Prediction of Selection Decision of Document Using Bibliographic Data at the National Library of France (BnF)

Ahmed Ben Salah, Geneviève Cron, Nicolas Ragot, Thierry Paquet

Prediction of Selection Decision of Document Using Bibliographic Data at the National Library of France (BnF)

Ahmed Ben Salah (1 2), Geneviève Cron (1), Nicolas Ragot (3), Thierry Paquet (2); (1) Bibliothèque Nationale de France, (2) Université de Rouen - Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes LITIS EA 4108 (F), (3) Université François Rabelais Tours - Laboratoire d'Informatique LI EA 6300.

Abstract

The selection process of the documents is a very important step in mass digitization projects. This is especially true at the BnF, where the digitization should include or not OCRization depending on the OCR results expected. Consequently, the selection task is very complex and time consuming due to the number of documents to be processed and the diversity of the selection criteria to consider.

trying to improve and simplify this task by automation, we studied the relationship between bibliographic data and the selection decisions of documents. We used two statistical analysis : a factor analysis of correspondence and a multiple correspondence analysis. Our analysis has shown that, for example, the documents in format "4 or GR FOL" and edited "between 1961 and 1990" in Morocco are more likely to be "Selected". However, the documents in format "16 or 8" and edited "between 1871 and 1800 in English or Spanish have a greater chance to be "Not Selected".

Introduction

Since 1992, the National Library of France has started a number of projects of mass digitization and since 2006, these programs include optical character recognition (OCR) of texts. The main objective of these projects is to preserve patrimonial documents and disclose their information to the public. These projects always begin by a selection step of documents which depends on physical and bibliographic criteria. This task is very complex due to, first, the time-consuming physical manipulation of documents ; second, the large number of documents in these projects and third the diversity of the selection criteria. Furthermore, our first study showed that only 10% of documents processed by the selection department at BnF are finally selected which make this task crucial. Physical aspects (such as quality of ink, Document opening, nature of paper....) are known as major impact factor on OCR efficiency. Unfortunatly, these data are not available since the document hasn't been manipulated by staff responsible for the selection process. This manipulation step by a human operator cannot be discarded. Consequently, the optimization of the selection should be done before and physical aspects criteria cannot be used easily. After several years of working in documents selection department, the staff in charge of selecting documents began to put some assumptions about the correlation relationship between some bibliographic data and the physical quality of documents. Trying to simplify and improve the task of documents selection, we took into account these hypothesis and we have studied the relationships between the bibliographic data and the documents selection decisions.

In this respect, we present first in this article how is performed the selection at the BnF. Next, the digitization process itself is shortly presented. Finally, we provide the results of our analysis based on Correspondence Factor Analysis (CFA) and Multiple Correspondence Analysis (MCA).

Documents selection process

The documents selection process is a major step in the process of scanning documents. Indeed, a compromise as to be done to select the "interesting" documents considering the mass of documents available and several criteria. At the BnF, the selection process begins with an annual planning of documents according to their themes and to their intellectual values. Then, the departments of document selection do three tests :

- Test of copyright : the scanned documents at the BNF must be royalty free. This criteria is essential in the selection decision of document.
- Test of the quality of bibliographic data : the bibliographic metadata of the documents must be correct complete and accurate. This criteria is also very important in the selection decision of document.
- Test of the physical quality of the documents : the physical quality of the document must be evaluated to avoid sending to providers documents with too poor physical quality since OCR will not work properly.

At the end of these tests, the documents selection decisions can be either "Not selected", "Selected for Raw OCR" or "Selected for High Quality (HQ)". Raw quality is obtained after the automatic processing of OCR systems, i.e. without human corrections. HQ is the corrected results of OCR systems to obtain an editorial quality. On these HQ digital documents, the BNF requires the final OCR rate to be higher or equal to 99.95%.

This selection task is very difficult and costly in terms of working time and staff number. Indeed, it requires great skill and a massive processing of documents. Moreover, a study showed that only 10% of documents processed by selection departments are selected (in Raw or HQ). This is the reason why the Bnf would like to improve this process to select more documents and at the same to save time and money (obtaining HQ documents cost more than Raw documents). Considering these objectives, optimal operations would affect the documents with a medium physical quality to the necessity of having a High Quality (HQ) digitization, and documents with very good physical quality to a digitization with Raw quality (see Figure 1).

Documents digitization process at BnF

Once the documents selected, they are sent to a digitization provider. For each scanned document, the provider builds a digi-

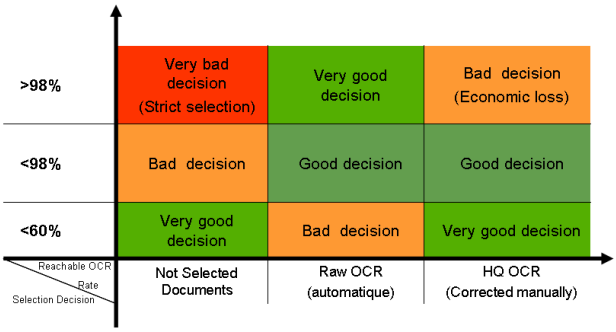

FIGURE 1. *The judgment of the selection decision of documents according to their reachable OCR raw rate*

tal directory that contains the *TIFF* images of document pages and a "*Refnum*" file containing documents metadata. After scanning, the provider sends the digital documents to the BnF. The computer department supports the integration of these digital documents in the internal servers and monitors the file Refnum through automatic tools for verifying physical structure (number and order of pages, orientation, etc.). In parallel, the digitization department at the BNF provides bibliographic control of the metadata of digital documents. The department also controls the conformity of the quality of document images with the BnF's sepecifications. Depending on the type and the number of errors detected during these controls the digital document will be validated or rejected. All documents containing at least one major error such as error in ALTO [1] file structure will be rejected and all documents that contain more than 7 minor errors such as curved effect of the image of the page will be rejected At the end of this process the document will be either returned to the provider for reprocessing or available on *Gallica* [2].

Bibliographic data analysis

To facilitate and improve the selection task, we intend to estimate the selection decision document before performing the selection task using only bibliographic data. So we tried to evaluate the relationships between these metadata and the possible selection decisions.

We used for this analysis a dataset of 20411 documents taken randomly from a database that contains 50000 documents obtained through an internal application used by selection documents department at the BnF. In this set, the amount of documents "*Selected in HQ*", is very low compared to the number of documents "*Not Selected*" and "*Selected in Raw*".This fact skew our study. To overcome this problem, we merged both "Selected" classes into one. This leads to only 2 variables for the decision value : "*Selected*" and "*Not Selected*".

A preliminary work guided us to focus on two factors involved in the selection decision of documents : the edition date and format of documents. The choice of these variables was also motivated by the observation of staff who are responsible for the documents selection that say :

- Old documents have more physical defects than new documents,
- Small documents have poor quality of writing and are more difficult to digitized.

Consequently, to conduct our study, we first applied a correspondence factor analysis (CFA) on two contingency tables :

- one table describing the link between the publishing date and the selection decision documents ;
- one table describing the link between the format and the selection decision documents

Then, in a second step, to have a more precise idea, we used a multiple correspondance analysis (MCA) on a data array that contains more qualitative variables.

Influence of edition dates and format on selection decision

The CFA is a statistical technique that analyses two qualitative variables intersected in a PivotTable. This method allows to study the correlation between modalities of each variable by referring to the model of independence of studied variables. According to [1][2], we begin our factorial analysis by the construction of the adjusted clouds of column profiles and line profiles. Then in the manner of Principal Component Analysis (PCA) and to explore the links between the two studied variables, we proceed to the determination of the factorial axes that maximize the inertia of the lines cloud and columns cloud. The simultaneous representation of profiles lines and profiles columns is possible due to the transformation equations from line profiles to columns profiles and vice versa. This representation simplifies a lot the interpretation of these connections since we can display the modalities of each variable on the same graph.

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i\bullet}} G_s(j) \tag{1}$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{\bullet j}} F_s(i) \tag{2}$$

where :

- $F_s(i)$ the coordinate of the profile i line on the axis of rank s.
- $G_s(j)$ the coordinate of the profile j column on the axis of rank s.
- λ_s the inertia of the cloud of lines (resp. cloud of columns) projected onto the axis of rank s in the values space of columns R_j (resp. the value space of lines r_i).

Edition dates study

We begin by studying the relationship between edition dates and the selection decisions of the documents. For this, we applied a test of independence on the table 1 that crosses dates of editions with the selection decisions of document.

The critical probability of this test is very low 10^{-16} which makes the interest of the CFA on these data beyond doubt. The two intersected variables are linked since the test of independence of χ^2 is different from zero ($\chi^2 = 359.703 \neq 0$). Nevertheless, this relationship is very low since the intensity of the connection measured by the Cramer's V is low ($V = 0.1370$). Anyway, we will analyze the Table 1 by a CFA to determine the modalities that correspond to more selected documents than rejected ones. The

1. ALTO file is an XML file that contains the results of OCR systems. This format is managed by the Library of Congress http ://www.loc.gov
2. http ://www.gallica.bnf.fr

Table 1 : Contingency table which crosses the edition dates of the document with the selection decisions of document

	Selected	Not Selected
1781-1810	17	230
1811-1840	27	389
1841-1870	34	1345
1871-1900	85	4369
1901-1930	59	3674
1931-1960	134	2350
1961-1990	405	4986

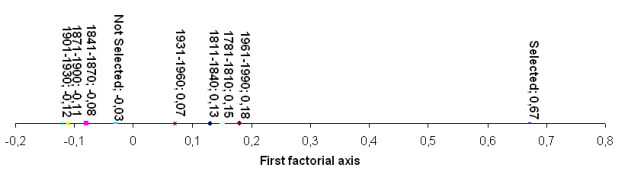

FIGURE 2. *Simultaneous representation of edition date and documents selection decision*

Table 2 : Contingency table which crosses the formats of the document with the selection decisions of document

	Selected	Not Selected
4	568	4466
8	25	5265
16	14	4944
FOL	118	4673
GR FOL	64	273

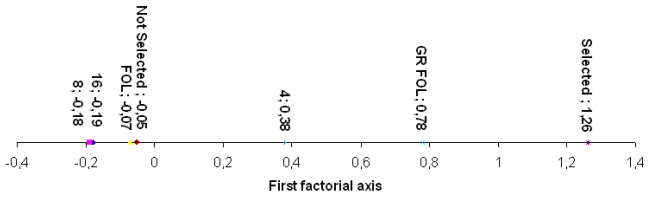

FIGURE 3. *Simultaneous representation of Format and documents selection decision*

maximum number of axes needed to represent the documents perfectly is equal to $\min\{(card(lines)-1),(card(columns)-1)\}$. Factor analysis of Table 1 gave two factorial axes. The percentage of inertia associated to the first axis is close to 100%. On the one hand, the terms of the variable edition dates "*1871-1900*", "*1901-1930*" and "*1961-1990*" have more weight (79%) in the construction of the first axis compared to the other edition dates. On the other hand, we also find that the selection decision "*Selected*" has a significant contribution of 96% in the construction of the first axis (the modality "Not selected" contributes only for 3.97%). Figure 2 shows at the same time the modalities in lines and the modalities in columns on the same figure. This allows us to build links between the terms "date of edition" and the modalities "decisions selection of the document". According to this simultaneous representation, we notice that all the intervals of the edition date are close to the decision "*Not selected*". This observation is due to the large number of documents rejected in each date of edition. Nevertheless, the representation of edition dates between "*1781-1840*" and between "*1931-1990*" are more likely to be selected than others. In the same way, documents published between 1841 and 1930 are much more rejected. However one should keep in mind that the intensity of these relationships are weak and we cannot use them alone as predicitive variables. This weakness cloud be easily explained by the fact that other criteria should be considered for the selection decision.

Document Formats study

In terms of bookbinding we often talk with the terms of format to describe the size of documents. According to internal documentation of the BNF, we use the format "*in-folio (FOL)*" when we fold the press papers once, "*in-quarto (8)*" in the case when we fold the press papers twice, "*in-octavo (4)*" in the case when we fold the press papers three times and "*in-sixteen (16)*" in the case when we fold the press papers four or more times. In the case where we do not fold the press papers we use the document format "*in-plano (GR FOL)*".

As we specified in the first part of this section, we want to study the relationship between document formats and the selection decisions. For this, we constructed Table 2 that intersects document formats with the selection decisions of documents.

The test of independence applied on Table 2 showed that the two variables are linked since the value of χ^2 is greater than zero ($\chi^2 = 1313.304$). The critical probability of this test is less than 2.2×10^{-16} which makes the interest of the CFA on these data beyond doubt.

Even if the intensity of the connection between the two variables is weak $V-Cramer = 0.25366 < 0.5$, we have chosen to analyse our data by a correspondence factor analyse in order to identify the relationships between formats and their selection decision.

The factorial analysis of Table 2 has given two factorial axes. The clouds of rows and columns are well presented by the first factorial axis since it expresses around 100% of the total inertia.

According to the Figure 3, the first factorial axis separates the two terms "*Not Selected*" and "*Selected*". We note also that the "*not selected*" documents are close to the origin of the first axis. In fact, the large numbers of the not selected documents compared with the number of selected documents influence the calculation of the independence model.

On this figure, we can also visualize and interpret the connections that exist between the two variables. We notice that the documents in "*16*", "*8*" and "*FOL*" have a tendency to be not selected since these representations are close to the representation of modality "*Not Selected*". This proximity can be interpreted by the lack of selected documents in the categories of document format.

Documents in format "*GR FOL*" and "*4*" have positive coordinates on the first axis. These formats have more selected documents compared to other formats. The calculation of the contributions of rows and columns in the construction of the factorial axes allows us to select the main points contributory in the construction of the principal axis. The modality "*Selected*" has an important contribution of 96.13% so we can say that the first axis represent perfectly the decisions of document's acceptance. The format "*4*" contributes decisively in the construction of the first axis. These document formats are a feature of selected documents. We can conclude, according to this analysis that document formats ("*GR FOL*" and "*4*") characterize the selected documents.

But, this proximity does not assert that all documents which possesses those format will be selected due, firstly, to the weakness of their relationship intensities and secondly to the complexity of the document selection decision which depends on several factors at the same time. This is the reason why we decided to operate a Multiple Correspondence Analysis.

Multiple correspondence analysis (MCA)

The MCA offer the opportunity to deal with many variables. In fact, it can be applied on complete disjunctive table that intersect individuals (documents) in lines with qualitative variables (Bibliographic data) in columns. The value of the elements x_{ik} is 1 if the individual i has the modality k and 0 if not. The size of this table is equal to $I \times K$ with I is the number of documents and K is the number of modalities of our variable. The main focus of the MCA is the study of the relationships between modalities and variables. As in Principal Component Analysis (PCA) [1][2][3], we try to draw the links between variables. These links can be studied either two by two as in CFA or globally.

MCA method

According to [1], studying the similarities between modalities, implies to define the distance which separates them. Let two modalities k and k' assimilated to each group of individuals. One way to compare these two modalities is to count the individuals who have one or both of these modalities.
As in the CFA, we should begin our analysis with the construction of the clouds of modalities and variables using the following formulas :

$$f_{ik} = \frac{x_{ik}}{I \times J} \tag{3}$$

$$f_{\bullet k} = \sum_{i=1}^{I} \frac{x_{ik}}{I \times J} = \frac{I_k}{(I \times J)} \tag{4}$$

$$f_{i\bullet} = \sum_{k=1}^{K} \frac{x_{ik}}{(I \times J)} = \frac{1}{I} \tag{5}$$

Where :

- $x_{ij} \in \{0,1\}$
- I the number of documents
- J the number of variables
- $i \in [1...I]$
- $j \in [1...J]$

The distance between two modalities of complete disjunctive table is calculated using the following formula :

$$d^2_{\chi^2}(k,k') = \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} \left(\frac{f_{ik}}{f_{\bullet k}} - \frac{f_{ik'}}{f_{\bullet k'}} \right) \tag{6}$$

Then, as in PCA, we try to find the axes that maximize the projected inertia of the cloud of variables and modalities. If we use orthogonal axes, we can have plans that maximize the inertia of the cloud of variables and modalities. The eigen vectors are by definition orthogonal. Moreover, the eigenvalue λ_s can also be interpreted as the inertia of the cloud projected on the axis of rank s. Therefore, the biggest eigenvalue represents the maximum inertia of our cloud of variables or modalities. To use this method, in our study, we constructed a data table which includes 5 variables

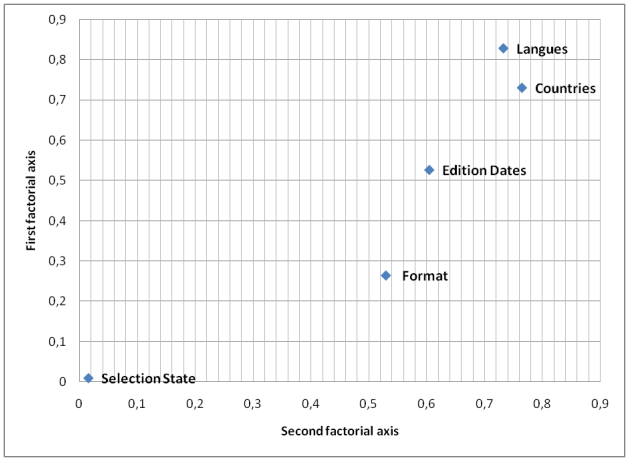

FIGURE 4. *Representation of qualitative variables according to the two first factorial axes*

which are : *Date, Language, Country, Format and Selection State.* we added to our CFA database some informations about the country [3] and language [4] of the documents because we know that the BNF's mission is to treat the French national heritage and especially the French documents. Then, theoretically, documents that contain these two criteria have more chance to be selected. The number of documents examined in this analysis is the same as in our previous studies (20 411 documents).

Results and discussion

We began our analysis by presenting the cloud of variables according to the first two factorial axes. Then, we have refined our analysis by a study of relationships between variables and we analyzed also the relationships between the terms of these variables to clarify our study.
The variables can be plotted by calculating the correlation ratios between the coordinates of individuals on one axis and each of the variables. According to [1], if the correlation ratio between the variable j and axis s is close to 1, this means that individuals (documents) with the same modality (for this qualitative variable) have neighbor coordinates in this axis s.
In our study, according to the plane representation of the cloud of the variables, we find, on the one hand, that variable *Language, Country and Edition date* are related to each of the first two axes (see, Figure 4). On the other hand we note that the qualitative variable *Format* is more related to the first axis than to the second one(see, Figure 4) According to [1], in MCA, the percentage of inertia associated to the first axis is lower than in PCA. This is because, in ACP for example, only linear combination are considered : this mean that only one axis can represent all variables if these are highly correlated with each other. Seemingly in MCA, we study much more general links : if we study two variables that have K_j and K_l modalities we need at least "$\min(K_j, K_l) - 1$" dimensions to represent the relation between two variables.

3. The National Library of France uses the iso code in order to name the country of document http ://www.iso.org/iso/fr/french country names and code elements

4. The National Library of France uses the iso code in order to name the language of document http ://fr.wikipedia.org/wiki/Liste des codes ISO 639-2

In our study, we note that the decrease of the eigenvalues became regular from the third factorial axis. As a consequence, we will interpret here only the first two factorial axes although it is interesting to interpret the following axes. The simultaneous representation of modalities of the qualitative variables and individuals (*Documents*) is not highly efficient. In fact, we obtain a cluttered representation which makes our interpretation difficult and inaccurate (see Figure 5).

Nevertheless, MCA statistics show a positive coordinates for modalities *4* and *16* of variable" *Format*" and a negative coordinate for modality *8*. Furthermore, on the same axis, we also notice that the coordinate of "*Selected*" modality is positive whereas the coordinate of the modality "*Not Selected*" is negative. The automatic description of the second axis also shows that the coordinates of formats *FOL* and *GR FOL* are positive whereas the coordinates of formats *4*, *8* and *16* are negative. Based on these results, we can say that the first axis represents the documents in *16* and *4* (resp. *8*) and which are selected (resp. not selected) after the selection process of documents. We can also deduce that the second axis represents the documents in formats *FOL* and *GR FOL* (resp. *4*, *8* and *16*) that are not selected (resp. selected) by selection departments at the BnF.

Often a multivariate analysis is completed by a univariate analysis to characterize some variables. In this step of our study, we constructed a χ^2 test of independence in order to evaluate the relationship between the variable of interest (in our case "selection state") and other variables. The more this test is significant, the more the modalities considered and the qualitative variables are related. According to [1], the critical probability (*p-value*) measures the significance of the χ^2 test . Therefore, the more the critical probability for the χ^2 test is smaller the more the assumption of independence is questionable and the more the qualitative variable characterizes the variable of interest.

In our study and according to the Table 3 we notice that the variable "*Format*" is the most linked to the variable of interest "*State of selection*".

Table 3 : Description of the variable "*Selection State*" by the other qualitative variables

	Critical probability
Format	4.341334e-283
Country	1.214759e-80
Date	3.869601e-72
Langue	3.660244e-16

Another way of interpreting links between variables is done by the study of links between the modalities of these variables.Then, we characterize in this part each modality of the variable of interest "Selection State" by the modalities of other variables. The "*test value*" is a tool that contributes to the exploratory and descriptive approach of large spreadsheets. To simplify our analysis, we divided our study in two steps. First, we focus on the analysis of the modality "*Not Selected*". For this, we adapted the method presented in [1] which gives a table that contains qualitative variables sorted from the most to the least significant when the modality is over-represented (value-test is then positive) and from the least to the most significant when the modality is under-represented in the class (the value-test is then negative).

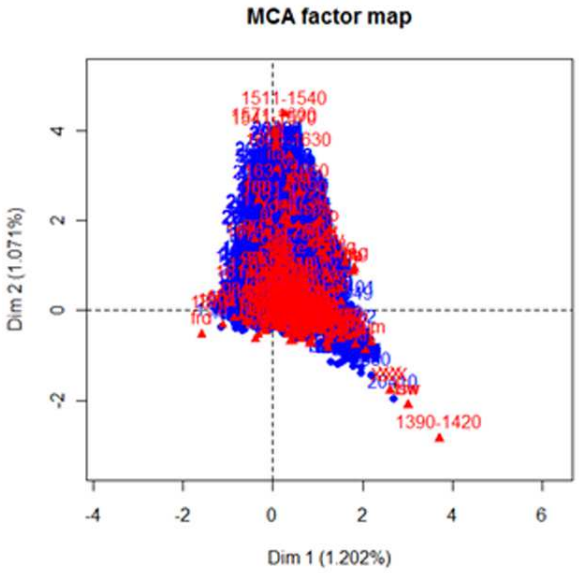

FIGURE 5. *The simultaneous representation of modalities of bibliographic variables and documents*

According to Table 4[5], we find that 99.71% of documents in format "*16*" are not selected ; 25.19% of the not selected documents are in format "*16*" and 24.29% of documents processed in our study are in format "*16*". The critical probability of the test (5.54×10^{-75}) is low and the associated value-test is 18.32. Therefore, we can conclude that format "*16*" characterizes very well the modality "*Not Selected*". We also find that the formats "*16*" and "*8*" and the editions dates "*1901-1930*" characterizes more the documents not selected. We remark also that contrary to what we thought at the beginning of our analysis the French documents characterize more the modality "Not Selected" than the modality "Selected". This result is logical since on the one hand the French documents are the majority of BnF collection (and then a huge part of them can be not selected), and on the other hand, the foreign documents are generally programmed in specials digitization projects in which the BnF tries to select most of the documents. As well to study the modality "Selected", we adopted the same previous analysis.

Table 5[6] shows the results of this operation. According to these results, we found that 11.28% of documents in format "*4*" are accepted ; 72% of documents accepted are in format "*4*" ; 24.66% of documents processed in our study are in format "*4*". The critical probability of our test is (1.4×10^{-179}) and the associated value-test is (28.57). Furthermore, 19% of the documents in "*GR*

5. *Cla/Mod* : The percentage of documents that have the modality "*i*" and that is not selected ;*Mod/Cla* : The percentage of not selected documents that have the modality "*i*" ; *Global* : The percentage of documents having the modality "*i*" ; *p-value* : The critical probability ; *V-test* : The distance between the individual's average having the modality "*i*" and the overall average

6. *Cla/Mod* : The percentage of documents that have the modality "*i*" and that is selected ;*Mod/Cla* : The percentage of selected documents that have the modality "*i*" ; *Global* : The percentage of documents having the modality "*i*" ; *p-value* : The critical probability ; *V-test* : The distance between the individual's average having the modality "*i*" and the overall average

FOL" are selected and 8.11% of selected documents are in format "*GR FOL*"; 1.65% of documents processed in our study are in format "*GR FOL*". The critical probability of this test is equal to ($2,601 \times 10^{-26}$) and the test-value associated is equal to (10.61).Therefore, we conclude that the format "*4*" and "*GR FOL*" characterizes the selected documents.
We notice also that 7.51% of documents printed between "*1961-1990*" are selected and 51.33% of selected documents are printed between "*1961-1990*". 26.41% of documents processed in our study are are printed between "*1961-1990*". The critical probability of our test is (6.3×10^{-52}) and the associated value-test is (15.16).

Table 4 :Description of "*Not Selected*" modality by the modalities of the others qualitative variables

	Cla/Mod	Mod/Cla	Global	p-value	v.test
16	99.71	25.19	24.29	$5.54\ 10^{-75}$	18.32
8	99.52	26.83	25.91	$2.33\ 10^{-69}$	17.6
1901-1930	98.41	18.72	18.29	$1.6\ 10^{-18}$	8.78
1871-1900	98.04	22.26	21.82	$2.23\ 10^{-16}$	8.2
fr	96.97	60.17	59.65	$7.39\ 10^{-14}$	7.4
FOL	97.53	23.81	23.47	$2.12\ 10^{-9}$	5.98
eng	98.34	7.87	7.69	$2.36\ 10^{-7}$	5.16
spa	98.40	4.09	4	$1.98\ 10^{-4}$	3.72
gb	98.86	2.66	2.59	$2.59\ 10^{-4}$	3.65
.....	.....	.....	.....	.....	.....

As a conclusion, from the multiple correspondences analysis made in this section, we can conclude that for example the documents in format "*4* or *GR FOL*" and edited "*between 1961 and 1990*" in "*Morocco(ma)*" are more likely to be "*Selected*". However, the documents in format "*16*" and edited "*between 1901 and 1930*" in "*English (eng)* or *Spanish (spa)*" in "*France (fr)*" have a greater chance to be "*Not Selected*".

Conclusion

In this paper we tried to estimate the selection decision made on BnF's documents. For this, we had studied the relationships between selection decision and bibliographic data. Our analysis has shown that there are relationships between bibliographic data (like Format, Edition date, country and Language) and selection decisions. For example, according to our analysis we have noticed that documents in format "*4*" and edited "*between 1961 and 1990*" in Morocco are more likely to be "*Selected*". However, the documents in format "*16*" and edited "*between 1991 and 2021*" in English (eng) or *Spanish (spa)*" in "*France (fr)*" have a greater chance to be "*Not Selected*".
These relations still weak. In fact, others factors, as the intellectual value of the document and the topic of document digitization

Table 5 :Description of "*Selected*" modality by the modalities of the others qualitative variables (XXX : Language not mentioned, XX : Country not mentioned)

	Cla/Mod	Mod/Cla	Global	p-value	v.test
4	11.28	71.98	24.29	$1.4\ 10^{-179}$	28.57
1961-1990	7.51	51.33	26.41	$6.3\ 10^{-52}$	15.16
GR FOL	19	8.11	1.65	$2.6\ 10^{-26}$	10.61
ma	26.31	3.8	0.55	$6.4\ 10^{-17}$	8.35
XXX	11.04	6.84	2.4	$8.06\ 10^{-12}$	6.83
XX	11	6.84	2.4	$9.46\ 10^{-12}$	6.81
sn	31.48	2.15	0.26	$2\ 10^{-11}$	6.70
dz	13.37	2.66	0.76	$1.45\ 10^{-6}$	3.98
1931-1960	5.39	16.98	12.17	$6.83\ 10^{-5}$	3.98
ota	33.33	0.38	0.04	$8.11\ 10^{-3}$	4.81
.....	.....	.....	.....	.....	.....

projects, are involved in the selection decision of documents. Unfortunately, these data are not annotated in BnF databases. This makes the estimation of selection decisions of documents with bibliographic data difficult. Therefore if we want to provide a tool able to help the selection department of the BnF, we must enrich our data especially with physical and intellectual characteristics.

Références

[1] F. Husson S. Le and J. Pages, Analyse de donnees avec R, Presses Universitaire de Rennes, Rennes, 2009.

[2] B. Escofier J. Pages, Analyses Factorielles simples et multiples, Dunod,Paris,2008.

[3] IT. Jolliffe, Principal Component Analysis, Springer-Verlag, New York (2002).

[4] A. Rafik G. Michael L. Carlo and P. Francesco, Discriminant Multiple Correspondence Analysis, Proceedings of The Italian Statistical Society, Session Factorial Methods, Caserta, 2008.

Author Biography

Ahmed Ben Salah is a PhD student at the University of Rouen. His PhD is performed at the national library of France (BnF) on prediction of OCR rate and automatic control of OCR results. More generally, his research interests include data analyse and pattern recognition.