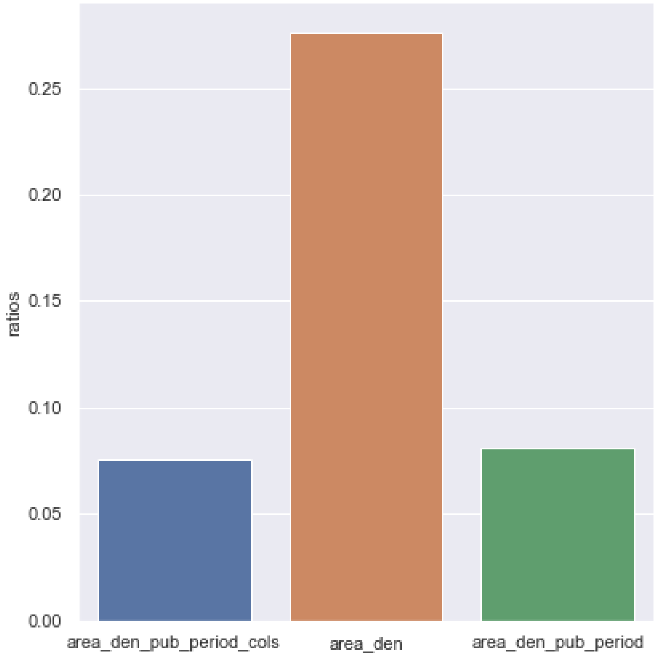


Figure 5.19: Combination of features causing newspaper change a cluster

The changes in different features as discussed above give a good understanding of how the cluster changed. Following is a case study showing overall change in different features when newspapers moved from one cluster to another because of a change in feature(s). The result shows that on average area changed 85% of the times, density changed 45%, columns changed 37%, publication period changed 25% and pages changed on average 4%.

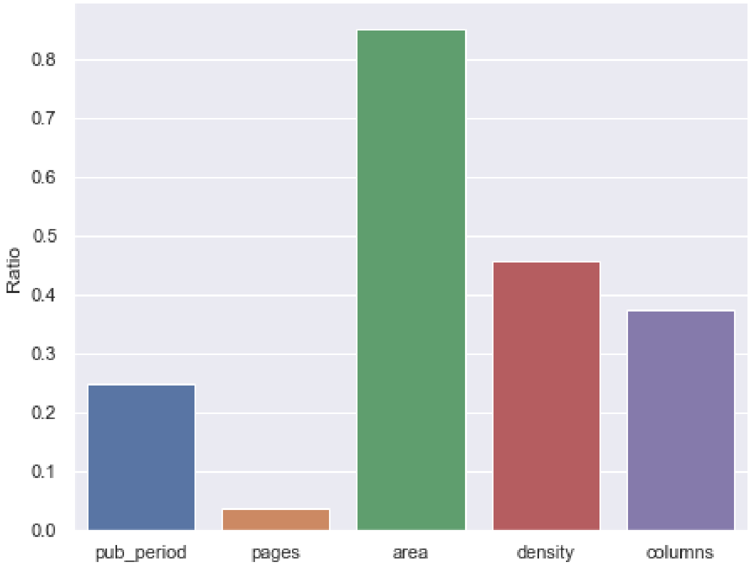


Figure 5.20: Ratio of average change in features

Next, the impact of changes in various features on the life of a newspaper is studied. Since newspapers' ratio of yearly cluster change is already computed, the newspapers are divided in three categories. Newspapers which changed their cluster less than 25% of their life time, the newspapers which changed their clusters less than 50%, and the newspapers which changed their clusters more than 50% of their life. The average life of newspapers in these three categories is shown below.

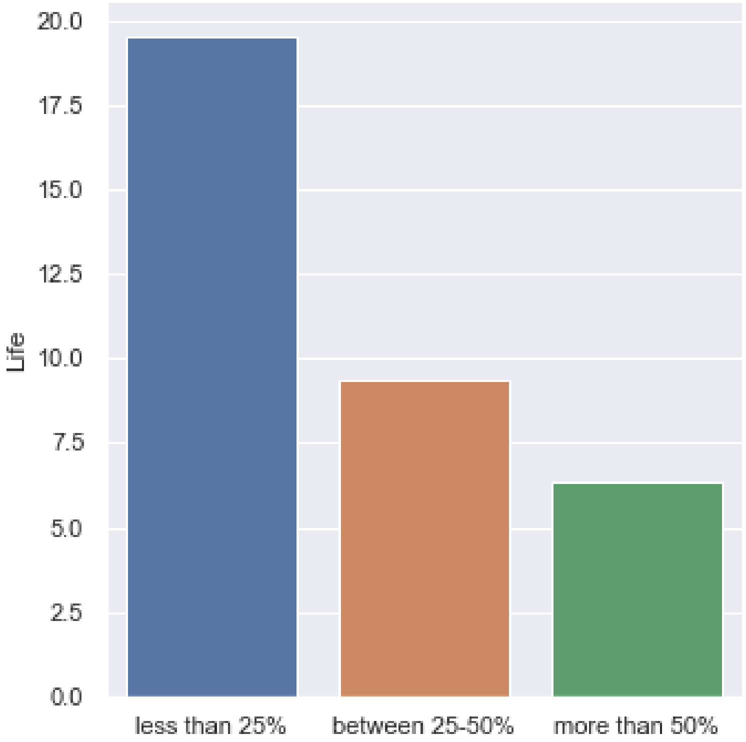


Figure 5.21: Comparison of newspaper's cluster change percentage and life

The newspapers which changed their cluster less than 25% of the time, on average lived for 19 years, while the newspapers which changed their cluster between 25% and 50% lived on average 9 years. Those newspapers which moved from one cluster to another every other year, lived on average 6 years. This case study shows the impact of stability in the features. The more stable the newspapers are with these features, the less chances of them to change clusters, which shows on average a longer life.

The above result is based on the ratio of how many times a newspaper changed its cluster in its entire life. A newspaper living four years and changed cluster 2 times, has a ratio of 0.5 whereas a newspaper living 20 years and changed clusters 5 times has a ratio of 0.4. By adding the absolute value of newspaper's life, the above representation seems problematic. To get a better understanding of newspaper's life and number of times it changed its cluster, absolute values are plotted below.

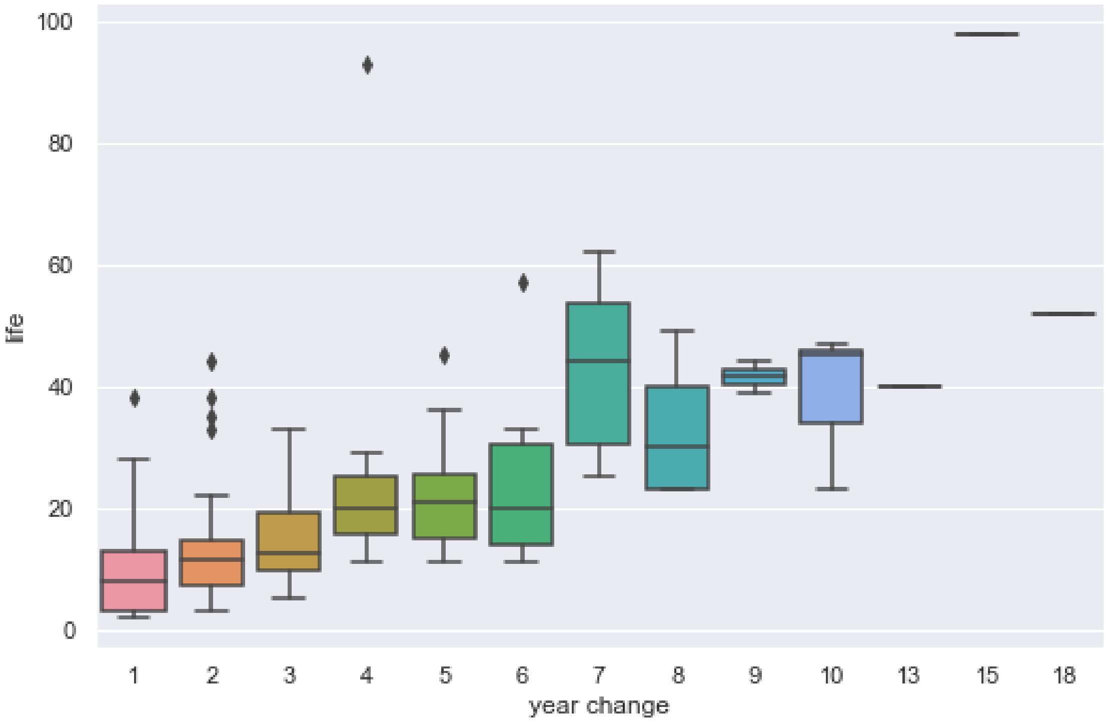


Figure 5.22: Comparison of number of times newspapers changed cluster and life

The result above shows life of newspaper and the number of times newspaper changed clusters. The x-axis masks the number of times as year change since the data is based on yearly averages. The newspapers which changed their clusters less than 7 times, mostly lived less than forty years. The average life of newspapers which changed cluster less than four times is between 10 and 15. The newspapers which changed their clusters more than 3 times but less than 7 times have an average life of 20 years. The newspapers which changed their clusters between 7 times and 10 times are those which lived more than 20 years with the average lives around forty years. The result shows that the newspapers which lived long are usually changing their clusters more than the short lived newspapers. This result is adequate, as it has already been discussed that with the time the publishing trends kept changing and the newspapers which lived long, generally followed those trend which caused the cluster change.

5.4 Hierarchical Clustering

The above analysis are based on different features, their values and their correlation with other features, and they give several new options for further analysis. One of the case study which emerges from above analysis is to understand the change in different features over the years, and then compare the newspapers based on these changes. The features selected for this case study are publication period, columns, density and area. As the data consists of yearly averages of these features, so each feature's change in percentage is calculated. Each feature is compared throughout its lifetime against a threshold value, and if the feature's values are decreased, increased or remained stable as compared to previous year, percentage of change is calculated. For example, an ISSN has the following area averages.

x = [118.81, 115.91, 116.56, 120.59, 128.24, 225.79, 161.97, 129.98, 128.18, 158.39]
A threshold value is set to be 10%, checking the increase in this feature, at index 4 (index starting from 0), value is 128.24 which increased to 225.79 which is more than 10%, and the last index has a value of 158.39 which is also more than 10% as compared to previous year. So the ratio of area increase is 2/10. Similarly the ratio of area decrease is calculated which is also 2/10, and area stability is calculated which is 6/10. Applying the same method on other features with their relevant threshold, ratios of change are calculated. The threshold value for columns and publication period is set to be zero, and for density it is set as one.

Since throughout this study, it has been observed that area feature is more vital for any analysis, so only this feature is selected for next case study. The objective is to cluster the newspapers based on this feature. Each newspaper is represented by only one data point, with three features Area Increase, Area Decrease and Area Stable. Another feature Volatility is added in this data set, which represents the ratio of change as a whole for a newspaper. Taking the above values of Area

x = [118.81, 115.91, 116.56, 120.59, 128.24, 225.79, 161.97, 129.98, 128.18, 158.39]

and using the same threshold, each value is compared to only it's previous year. Since the threshold value here is 10% also, so index are labeled as Increase, Decrease or Stable. The above data has following labels.

X = [Stable, Stable, Stable, Stable, Increase, Decrease, Decrease, Stable, Increase]

By comparing each label with its previous label, it seems that this newspaper changed its area 5 times in its life. For first five years the area was stable, then increased, then decreased. In the penultimate year it got stable and increased again in the last year. So this newspaper lived for 10 years and its Volatility is 5/10.

Now the dataset has four features to be clustered. Though density-based and centroid-based algorithms are considered for this case study also, but the conditions of selecting parameters in advance in those algorithms created problems with this dataset. Initial exploratory analysis with K-means and DBSCAN could not cluster the data with varying range of parameters. Since the objective of this case study is to build a hierarchy based on area change and study at what distance the newspapers start merging into clusters, Hierarchical clustering seems an appropriate method.

The hierarchical clustering method has been studied for this set-up. The reason for applying hierarchical clustering is, since it takes each data point as a single cluster and then starts creating clusters of similar objects, newspapers with the same ratio of area change (increase, decrease and stable) and volatility will be clustered together. The hierarchical clustering works based on the proximity matrix which contains the distance between data points. At its core the algorithm for hierarchical clustering is given below.

Algorithm 3 Hierarchical Clustering Algorithm [30]

```
Initialization: $P_1 = \{\{x_1\}, \{x_2\},. . . , \{x_n\}\}$
for t = 1 to n  1 do
    Compute pairwise linkage values between clusters of the current partition Pt
    Merge the two clusters with minimal linkage value to obtain the next partition $P_{t+1}$
return $P_1, P_2, ..., P_n$
```

The vital element of the above algorithm is linkage method to calculate the distances between data points. There are several methods to find the distance between clusters. Following is the brief description of three main linkage methods, which are studied for this research.

Single Linkage

It is the simplest hierarchical method to find the distance between data points. Given two clusters C1 and C2 , single linkage computes the minimum distance between two objects/clusters [19]. The mathematical notations are as follow

$$f(C1, C2) = \min_{u \in C1, v \in C2} d(u, v) \tag{5.6}$$

where d is the distance between point u and v [19] . Following figure represents this method in more elaborating way.

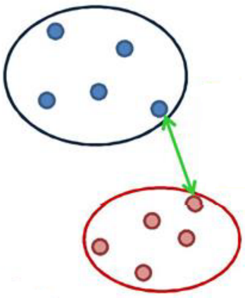


Figure 5.23: Single Linkage

Complete Linkage

This is the maximum linkage, where distance between two clusters is based on the largest link from one cluster to another [19]. Given two clusters C1 and C2 the formula to calculate complete linkage is given as

$$f(C1, C2) = \max_{u \in C1, v \in C2} d(u, v) \tag{5.7}$$

where d is the Euclidean distance [19]. Figure below explains the complete linkage method in a more concrete manner.

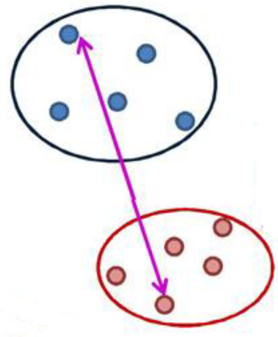


Figure 5.24: Complete Linkage

Ward Linkage

Ward linkage is different from the above two methods. The distance between two clusters in ward linkage method is the sum of the squares of the distances between all data points in the cluster and the centroid of the cluster [19]. For two clusters C1 and C2, the distance is calculated as,

$$f(C1, C2) = \sum_{x \in C1 \cup C2} d(x, \mu C1 \cup C2)^2 \tag{5.8}$$

where d is the distance and μ is the centroid of the new cluster merged after the union of C1 and C2. [19]

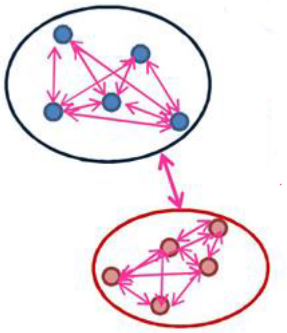


Figure 5.25: Ward Linkage

To compare the results of these methods, Cophenet Score [32] is calculated and based on the score, the method with maximum score is chosen for clustering. Cophenet (X,Y) computes the correlation coefficient for hierarchical cluster (X), which is the output of applied linkage method. Y contains the distances used to construct X. A dendogram represents the cophenetic distance between two observations. This distance is the height of the node at which these two points are first joined together. [32]

$$cophenet\_scores = \frac{\sum_{i<j}(Y(i,j)-y)(X(i,j)-x)}{\sqrt{(\sum_{i<j}(Y(i,j)-y)^2 \sum_{i<j}(X(i,j)-x)^2)}} \tag{5.9}$$

where $Y(i,j)$ is the distance between point i and j in Y and $X(i,j)$ is the distance between point i and j in X. y and x are mean values of Y and X respectively [32]. The values of cophenet function ranges from 0 to 1, the higher the magnitude of this correlation coefficient, better the solution is. The following figure shows the scores of three methods, and based on this result Hierarchical clustering is performed with the Ward Linkage method.

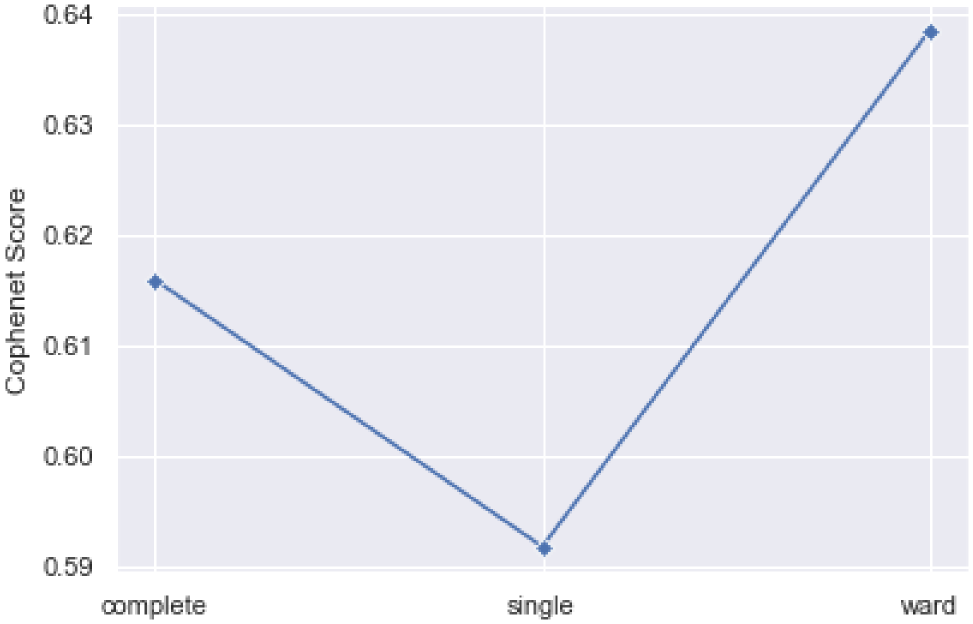


Figure 5.26: Cophenet method to find the best linkage method

Using Euclidean distance, Hierarchical clustering is performed with Ward method and for representational purpose dendogram is used. The following dendogram shows the complete hierarchy of the clusters. At the leaf nodes, number of newspapers are given which falls on one side of the cutting point or other.

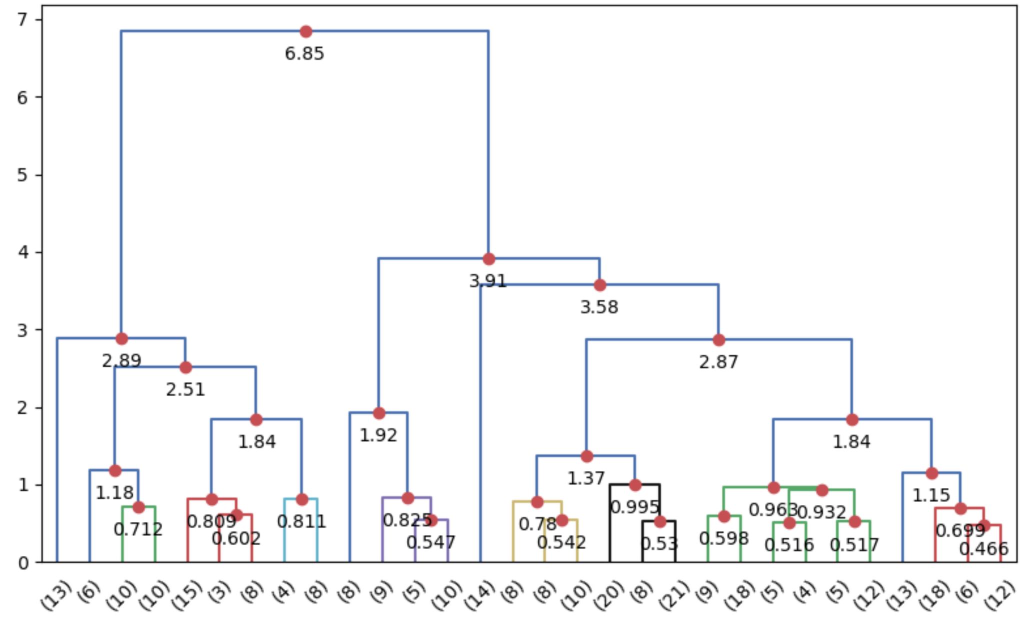


Figure 5.27: Hierarchical clustering dendogram of Area

Since this is a dense representation of clusters, so the dendogram is cut at 3.58 distance. The following figure is more interpretable and understandable. The cut off points in the figure are distances, which are calculated by Ward method. The maximum distance between clusters is 6.85, which divides the data into two clusters, grouping 77 newspapers in one cluster and 223 in another. With the distance of 3.91, the newspapers in second clusters are further divided in two groups, with 32 newspapers in one cluster and 191 newspapers in another. Similarly, using the distance 2.87 between data points, 6 clusters are formed. By cutting the tree at 3.58, number of clusters can be 4, but in that case 177 newspapers fall in fourth cluster, which is 59%of the data. So cutting it at 2.87, gives six cluster with reasonable cluster sizes.

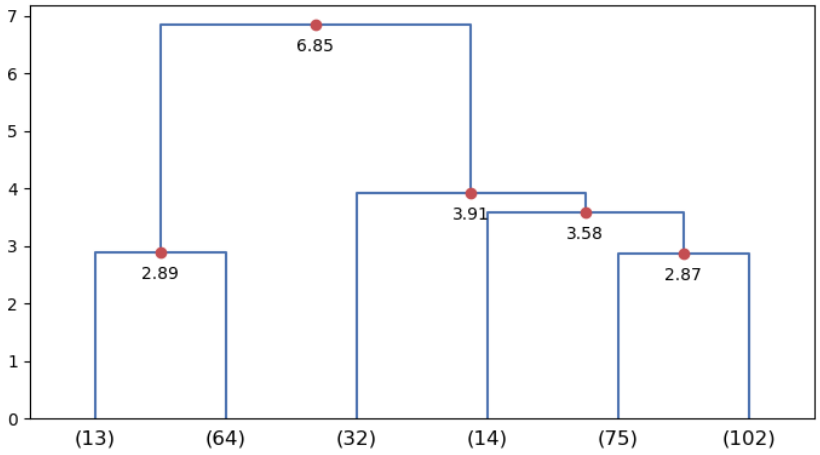


Figure 5.28: Truncated Hierarchical clustering dendogram of Area

Following is a representation of clusters for each feature. The result shows that cluster one has newspapers with more area stability, whereas cluster two has a mix of features. Cluster two has newspapers with average area increasing, decreasing and moderate volatility. Cluster three consists of newspapers with more decrease in area and cluster four has newspapers with more increase in area. Cluster five has papers with less stability, while cluster six has newspapers with more volatility.

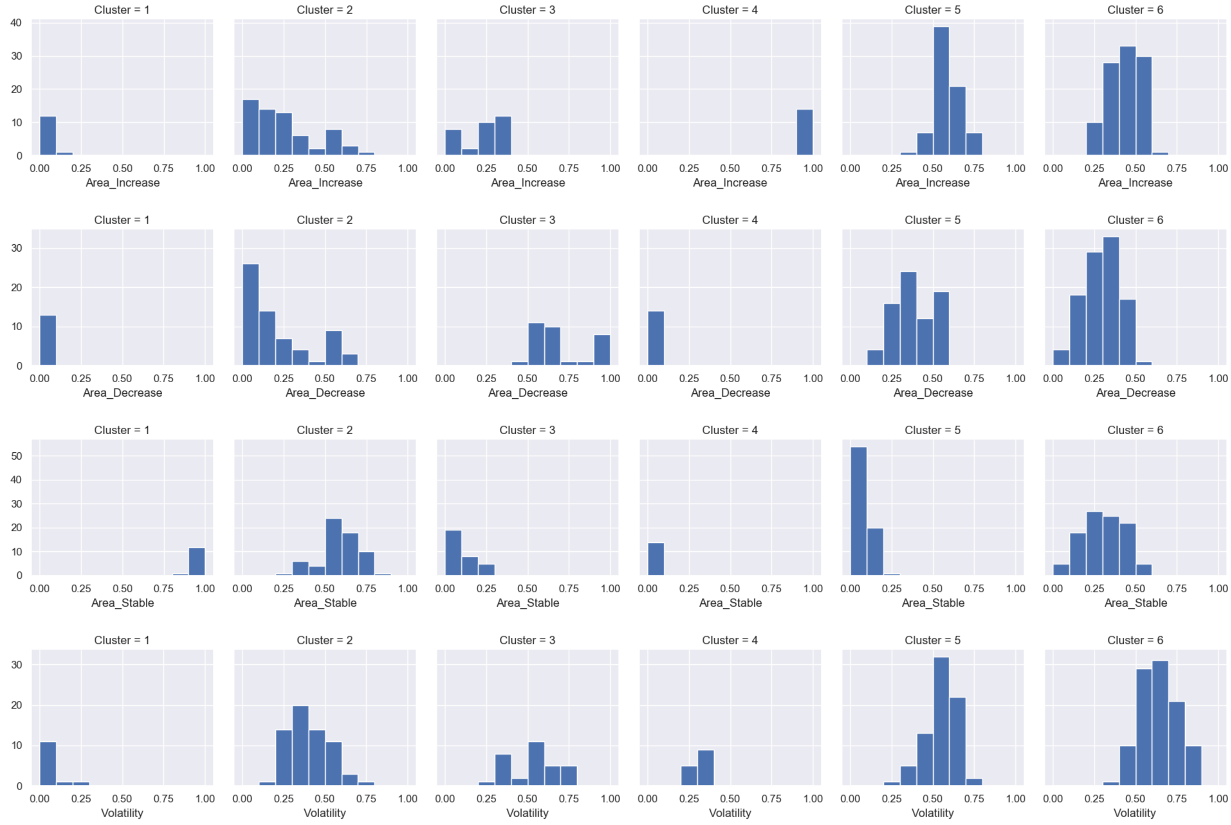


Figure 5.29: Comparison of Area changes for six clusters

Since the clustering method assigned a cluster label to each data point, the original data is extended with another feature, Cluster, which represents each newspaper with its cluster number. Following plot shows the same results as discussed above by comparing features against each other. Feature Life is also added in this figure to study the clustering based on area changes against life of the newspaper. The results indicate that the newspapers in cluster six, which has more volatility, have on average longer lives while the cluster one and two have newspapers with more stability but comparatively shorter lives.

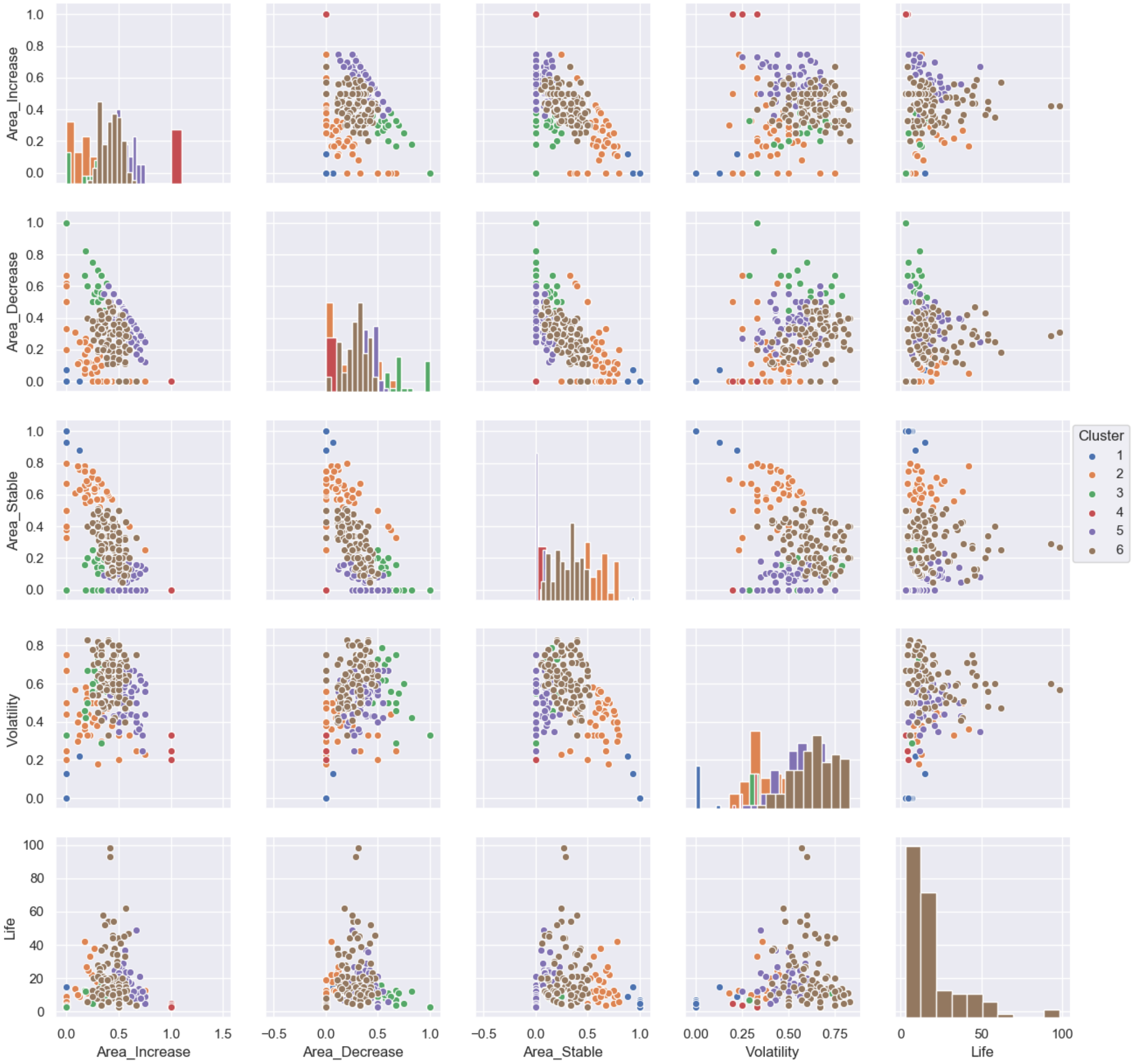


Figure 5.30: Pair plot of Area changes comparison of six clusters

A total of 300 newspapers are clustered. Total number of Swedish and Finnish language newspapers are 89 and 211 respectively. Since the underlying structure of clusters has been discussed above, the next step is to find the number of Swedish and Finnish newspapers in each clusters. The result below shows that overall cluster one and four have fewer number of newspapers as compared to other clusters. Cluster one and cluster four are the ones with more stability and increasing area respectively. Cluster two has the same number of newspapers for both languages. Cluster five and six which have less area stability and more volatility are showing a large number of Finnish newspapers. Since the total number of Swedish newspapers is very less than Finnish newspapers, they are less in numbers in all the clusters except cluster two.

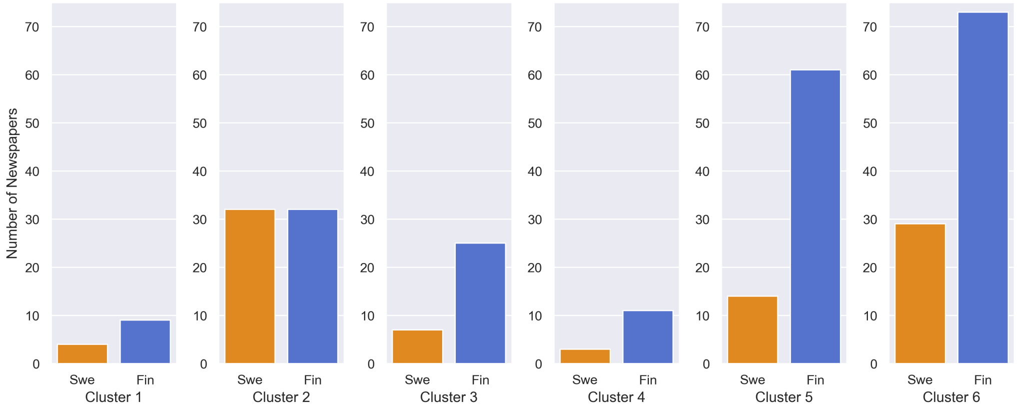


Figure 5.31: Number of Swedish and Finnish newspapers in six clusters

As it has been discussed above, the clusters are based on the area and volatility features, and cluster six has the newspapers with longer lives, but within the clusters, lives of newspapers based on their language is the focus of next analysis. The result below shows that Swedish newspapers have longer lives than Finnish newspapers in all clusters but cluster three and four. As the description of cluster three suggests that it has newspapers with less stability and more area decrease. Similarly cluster four's description confirms it contains newspapers with less stability and more area increase. In these two clusters Finnish newspapers have longer or equal lives as compared to Swedish newspapers. This states the instability of area in Finnish newspapers. These two clusters show shorter lives of Swedish newspapers as compared to other clusters. Cluster one, two and five show overall long lives of Swedish newspapers. Cluster six, which has newspapers with more volatility, comprises of long-lived newspapers though the average life of Swedish newspapers in this cluster is lower than cluster five.

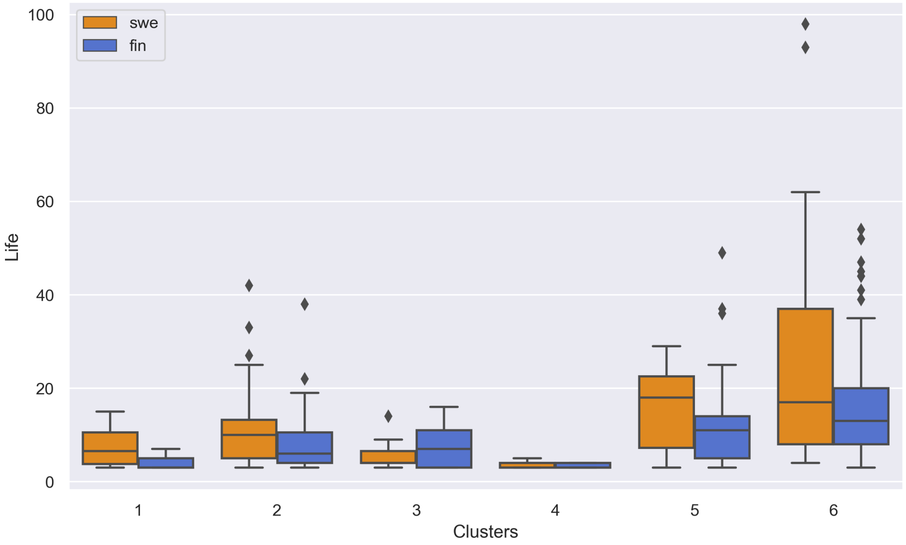


Figure 5.32: Life of Swedish and Finnish newspapers in six clusters

Since the area feature is divided into three sub-features Increase, Decrease, and Stable to understand it more thoroughly, similarly other features columns, density, and publication period have also been divided into sub-features to compute their ratio of change. The rationale to select these three features is based on the result of the Random Forests algorithm in classifying page size. The result in (Fig. 4.37) shows the important features to classify page size. This result indicates that the number of pages has the least impact on page size. Since page size is derived from the area feature, and this case study is based on changes in area, the three important features density, columns, and publication period are considered for further comparative analysis.Based on the clustering achieved above, stability of density, columns and publication period is compared for newspapers published in Swedish and Finnish language. The result below shows the density stability of newspapers of both the languages in six clusters. All the clusters have almost stable density except cluster four and five, where Swedish newspapers are showing less stability. An important outcome of this result is that it shows average density of 80 or above for all the clusters except cluster four, which contains the newspapers with increasing area.

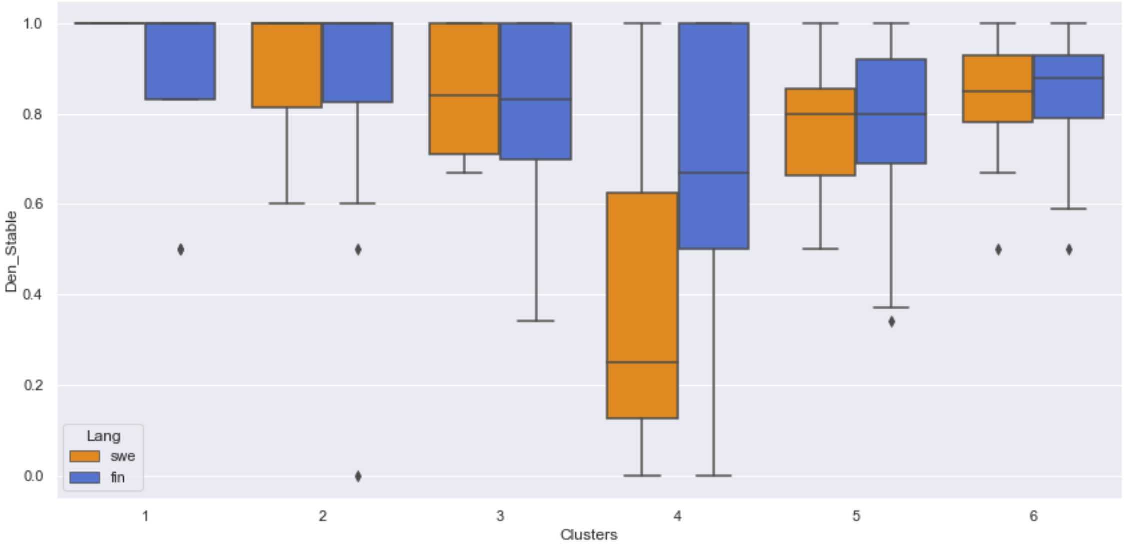


Figure 5.33: Density stability of Swedish and Finnish newspapers in six clusters

The figure below shows the columns stability for six clusters. Overall, Swedish newspapers have more columns stability than Finnish newspapers in almost all the clusters except cluster four. On average Swedish newspapers have columns stability of more than 80 whereas Finnish newspapers have fluctuating average in all clusters. Cluster four is the only cluster where Finnish language newspapers are showing slightly more stability than Swedish language newspapers, though in this cluster, Finnish language newspapers have less average columns stability than in other clusters.

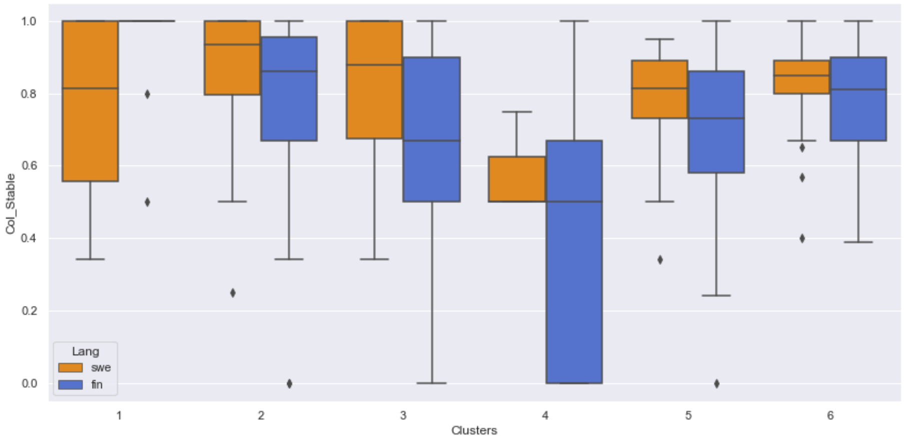


Figure 5.34: Columns stability of Swedish and Finnish newspapers in six clusters

Similarly publication period stability is shown below for all clusters. Following the above patterns, Swedish newspapers are showing more publication period stability than Finnish newspapers in all clusters except cluster four. Above three figures have shown almost the similar results. Finnish newspapers are ahead in all features in cluster four whereas in all other clusters Swedish newspapers are showing more stability. The features' average values in most of the clusters for Swedish newspapers are also higher than Finnish newspapers.

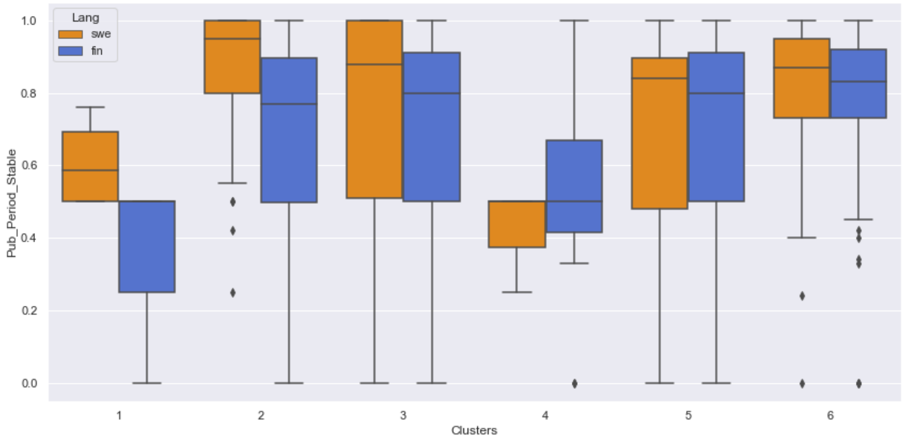


Figure 5.35: Publication Period stability of Swedish and Finnish newspapers in six clusters

Various features along the publishing language of the newspapers for each cluster have been discussed above but the place of publication is still missing in this case study. The figure below compares newspapers from different cities and in different languages which fall in cluster one. Since this cluster has newspapers with more area stability, so density, columns and publication periods are compared. The figure shows variation among newspapers and these features. Even though this cluster has newspapers with more stability in area, but some cities such as Salo and Mikkeli shows less density stability, while Kristiinakaupunki shows less Columns stability. There are though newspapers from Helsinki, Vaasa and Huittinen which have more stable features.

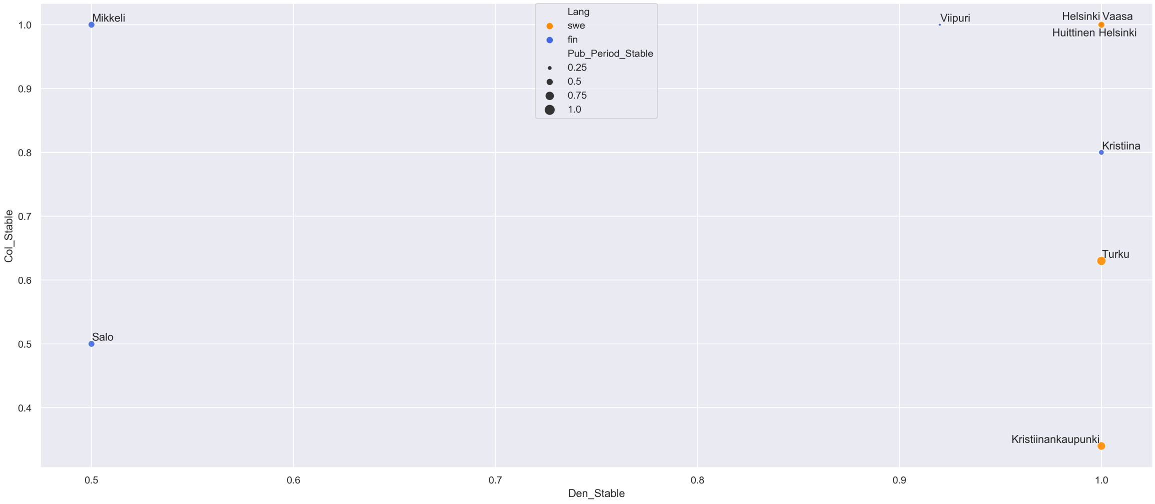


Figure 5.36: Cluster 1: Language and cities based comparison of newspapers' Density, Columns and Publication Period stability

Cluster two has newspapers with moderate values of area changes, and volatility not on any extreme, figure below shows more stability in all the features. Newspapers from both languages have shown more density and columns stability.

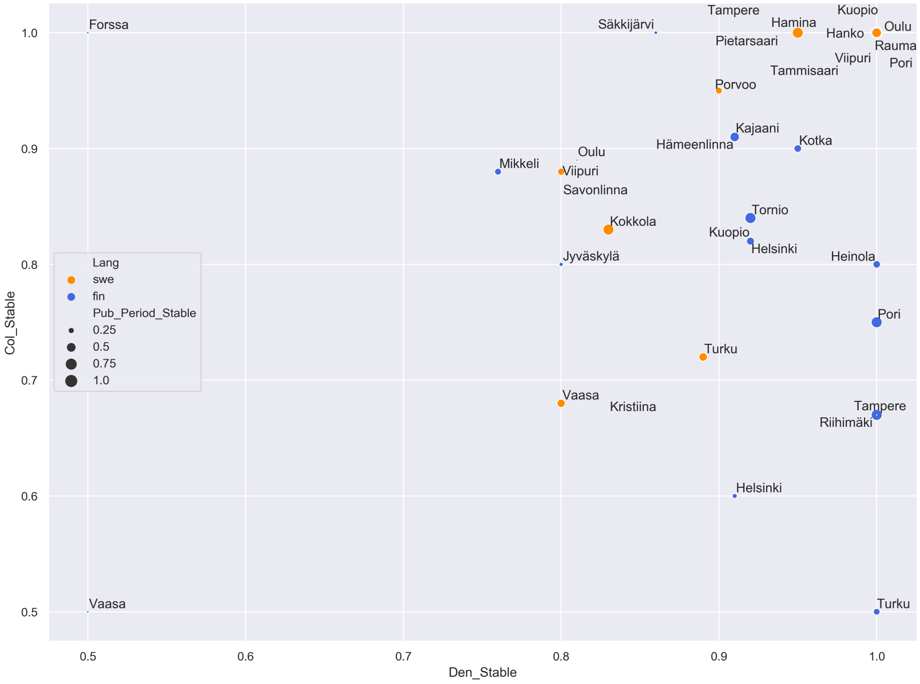


Figure 5.37: Cluster 2: Language and cities based comparison of newspapers' Density, Columns and Publication Period stability

As discussed earlier, cluster three has newspapers with more decrease in area, following figure shows less stability in other features also. Though newspapers from different cities are leaning more towards stable columns but the variation in density stability is showing more saturation here.

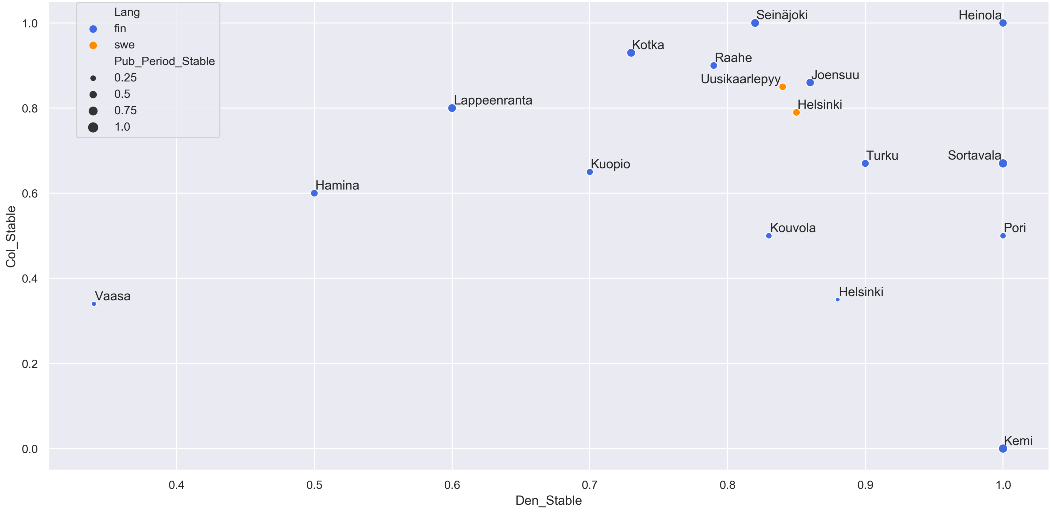


Figure 5.38: Cluster 3: Language and cities based comparison of newspapers' Density, Columns and Publication Period stability

Cluster four consists of newspapers with more Area increase, but the figure below shows there are several cities publishing newspapers in both languages with less stable features. For example, Finnish and Swedish newspapers from Helsinki have less stability in density and publication period. This cluster is dominated by the Swedish newspapers.

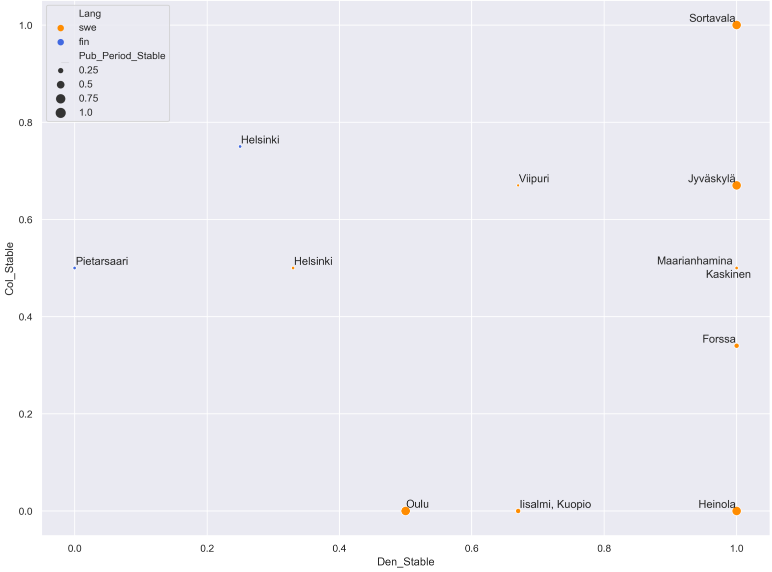


Figure 5.39: Cluster 4: Language and cities based comparison of newspapers' Density, Columns and Publication Period stability

Cluster five has newspapers with less stable area and more fluctuation between increase and decrease. The result below shows more column stability in this cluster. The density stability is also high for almost all the cities for both the publishing languages. It is an interesting finding, though area is less stable, but other features are stable in this cluster.

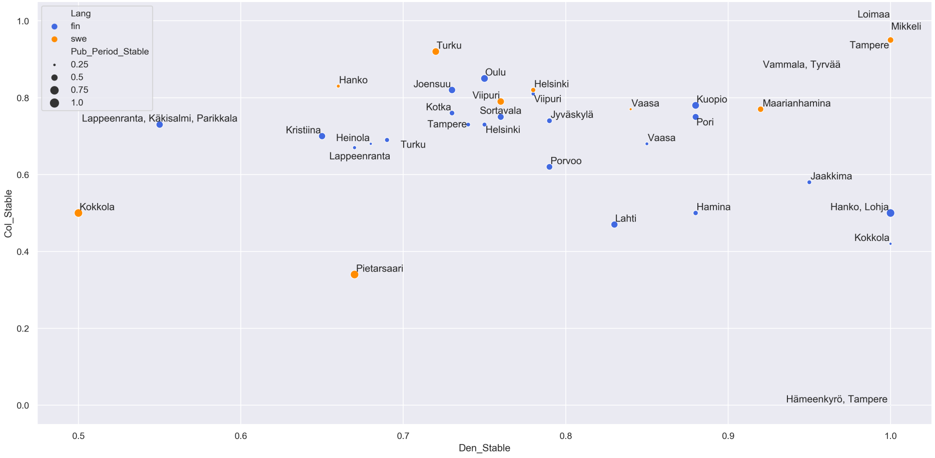


Figure 5.40: Cluster 5: Language and cities based comparison of newspapers' Density, Columns and Publication Period stability

Volatility is high in cluster six, which eventually shows high-density stability and column stability. This cluster has the most number of newspapers, which is the reason of having more cities in it. This cluster shows almost the same publication period stability for both Swedish and Finnish newspapers, which indicates that even though the newspapers have high volatility in area size but they kept their publication period stable more or less throughout the country.

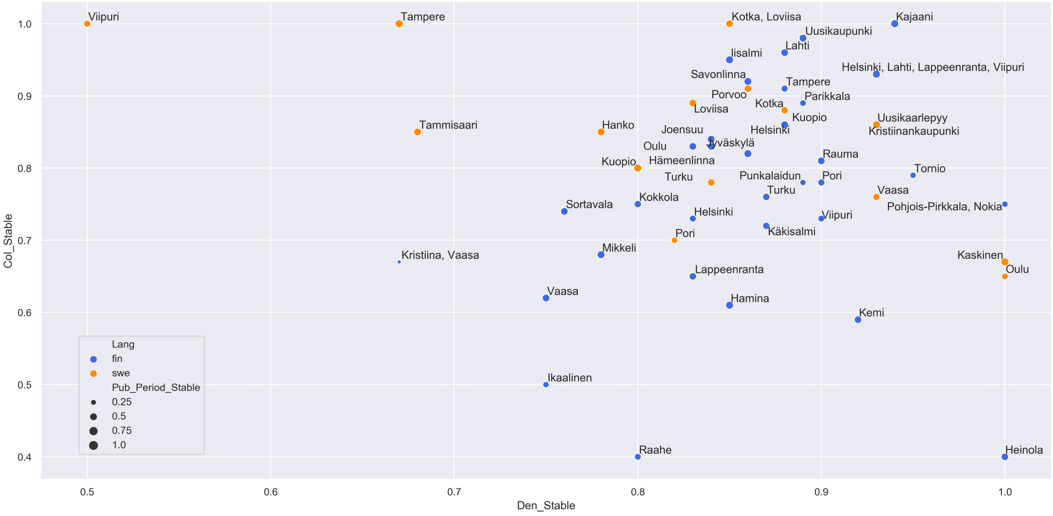


Figure 5.41: Cluster 6: Language and cities based comparison of newspapers' Density, Columns and Publication Period stability

The above results show three features' stability in six clusters for Swedish and Finnish language newspapers in all the cities. Since the number of newspapers in all clusters vary, and understanding which city has more columns, density or publication period stability in which publishing language is not an easy task in the above results. To compare the cities and features for six clusters, the data is filtered. A total of 51 cities are present in this data, and visualizing all of them is not very useful, as 13 cities have only one newspaper. For that reason, only those cities are filtered which published newspapers in both the languages. A total of 14 cities are found with the Swedish and Finnish languages newspapers. The result below shows the columns stability of these cities in six clusters. The result shows Helsinki is the only city with newspapers in all clusters, whereas Turku, Vaasa, Kuopio and Viipuri has newspapers in five clusters. Except Helsinki, no other city has columns stability of less than 0.4 for Swedish language newspapers, on the other hand, Finnish language newspapers have more fluctuations in columns stability within clusters and cities.

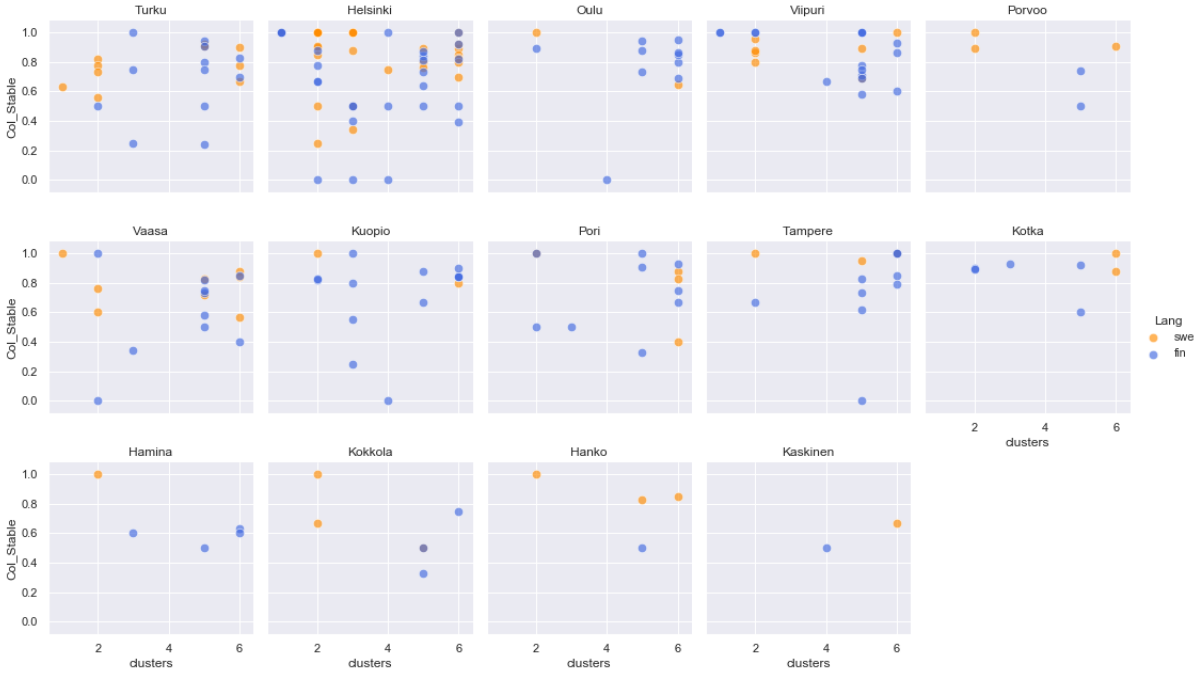


Figure 5.42: Comparison of clusters and columns stability of cities publishing in two languages

Similarly, density stability in six clusters is compared for cities. The newspapers of both the languages have overall more density stability regardless of the cluster they are in. Except the newspapers in cluster four, Helsinki has density stability of over 0.4 in all the clusters. Porvoo, Pori, Kotka, and Hanko have density stability of more than 0.6 for the newspapers in all the clusters. No newspaper in cluster one has density stability of less than 0.8 and except a Swedish language newspaper from Viipuri, newspapers in cluster six have density stability of at least 0.6 for all the cities which is interesting as the cluster six has newspapers with more area volatility.

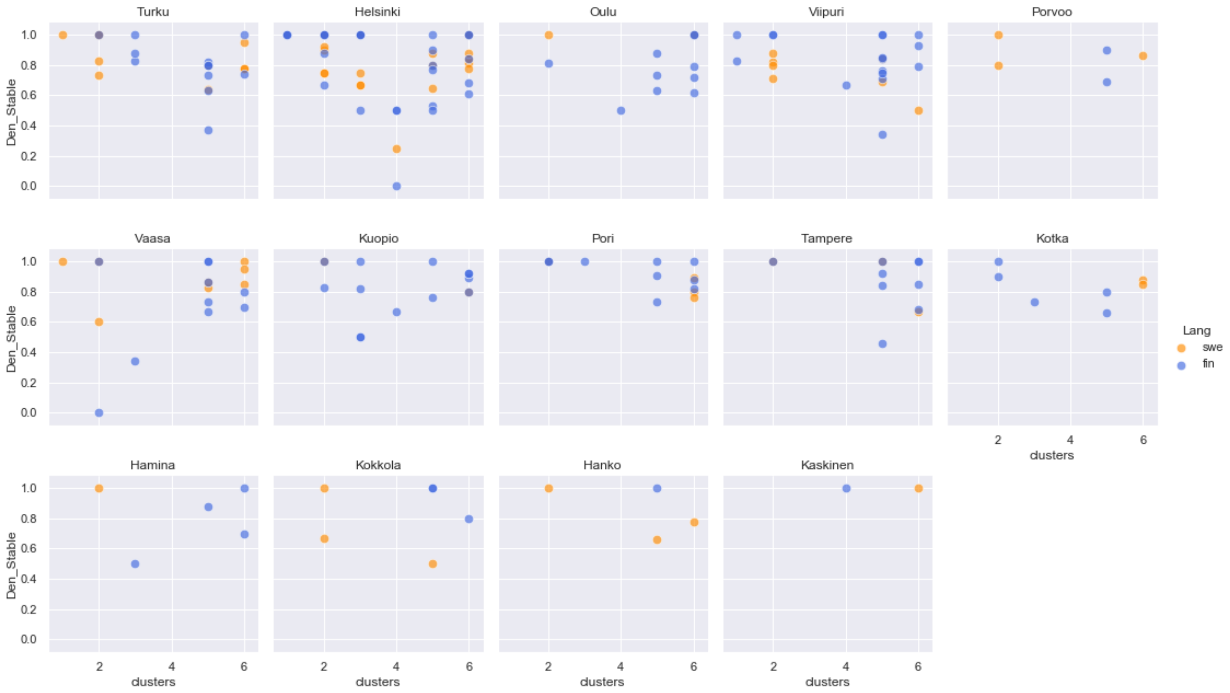


Figure 5.43: Comparison of clusters and density stability of cities publishing in two languages

Stability of publication period in six clusters for the same cities as above is compared and shown in the result below.

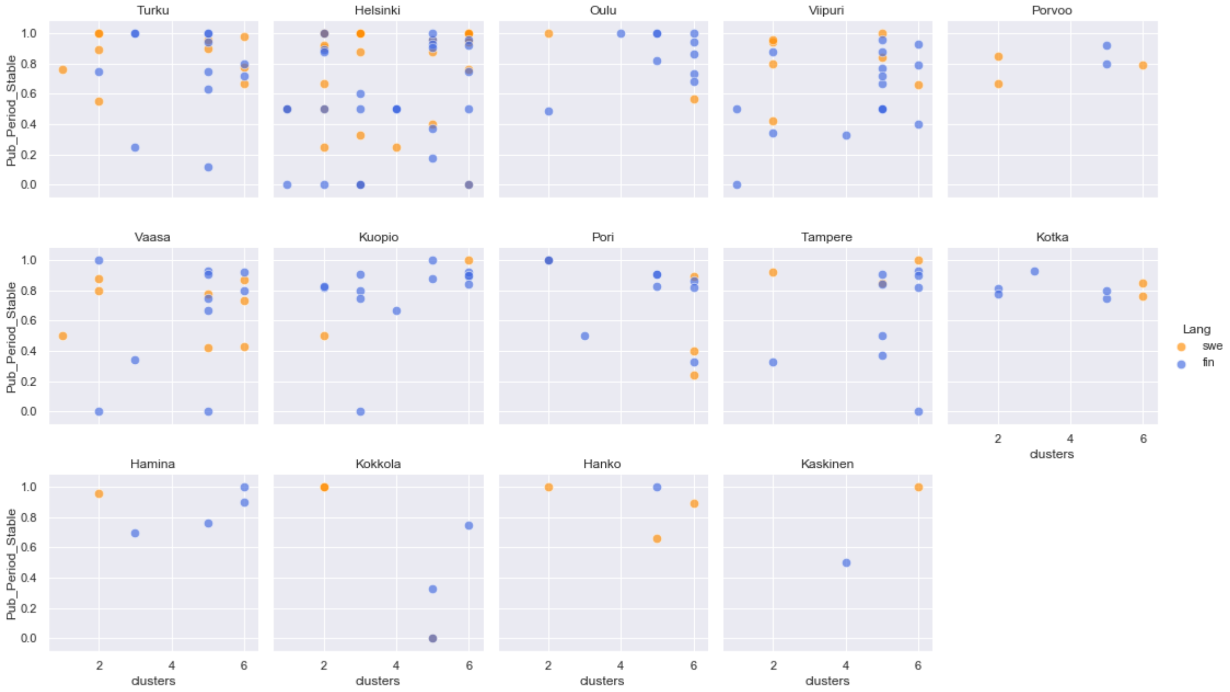


Figure 5.44: Comparison of clusters and publication period stability of cities publishing in two languages

The above result shows that Porvoo, Kotka, Hamina, and Hanko are the cities with publication period stability of more than 0.6 in all clusters. Almost all other cities have some newspapers with very less publication period stability and some with very high stability. All the clusters in this result have newspapers with varying publication period stability.

The above two case studies, results from (Fig. 5.36) to (Fig. 5.44) were focused on the comparative analysis of cities, languages and features' stability in six clusters, and they give a very detail understanding of the data. Since only the cities with newspapers in both the languages were filtered for the last case study, several cities with a number of newspapers were omitted. So as a last case study based on hierarchical clusters, regions as discussed in the exploratory analysis in (Chapter:4), are compared for the three features and six clusters. The result below shows that the regions have overall more columns stability in cluster two, five and six whereas small towns have more average columns stability in cluster three as compared to other regions.

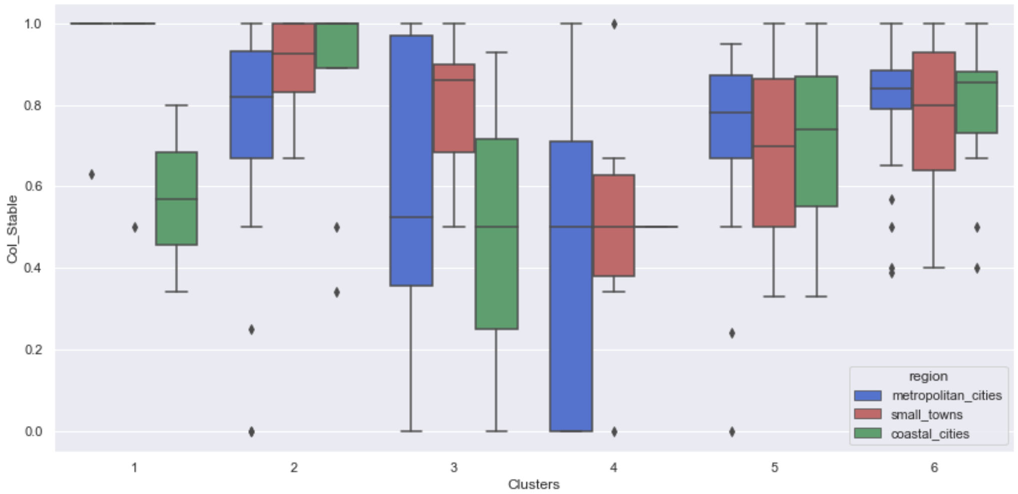


Figure 5.45: Comparison of regions and columns stability in six clusters

The result below shows the density stability in three regions. Interestingly, all the regions have very high density stability in all the clusters except cluster four, where metropolitan cities have an average of 0.5. Cluster three has newspapers with more decrease in area but the result below indicates that all the regions have average density stability of more than 0.8 whereas cluster six which represents the newspapers with more volatility has also very high density stability in the three regions. Newspapers from Small towns are present in each cluster, and they have on average same density stability as coastal and metropolitan regions.

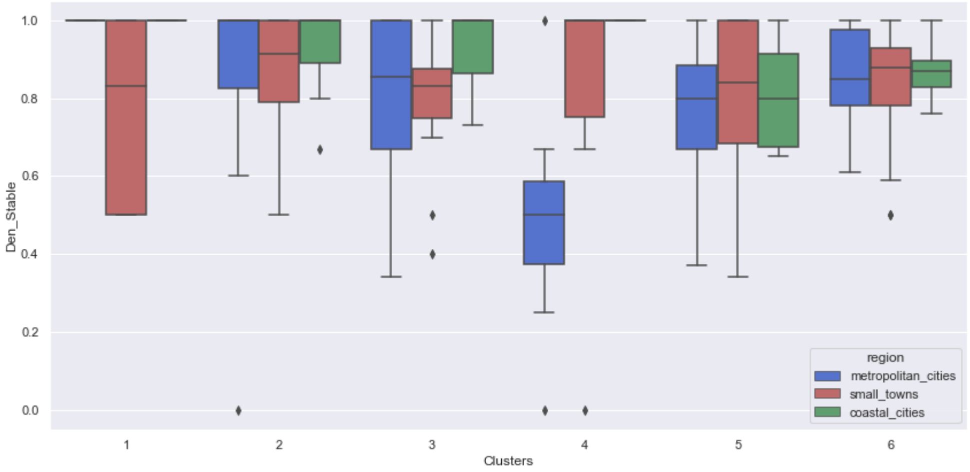


Figure 5.46: Comparison of regions and density stability in six clusters

In the last, publication period stability is compared for all the regions. The result shows that cluster two and six have average publication stability over 0.8 for all the regions whereas cluster one and four have the least stability. In cluster three metropolitan have the lowest average and cluster five has the lowest average stability for small towns.

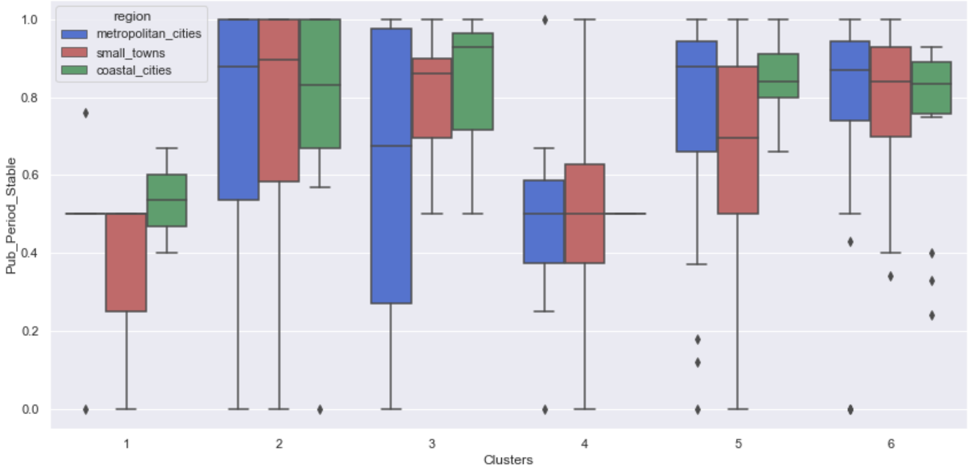


Figure 5.47: Comparison of regions and publication period stability in six clusters

Summing up the results of hierarchical clustering, area feature is divided into three sub-features. Hierarchical clustering created six clusters based on the changes in area. Cluster six which had newspapers with more volatility had the highest number of total newspapers whereas cluster one and four had fewer newspapers. The Swedish language newspapers had on average long lives, more columns stability and publication period stability. Mapping the cities to their regions, newspapers from metropolitan regions have more stable features in cluster 2, 5 and 6 whereas the newspapers from small towns had more density stability in all six clusters. By comparing these discussed results with

(Fig. 4.9), it is clear that even though the ratio of following a popular publishing trend decreased with time, but the newspapers from different regions kept their features' stable regardless of the trend they were following.

5.5 Decision Tree

The meta-data of newspapers have been studied in detail in this research work from individual development of features, to language and place of publication. Newspapers are clustered based on their area sizes and other features. Regression methods have been applied on the features and decision trees has been created for features. The last case study is to create decision trees based on the feature change.

The changes in features, computed in the above scenarios, are describing either feature has increased its value, decreased or kept it stable. This change is calculated over the whole life of a newspaper. For this case study, a newspaper's features' changes are computed for each year and studied those changes until the newspapers stopped publishing. For example, a newspaper has an average area of 600mm^2 for the year 1820, and in year 1821 it has an average area of 500mm^2. To calculate the area change, current year's area is subtracted from previous year's and then divided by the previous year. So area is changed 16% in this case. Since area is decreased, a minus sign is prefixed with the percentage of change, indicating a decrease in the feature.

After preparing the data for this case study, random forest is applied, and the results are shown for only depth of four. Only those newspapers which died within the time period of our data, are considered for this case study. Though the percentage of change in a feature is measured before applying the model, still the variation of the data is incomprehensible. For example, the result below shows that if the publication period of a newspaper has changed less than 13.40%, there is a possibility the newspaper is dying this year. Now this is a very rigid hypothesis in itself because out of 6.2% samples based on this change limit, 0.2% shows the death of a newspapers and 6.0% proves otherwise, but considering a newspaper lived for 30 years, one record of newspaper is with label 'yes' and remaining 29 are with label 'no' for the feature 'Dying'. The newspapers had so much variation among their features, that finding any significant pattern to prove that a newspaper is dying is not an easy task. The model is trained with 80% of the data, and 20% is reserved for testing. The accuracy of the model is 0.827, which is good, considering the variation in the data. Since the tree is cut at the depth of 4, so the features' changes before the newspapers ended are varying. Starting from the left, if a newspaper changed its publication period less than 11.8%, density is changed less than 5.87% and area is decreased but not less than 3.6%, there are 0.1% samples proving that the newspaper is dying. Similarly, there are 0.2% samples suggesting that the newspaper is dying if the publication period is increased but not more than 11.8%, area is decreased more than 3.46%, density increased more than 5.87%, and column decreased more than 18.34%. In another scenario, if a newspaper's area is decreased but not more than 3.46% but the publication period is decreased up to 95.72%, there are 0.1% samples where newspaper died.

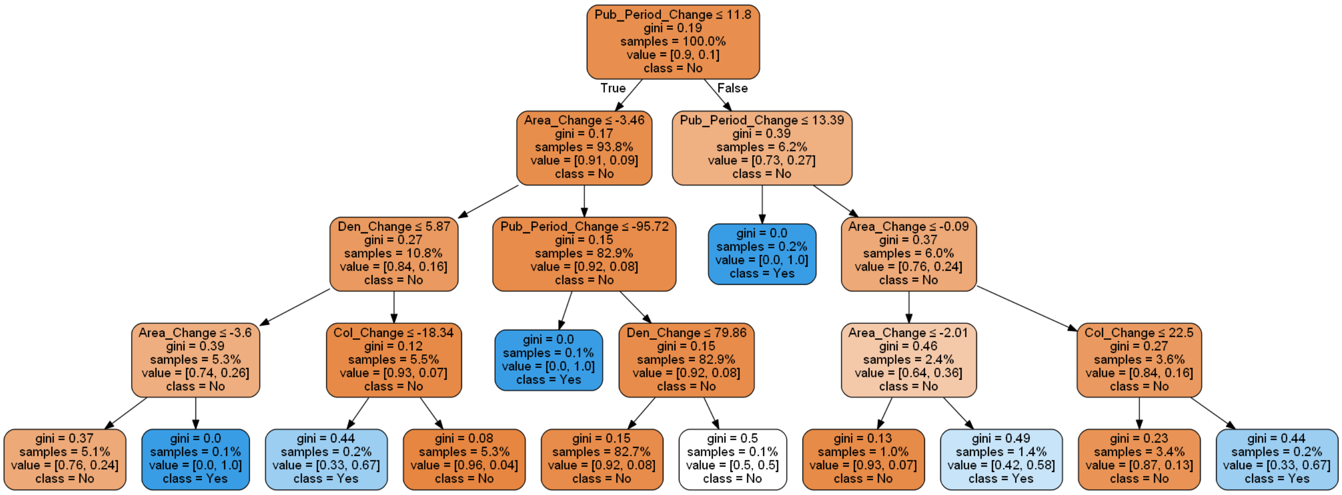

Pub_Period_Change ≤ 11.8 gini = 0.19 samples = 100.0% value = [0.9, 0.1] class = No
True
False
Area_Change ≤ -3.46 gini = 0.17 samples = 93.8% value = [0.91, 0.09] class = No
Pub_Period_Change ≤ 13.39 gini = 0.39 samples = 6.2% value = [0.73, 0.27] class = No
Den_Change ≤ 5.87 gini = 0.27 samples = 10.8% value = [0.84, 0.16] class = No
Pub_Period_Change ≤ -95.72 gini = 0.15 samples = 82.9% value = [0.92, 0.08] class = No
gini = 0.0 samples = 0.2% value = [0.0, 1.0] class = Yes
Area_Change ≤ -0.09 gini = 0.37 samples = 6.0% value = [0.76, 0.24] class = No
Area_Change ≤ -3.6 gini = 0.39 samples = 5.3% value = [0.74, 0.26] class = No
Col_Change ≤ -18.34 gini = 0.12 samples = 5.5% value = [0.93, 0.07] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
Den_Change ≤ 79.86 gini = 0.15 samples = 82.9% value = [0.92, 0.08] class = No
Area_Change ≤ -2.01 gini = 0.46 samples = 2.4% value = [0.64, 0.36] class = No
Col_Change ≤ 22.5 gini = 0.27 samples = 3.6% value = [0.84, 0.16] class = No
gini = 0.37 samples = 5.1% value = [0.76, 0.24] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
gini = 0.44 samples = 0.2% value = [0.33, 0.67] class = Yes
gini = 0.08 samples = 5.3% value = [0.96, 0.04] class = No
gini = 0.15 samples = 82.7% value = [0.92, 0.08] class = No
gini = 0.5 samples = 0.1% value = [0.5, 0.5] class = No
gini = 0.13 samples = 1.0% value = [0.93, 0.07] class = No
gini = 0.49 samples = 1.4% value = [0.42, 0.58] class = Yes
gini = 0.23 samples = 3.4% value = [0.87, 0.13] class = No
gini = 0.44 samples = 0.2% value = [0.33, 0.67] class = Yes

Figure 5.48: Decision tree explaining the newspaper ending scenarios

Since the tree is cut at the depth of four for representational purpose, the decisions nodes are not really explainable, but if the tree grows until the depth of 8 or 9, a more detailed representation of the data appears. Following is the tree with depth 8, proving to be more detailed.

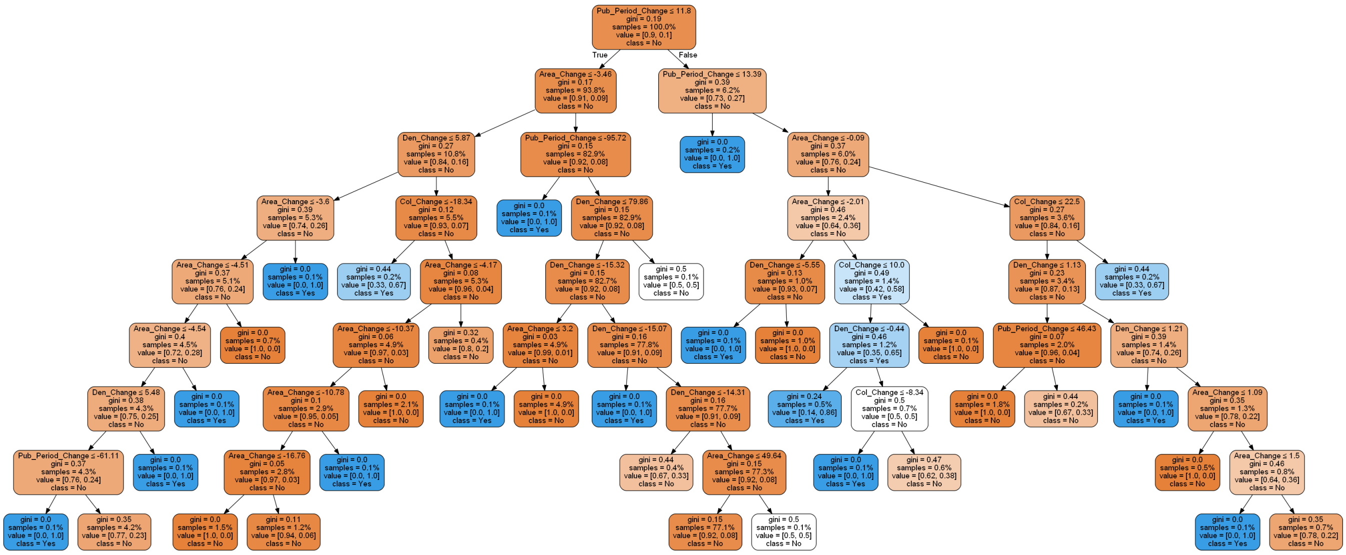

Pub_Period_Change ≤ 11.8 gini = 0.19 samples = 100.0% value = [0.9, 0.1] class = No
True
False
Area_Change ≤ -3.46 gini = 0.17 samples = 93.8% value = [0.91, 0.09] class = No
Pub_Period_Change ≤ 13.39 gini = 0.39 samples = 6.2% value = [0.73, 0.27] class = No
Den_Change ≤ 5.87 gini = 0.27 samples = 10.8% value = [0.84, 0.16] class = No
Pub_Period_Change ≤ -95.72 gini = 0.15 samples = 82.9% value = [0.92, 0.08] class = No
gini = 0.0 samples = 0.2% value = [0.0, 1.0] class = Yes
Area_Change ≤ -0.09 gini = 0.37 samples = 6.0% value = [0.76, 0.24] class = No
Area_Change ≤ -3.6 gini = 0.39 samples = 5.3% value = [0.74, 0.26] class = No
Col_Change ≤ -18.34 gini = 0.12 samples = 5.5% value = [0.93, 0.07] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
Den_Change ≤ 79.86 gini = 0.15 samples = 82.9% value = [0.92, 0.08] class = No
Area_Change ≤ -2.01 gini = 0.46 samples = 2.4% value = [0.64, 0.36] class = No
Col_Change ≤ 22.5 gini = 0.27 samples = 3.6% value = [0.84, 0.16] class = No
Area_Change ≤ -4.51 gini = 0.37 samples = 5.1% value = [0.76, 0.24] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
gini = 0.44 samples = 0.2% value = [0.33, 0.67] class = Yes
Area_Change ≤ -4.17 gini = 0.08 samples = 5.3% value = [0.96, 0.04] class = No
Den_Change ≤ -15.32 gini = 0.15 samples = 82.7% value = [0.92, 0.08] class = No
gini = 0.5 samples = 0.1% value = [0.5, 0.5] class = No
Den_Change ≤ -5.55 gini = 0.13 samples = 1.0% value = [0.93, 0.07] class = No
Col_Change ≤ 10.0 gini = 0.49 samples = 1.4% value = [0.42, 0.58] class = Yes
Den_Change ≤ 1.13 gini = 0.23 samples = 3.4% value = [0.87, 0.13] class = No
gini = 0.44 samples = 0.2% value = [0.33, 0.67] class = Yes
Area_Change ≤ -4.54 gini = 0.4 samples = 4.5% value = [0.72, 0.28] class = No
gini = 0.0 samples = 0.7% value = [1.0, 0.0] class = No
Area_Change ≤ -10.37 gini = 0.06 samples = 4.9% value = [0.97, 0.03] class = No
gini = 0.32 samples = 0.4% value = [0.8, 0.2] class = No
Area_Change ≤ 3.2 gini = 0.03 samples = 4.9% value = [0.99, 0.01] class = No
Den_Change ≤ -15.07 gini = 0.16 samples = 77.8% value = [0.91, 0.09] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
gini = 0.0 samples = 1.0% value = [1.0, 0.0] class = No
Den_Change ≤ -0.44 gini = 0.46 samples = 1.2% value = [0.35, 0.65] class = Yes
gini = 0.0 samples = 0.1% value = [1.0, 0.0] class = No
Pub_Period_Change ≤ 46.43 gini = 0.07 samples = 2.0% value = [0.96, 0.04] class = No
Den_Change ≤ 1.21 gini = 0.39 samples = 1.4% value = [0.74, 0.26] class = No
Den_Change ≤ 5.48 gini = 0.38 samples = 4.3% value = [0.75, 0.25] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
Area_Change ≤ -10.78 gini = 0.1 samples = 2.9% value = [0.95, 0.05] class = No
gini = 0.0 samples = 2.1% value = [1.0, 0.0] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
gini = 0.0 samples = 4.9% value = [1.0, 0.0] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
Den_Change ≤ -14.31 gini = 0.16 samples = 77.7% value = [0.91, 0.09] class = No
gini = 0.24 samples = 0.5% value = [0.14, 0.86] class = Yes
Col_Change ≤ -8.34 gini = 0.5 samples = 0.7% value = [0.5, 0.5] class = No
gini = 0.0 samples = 1.8% value = [1.0, 0.0] class = No
gini = 0.44 samples = 0.2% value = [0.67, 0.33] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
Area_Change ≤ 1.09 gini = 0.35 samples = 1.3% value = [0.78, 0.22] class = No
Pub_Period_Change ≤ -61.11 gini = 0.37 samples = 4.3% value = [0.76, 0.24] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
Area_Change ≤ -16.76 gini = 0.05 samples = 2.8% value = [0.97, 0.03] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
gini = 0.44 samples = 0.4% value = [0.67, 0.33] class = No
Area_Change ≤ 49.64 gini = 0.15 samples = 77.3% value = [0.92, 0.08] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
gini = 0.47 samples = 0.6% value = [0.62, 0.38] class = No
gini = 0.0 samples = 0.5% value = [1.0, 0.0] class = No
Area_Change ≤ 1.5 gini = 0.46 samples = 0.8% value = [0.64, 0.36] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
gini = 0.35 samples = 4.2% value = [0.77, 0.23] class = No
gini = 0.0 samples = 1.5% value = [1.0, 0.0] class = No
gini = 0.11 samples = 1.2% value = [0.94, 0.06] class = No
gini = 0.15 samples = 77.1% value = [0.92, 0.08] class = No
gini = 0.5 samples = 0.1% value = [0.5, 0.5] class = No
gini = 0.0 samples = 0.1% value = [0.0, 1.0] class = Yes
gini = 0.35 samples = 0.7% value = [0.78, 0.22] class = No

Figure 5.49: Newspaper ending scenarios with more depth

6. Discussion

Different methods have been studied for the development of newspapers in this research work. The features are explored individually and in combination. The results of these methods are discussed in some detail in the relevant sections above. Following is the detail discussion based on the research questions presented in the beginning of this thesis.

RQ1: Newspaper metadata categories and fluctuation among them based on language, time and place? Since this research work has taken into account only the newspapers published in Swedish or Finnish language, so a clear division of the data is based on language. Considering the newspapers' publishing language, number of Finnish newspapers increased over time but the Swedish newspapers were ahead of Finnish newspapers in the development of various features such as area, publication period, density and columns. Swedish newspapers had longer lives, they had more stable features, and they were dominating in the trends for the most part of nineteenth century. By dividing the geographical locations into small towns, coastal cities, and metropolitan area, it is proven that the small towns had more fluctuations in the publishing trends as compared to coastal cities. There is symmetry in the small towns' trend following ratio between capital region and their nearby main cities whereas coastal cities followed the trends of their nearby main cities. With the increase in number of newspapers, various publishing trends developed but none of them had common popularity among the newspapers of different languages and locations, which shows the widespread changes in features over time.

RQ2: How do newspapers in Finland developed? The initial exploratory analysis suggest that newspapers started publishing once or twice a fortnight but then the publication frequency increased with the passage of time. One reason for this increase presumably be the newspaper as an important source of information at the time. With the number of newspapers increase, various publishing trends emerged. Newspapers started publishing on different page size with varying number of pages. With the time, newspapers started publishing more columns per page though the results suggest that this does not affect the text density ordinarily. The detailed study of these features suggest that the most impacting feature turned out to be 'area'. This feature is the baseline for other feature changes and is the vital element in the development of newspapers. Though there is correlation among features individually, but the results obtained with the regression methods indicate that studying the features jointly, better explains the development of features. With 68% accuracy columns can be predicted with publication period and area and with an accuracy of 80% area can be estimated using publication period, columns and density. An interesting factor, which has already been discussed is, that the number of pages are not effected by change in page size, which shows the trend of publishing a fixed number of pages regardless of the page size, density and columns.

RQ3: What are the common paths of development for long running newspapers? Regression methods are applied to study the feature development of newspapers as discussed in the above sections but to understand the common paths of development it is important to group the newspapers with similar features. Spectral and Hierarchical Clustering methods are applied to group the newspapers. The clustering results suggest that the clusters do not have clear boundaries on single feature because of the widespread values of them. The features' values vary over the time, which is why the clusters have mix values of features, but in the combination of features, the clusters give a better understanding. For example, publication period in cluster one varies from one to eight days, and same is the publication period for cluster two, but when added more features, the clusters show an understandable decision boundaries. The results show when the area increased, the newspapers started publishing more frequently and the newspapers with more density lived short lives as compared to the others. By the end of nineteenth century and in the beginning of twentieth century, the newspapers had more columns, bigger area and frequent publications. There are instances of newspapers' changing their features causing a change in the clusters also, and area is the most important feature in this hypothesis. When other features such as density and publication period are added with area, the newspapers changed the clusters which makes them imperative also. Since area is effecting newspapers throughout the time period, its effects have been studied separately. The changes in area show also how volatile the newspapers were and the results show that the Swedish newspapers were more stable in area compared to Finnish. On the contrary, results show that the cluster with more volatile newspapers have on average longer lives, which indicates the importance of changing features with time. This hypothesis is observed by looking at the data from the start, eighteenth century, until the twentieth century. The newspapers were publishing once a week with one column in the beginning, but with the passage of time, the publishing trends change to more frequent publication and more columns, so the newspapers which were following the change in trends lived longer as compared to the newspapers with not adapting the changes. This hypothesis is purely based on the hierarchical clustering of area change.

RQ4: Study the dynamics of processes and find out different kinds of lives of patterns in the development of newspapers? The changes in various features such as, columns, publication period, density and pages, were also studied to predict the page size. The logical decisions on each feature are made, which classifies the page sizes of the newspapers. Publication period and columns are the certain reasons to switch between A4 and A5 page sizes. The results show that publication period, density and columns are effecting to switch to A2 page size whereas when newspapers switched to A3, among other features, number of pages were also decisive. These results are obtained by applying decision trees and the accuracy of this method is 77% which is presumably a good estimate regarding the heterogeneity of the data. Similarly, the decision trees are applied on the features to understand the changes, which lead to the end of newspaper publication. Going in depth, each newspaper has different set of changes before ending the publication. There are newspapers, which decreased their publication period 96% and changed the area significantly before stopped publishing, while in other instances, changes in density and columns were noted before the end of newspaper. The newspapers which changed their columns more than 22% along other features change, stopped publishing after that and in some cases the newspapers which changed their density but not the area, ended up their publication. These were just few scenarios but the overall variations in the features' change do not provide any fixed set of rules, which will lead the newspaper to stop publication, but the decision trees give some estimate of the changes in features, which can be the reasons for the end of a newspaper. This research work gives the insight of the development of features, differences in newspapers based on language and publication place and effects of features on each other. The decision trees give the logical divisions of the features and clustering helps in evaluation of features similarities.

RQ5: Study and apply statistical methods to interpret the newspaper metadata and find the similarities among newspapers based on features?

Various statistical methods are studied and evaluated on this dataset. The selection of these methods was based on the literature review, research questions and exploratory analysis. The literature review rationalizes the appropriateness of clustering and decision tree algorithms for this dataset. The research questions are motivated to analyze the development of features and their correlation and based on the results, regression algorithms are evaluated. The Pearson and Spearman correlation methods confirm the correlations between different features. For the features which are linearly dependent on each other, such as density and area, linear regression is applied. This method is the simplest regression algorithm which suits this study. For the experimental purpose, Polynomial Regression algorithm is applied on density and area, but the results were not satisfactory. Another rationale to apply polynomial regression is correlation coefficient value between density and area, which does not indicate a strong relation but the sensitivity of this algorithm towards outliers could not fit the data well. The best possible results are achieved by linear regression algorithm which are discussed in detail in chapter

4. Other features such as columns, publication period and area are dependent on more than one features, for that purpose OLS method is applied. Since the OLS method tries to minimize the sum of all squared residuals, this method is apt for the multiple regression analysis.

Since the data is collected and mapped from multiple sources, it inherited the complexity and ambiguity. To analyze the common patterns among this data, clustering seems the right method. There are different clustering algorithms, so multiple algorithms are studied in detail and based on their properties and exploratory analysis of data, four clustering algorithms, K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Spectral clustering, and Hierarchical Clustering were selected for initial analysis. K-means and DBSCAN were discarded after the initial analysis as they could not achieve satisfactory results. Though k-means is a widely used algorithm, but it has trouble clustering data when the size and density of clusters vary. Another reason for not using this method is that its centroid can be dragged by outliers which gives poor results. DBSCAN comes with the benefit of not selecting the number of clusters beforehand, but it has a drawback, if the objects are placed too close and the ϵ parameter cannot be estimated easily, the algorithm will not be able to cluster the data well. Since the data represents similarities between pair of points, similarity based clustering seems an appropriate choice. Considering this aspect, spectral clustering algorithm is studied in detail and evaluated on the given dataset. The intuition behind spectral clustering is to form a similarity matrix where an entry is the similarity distance between data points instead of directly clustering them in their native data space. The given dataset contains feature values within same range with minor changes, for example, two newspapers publishing eight pages every fourth day with three columns per page but with varying density. So spectral clustering considers these two data points as nodes and finds the similarities among them. The mathematical formulation of this algorithm is given in chapter 5.

The exploratory analysis suggests the importance of area feature. Using this feature only, hierarchical clustering algorithm is evaluated to cluster the newspapers. This algorithm offers a complete range of nested cluster solutions based on group similarities. The rationale to use this algorithm to study area feature is to analyze at what point two newspapers with fluctuating area merge. It is a powerful technique to build tree structures using data similarities as shown in chapter 5. This algorithm tells how different sub-clusters are related and how far apart data points are. Since hierarchical clustering algorithm does not need to know the number of clusters beforehand, it suited well with the area feature data. It is easily interpretable at different level by just cutting the dendogram at a specific point.

For two case studies, classify page (A2, A3, A4, A5) and classify the end of newspapers, decision trees were evaluated. The decision trees help to reach on a decision with logical conditions. The reason to use decision trees for the two case studies was to not only classify the page and end of a newspaper but to understand the changes in different attributes. It helps to learn the path of development for these attributes. Out of several available decision tree algorithms, Random Forest was selected for both case studies. The selection of this algorithm is based on many features such as it is not influenced by outliers, it does not make any assumption about the underlying distribution of data, and it typically provides high accuracy without overfitting. To classify page, four features are selected. The objective is to learn which changes in these features lead to a specific page. This algorithm also helps in answering the research question of feature development. For the second case study, this algorithm suits really well as the classification problem becomes binary now. The algorithm learns the attribute changes and classifies the end of a newspaper. One option to solve this classification problem is to use a simple logistic regression, which might perform better than this algorithm, but prediction is not the point of interest in this case study. The objective is to learn the changes in different features and understand what might be the cause of newspaper ending. The random forest algorithm performs well for both the case studies as shown in chapter 4 and 5, and also helps in understanding the behavior of the attributes.

The exploratory analysis and different statistical algorithms are studied and based on their evaluation and discussion, five research questions are answered in detail. Though these research questions are at a more general level, but the process of finding the answers not only gives a comprehensive description of this dataset but also helps in understanding the behavior of various machine learning algorithms.

7. Conclusions and Future Work

Digitizing historical newspapers is challenging because of the complex structures, nature of the documents and varying layouts. Working with the historical heterogeneous raw data can be tedious and for that purpose, the Finnish National Library has processed and extracted content and metadata of the newspapers. The historical metadata from late eighteenth century to early twentieth century has been studied to explore various features and analyze the trajectories among them. Various statistical methods are applied on this data set to gain meaningful insights of the data. Newspapers are clustered based on the similarities of their features and decision trees are constructed based on logical divisions of the features. The results give a good understanding of the importance of various features and a good perspective of newspapers published in different languages and in different cities. Along the exploratory analysis and clustering of newspapers, feature development leading to the end of newspapers have been studied also to understand the trajectories in the last years of newspapers.

Future Work:

The layouts and development of features, which are the results of this research can be used for further research by merging them with content of the newspapers. This lays out the basis for numerous research questions, such as, 1) Automatically analyze structural changes and classify the genres, 2) Identify latent categories based on the feature development and structural changes, 3) Analyze the changes in features to identify visual elements in advertisements.

Bibliography

[1] A. Abdullin and N. Olfa. Clustering heterogeneous data sets. *2012 Eighth Latin American Web Congress, Cartagena de Indias*, pages 1–8, 2012.

[2] F. Bach and M. Jordan. Learning spectral clustering. *Advances in Neural Information Processing Systems 16 (NIPS), MIT Press, Cambridge*, pages 305–312, 2004.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *in Proceedings of the eleventh annual conference on Computational learning theory, New York, NY, USA: ACM*, pages pp.92–100, 1998.

[4] D. Breitkreutz. Clusterers: a comparison of partitioning and density-based algorithms and a discussion of optimisations. 2008.

[5] J. Brownlee. A gentle introduction to normality tests in python. *Machine Learning Mastery*, 2019. https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/.

[6] S. Dang. Performance evaluation of clustering algorithm using different datasets. *International Journal of Advance Research in Computer Science And Management Studies*, pages 167–173, 2015.

[7] M. Elbatta. An improvement for dbscan algorithm for best results in varied densities. *Dissertation, Gaza (PS): Islamic University of Gaza.*, 2012.

[8] a. K. H. Ester, M, J. Sander, and X. X. A density-based algorithm for discovering clusters in large spatial databases with noise. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press*, pages 226–231, 1996.

[9] M. Gashler, C. Giraud-Carrier, and T. Martinez. Decision tree ensemble: Small heterogeneous is better than large homogeneous. *Seventh international conference on Machine Learning and Applications*, pages 900–905, 2008.

[10] M. Gonen. Embedding heterogeneous data by preserving multiple kernels. *In ECAI 2014 - 21st European Conference on Artificial Intelligence, Including Prestigious Applications of Intelligent Systems, PAIS 2014, Proceedings*, 263:pp 381–386 (Frontiers in Artificial Intelligence and Applications; Vol. 263), 2005.

[11] K. Grabczewski and W. Duch. Heterogenous forests of decision trees. *in Lecture Notes in Computer Science Artificial Neural Networks, U.K., London:Springer-Verlag*, volume 2415:pages 504–509, 2002.

[12] A. Halevy. Why your data don't mix. *ACM Queue*, 3(8):50–58, 2005.

[13] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: Data mining, inference, and prediction. 2nd edition new york city, usa:. *Springer*, 2009.

[14] P. J, Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, volume 20:pages: 53–65 ISSN:0377–0427, 1987.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani. An introduction to statistical learning. *Springer, New York*, 2013.

[16] R. Jin and H. Liu. A novel approach to model generation for heterogeneous data classification. *In IJCAI International Joint Conference on Artificial Intelligence*, pages 746–751, 2005.

[17] V. Jirkovský and M. Obitko. Semantic heterogeneity reduction for big data in industrial automation. *Information Technologies - Applications and Theory ITAT*, 2014.

[18] W. Koehrsen. Random forest simple explanation. 2017. https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d.

[19] Z. Li and M. deRijke. The impact of linkage methods in hierarchical clustering for active learning to rank. *In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 941–944, 2017.

[20] v. Luxburg. A tutorial on spectral clustering. *Stat Comput 17*, pages 395–416, 2007.

[21] M. M, Mukaka. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi*, 24,3, 2012.

[22] T. M, Kodinariya and P. R, Makwana. Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies, 1(6)*, pages 90–95, 2013.

[23] E. Mäkelä, M. Tolonen, J. Marjanen, A. Kanner, V. Vaara, and L. Lahti. Interdisciplinary collaboration in studying newspaper materiality. *Digital Humanities in the Nordic Countries, Copenhagen*, CEUR-WS 2365, 2019.

[24] A. Mirzaei and M. Rahmati. A novel hierarchical-clustering-combination scheme based on fuzzy-similarity relations. *Fuzzy Systems, IEEE Transactions*, volume 18:pages 27–39, 2010.

[25] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering. *analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14. MIT Press*, 2002.

[26] T. Pääkkönen, J. Kervinen, A. Nivala, K. Kettunen, and E. Mäkelä. Exporting finnish digitized historical newspaper contents for offline use. *D-Lib Magazine*, 22, 2016.

[27] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, 2000.

[28] J. Podani. Simulation of random dendrograms and comparison tests: Some comments. *Journal of Classification*, volume 17:pages 123–142, 2000.

[29] N. Rahmah and I. S, Sitanggang. Determination of optimal epsilon (eps) value on DBSCAN algorithm to clustering data on peatland hotspots in sumatra. *IOP Conference Series: Earth and Environmental Science*, 31:012012, 2016.

[30] N. Randriamihamison, N. Vialaneix, and P. Neuvial. Applicability and interpretability of hierarchical agglomerative clustering with or without contiguity constraints, 2019.

[31] J. Sander. Density-based clustering. *In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA*, 2011.

[32] S. Saraçli, N. Doğan, and I. Doğan. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, pages 1–8, 2013.

[33] P. Schober, C. Boer, and L. Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia Analgesia*, 126 - Issue 5, May 2018.

[34] K. Singh, M. Dimple, and N. Sharma. Evolving limitations in k-means algorithm in data mining and their removal. *IJCEM International Journal of Computational Engineering Management*, pages 2230–7893, 2011.

[35] L. Wang. Heterogeneous data and big data analytics. *Automatic Control and Information Sciences*, 3(1):8–15, 2017.

[36] L. Zheng, T. Li, and C. Ding. A framework for hierarchical ensemble clustering. *ACM Trans. Knowl. Discov. Data 9, 2, Article 9 (September 2014), 23 pages.*, 2014.

[37] D. Zhou and C. J. C, Burges. Spectral clustering and transductive learning with multiple views. *in Proceedings of the 24th international conference on Machine learning, ICML 07. ACM,*, pages pp.1159–1166, 2007.