UNIVERSITAT
ROVIRA i VIRGILI

Assignment 3:

Fuzzy Expert System for Book Sales Prediction

Planning and Approximate Reasoning

João Valério

Eirik Grytøyr

Date: 12/12/2022

Index

1. INTRODUCTION

This report describes the result of an assignment, which aims to design a fuzzy expert system to predict future book sales based on data related to the author, publisher and the time of the year. The process is divided into four tasks: the definition of input and output variables, which rules will be applied to the system, how the system is implemented in MatLab, the results from four different test cases, and, in the last point, some general questions about the system will be discussed.

The choices made in the study are based heavily on data obtained from these two studies [1][2]. It is important to denote that the data in [1] is split into nonfiction and fiction, while in this study, those are combined to an approximated average.

2. TASK 1: DEFINITION OF THE VARIABLES.

Specified in the assignment are four concrete linguistic variables which are shown in the subsequent chapters. The important aspects considered in the creation of the input variables are

- That they satisfy fuzzy partition, meaning that each value of the reference domain adds up to 1.
- That the division of the predicates, as best as possible follows natural clusters in the data.

Finally, the scales are logarithmic as used in the study [1] and the plots are obtained from the mentioned paper.

a. Author visibility

Paper [1] utilizes data from Wikipedia to determine the visibility of the author, in which the average number of daily page views and cumulative visibility is considered.

In the current project, the same strategy will be performed. Figure 2-1 shows the cumulative visibility and its respective book sales after one year.

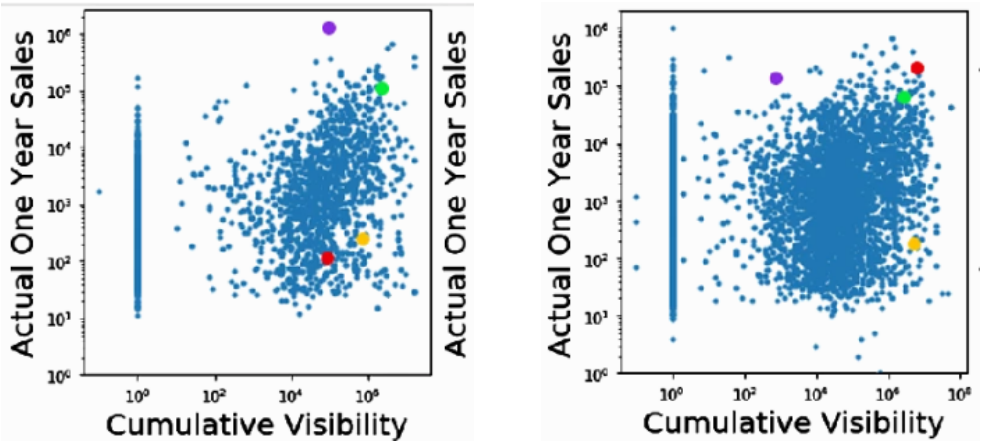


Figure 2-1: Cumulative visibility vs actual book sales after one year for fiction and nonfiction.

In the following plot, it is defined that the scale of the new function should range from 0 to 10^8 cumulative page reviews. The membership function is shown in figure 2-2

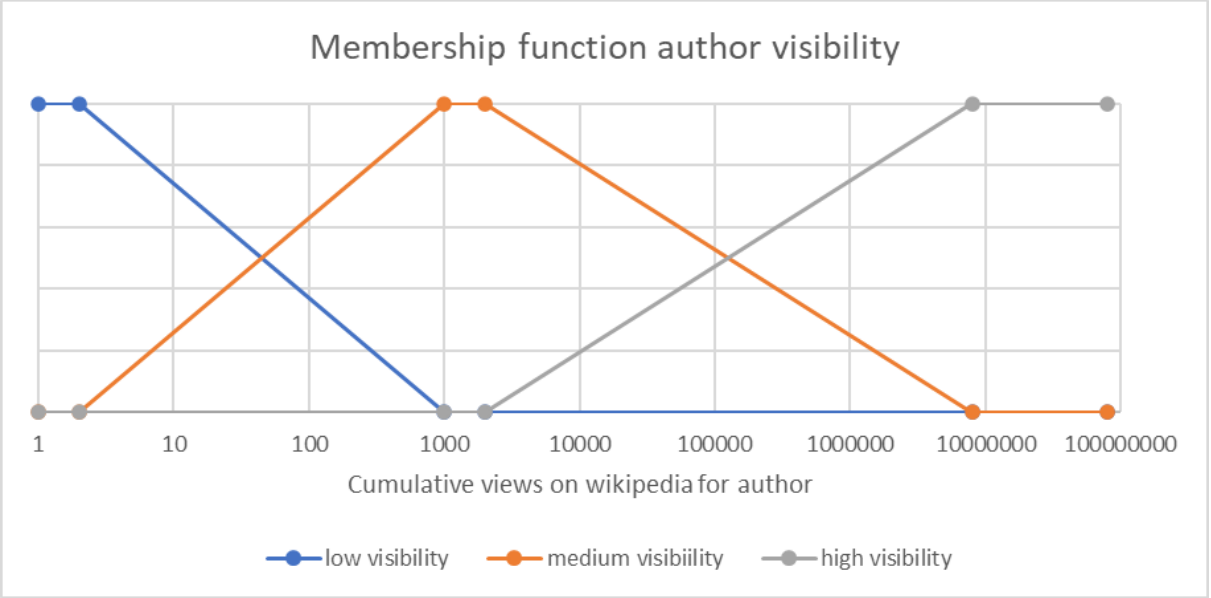


Figure 2-2: Membership function of author visibility.

The variables are divided into low, medium and high visibility. Figure 2-1 shows that it is a large number of observations at around 1 view, while the other cluster is between 1000 and 10^7. It is decided to split the second cluster into medium and high visibility, due to the large number of iterations. Furthermore, as the distribution between those categories is relatively even, the intersection regions are large.

b. Publisher prestige

Paper[1] attributes to the publisher prestige an imprint value as shown in Figure 2-3, where each value represents the one-year median sale of books for the imprint which can be seen as the publisher.

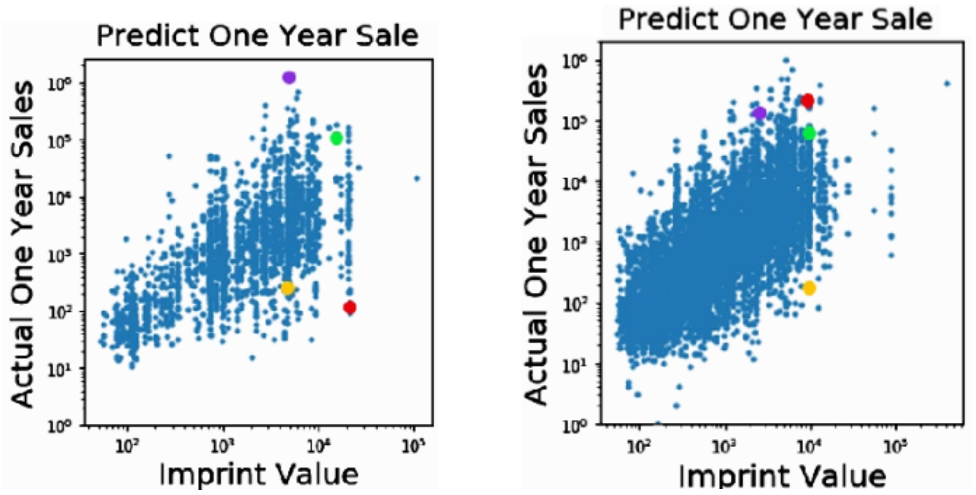


Figure 2-3: Imprint Value vs actual book sale for fiction and nonfiction.

From the plots, it is demonstrated that the majority of examples are in the range of approximately 100 to 11000 average books sold. Therefore as shown in Figure 2-4, the

linguistic variables are defined in this range. Since the distribution in this cluster is relatively even, the medium area without intersection is small.

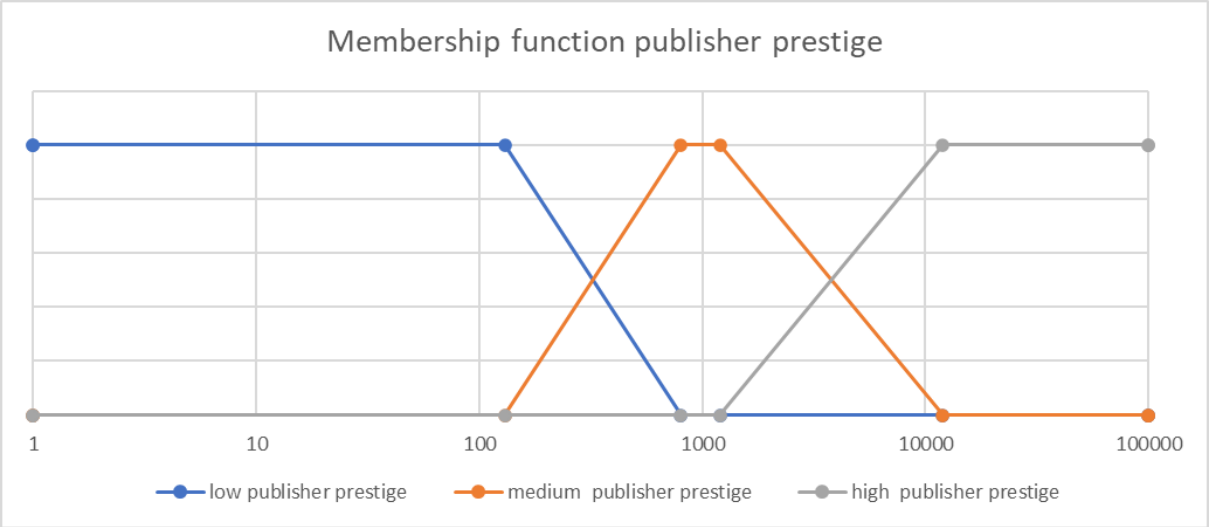


Figure 2-4: Membership function of publisher prestige.

c. Previous sales

Firstly, it is noted that previous sales can be measured in different ways. The study [1] bases this on both previous sales in general for the author in one category, and the sale based on the same genre as another. In this project, the previous sale based on the same genre is used, since it provides more information based on the case where the author has changed genre during his/her career. Thus, Figure 2-5 shows these distributions.

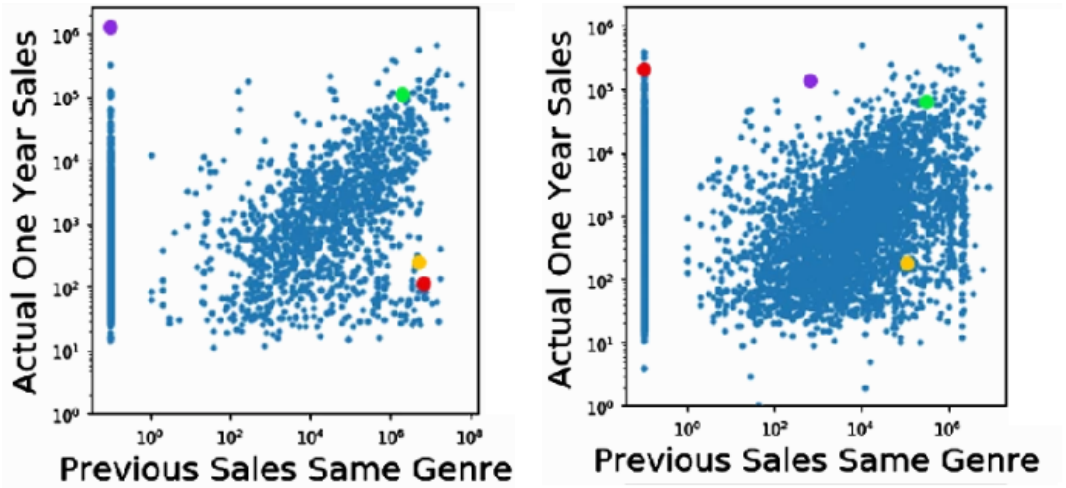


Figure 2-5: Previous sales in the same genre vs actual book sales for fiction and nonfiction.

As the figure illustrates, there is a distinction between the case where the author hasn't published any books and the remaining. Therefore, the membership function will have an extra category for this situation, while low, medium and high will be distributed from 1 to 10^8. Mainly, the distribution will be between 100 and 10^7 as shown in Figure 2-6.

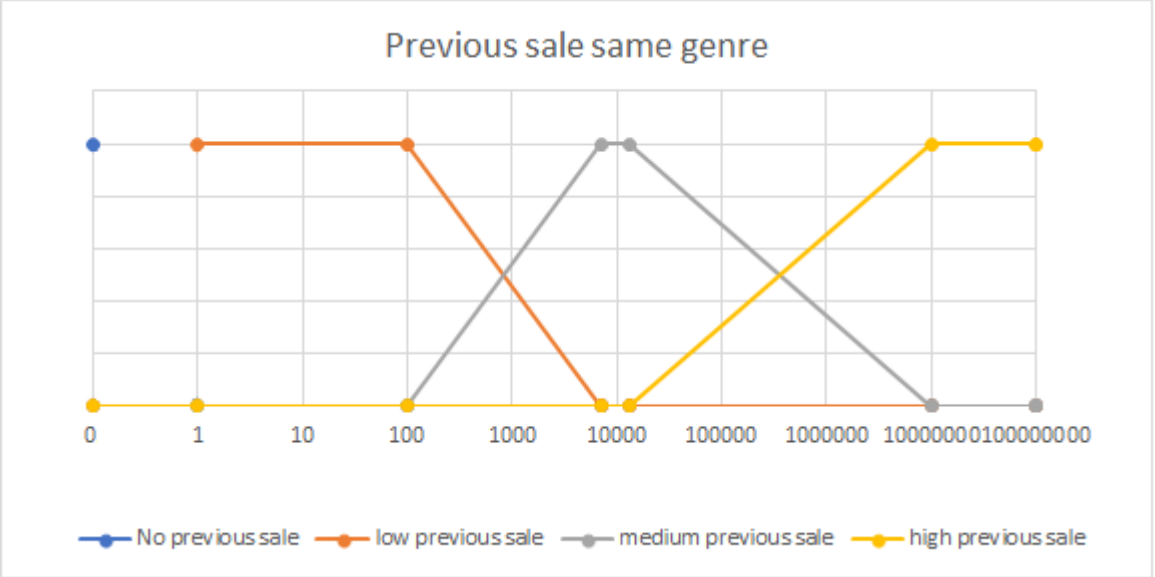


Figure 2-6: Membership function of the previous sale for the same genre.

d. Publishing period of year

The publishing period has an impact on general book sales. As figure 2-7 shows, the sale in October can be more than three times as large as in December, and the trend is fluctuating throughout the year.

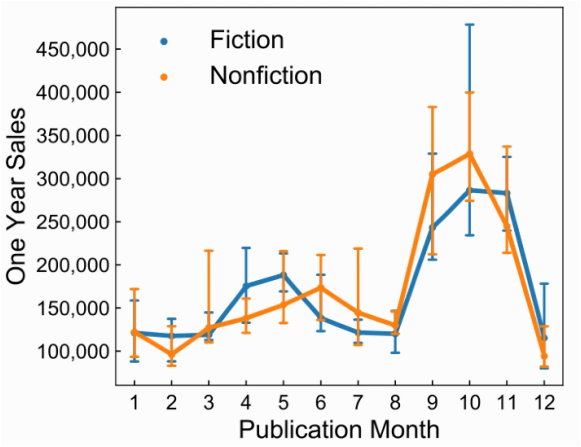


Figure 2-7: Book sale based on the publication month.

One solution to compensate for this would be to cluster all the similar months and convert them into a new scale. So, to facilitate the conversion, each month will be represented by the number from 1 (January) to 12 (December).

The season will be divided into 4: Spring, Summer, Autumn, and Winter where the latter will be divided into February - March and January, in order to keep the scale and the month numbers the same. The result is shown in Figure 2-8.

Figure 2-8: Membership function of seasonal sale.

e. The output variable

The output function that will be linked to the different rules will be a prediction of the sale. The number of linguistic values will be 4 since the study [1] has shown from clustering that the behaviour is preferable to the device into 4. This was also desired by the experts in the study.

The range will be logarithmic and have values between 1 and 10^6, in order to reflect the data from [1] as shown in figure 2-9

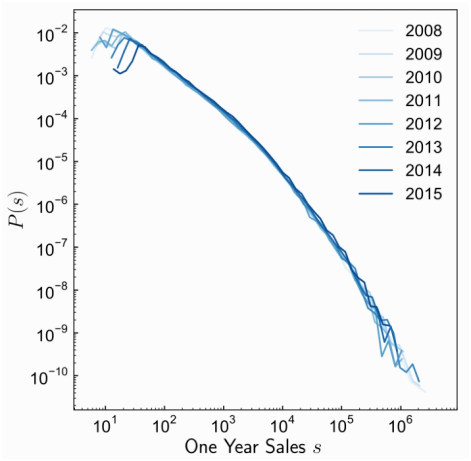


Figure 2-9: Sales distribution of books published between 2008–2015.

The linguistic labels will be low, low intermediate, high intermediate and high sale as shown in figure 2-10.

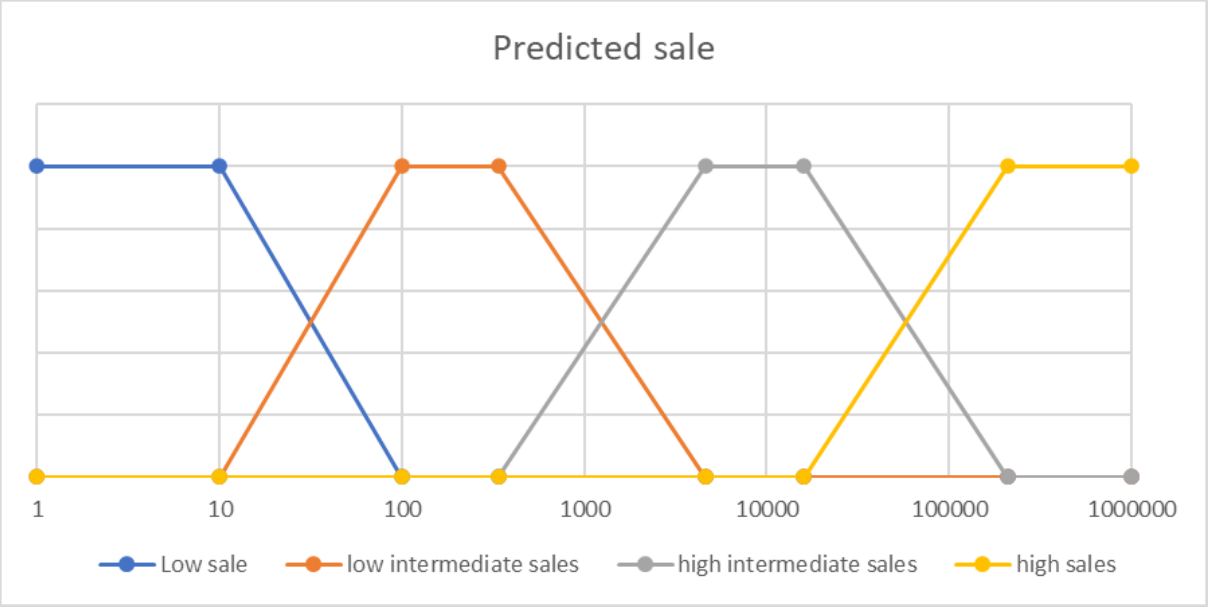


Figure 2-10: Output function of the predicted book sale.

3. TASK 3: DEFINITION OF RULES.

To define the rules, the most significant factor is the importance of each feature and the relations to the output in the study [1]. From this analysis, the importance of each feature is shown in figure 3-1:

Figure 3-1: Importance of each feature in the study [1].

In this fuzzy study, only four of them are used, and the importance of those will be as follows:

1. Publisher prestige/imprint Category
2. Previous sale
3. Visibility of the author

4. publishing period of year

Table 3-1 shows the resulting rules that are explained afterwards.

Table 3-1 – The rules

	Autor visibility	Publisher prestige	Previous sale	Publishing period	weight	Output sale
1	low				0.21	low
2	high				0.21	high
3		low			0.25	low
4		high			0.45	high
5			low		0.23	low
6			no		0.23	low intermediate
7			medium		0.14	high intermediate
8			high		0.35	high
9				Jan- Feb	0.19	low
10				Spring	0.19	high intermediate
11				Autumn	0.19	high
12				January	0.19	low
13				Summer	0.19	low intermediate
14	medium	not high	not high	not Autumn	0.25	low intermediate
15	medium	not low	not low		0.25	high intermediate
16	not high	medium	not high	not Autumn	0.25	low intermediate
17	not low	medium	not low		0.25	high intermediate
18	low	low			0.5	low
19	low		low	Summer	0.5	low intermediate
20	medium		no		0.5	high intermediate
21		high		Autumn	0.5	high
22	low		low	Jan- Feb	0.75	low
23	low	medium	no		0.75	low intermediate
24	medium	medium		Autumn	0.75	high intermediate
25	high	high	high		0.85	high
26	low	low	low	January	1	low
27	medium	low	no	Summer	1	low intermediate
28	high	medium	medium	Spring	1	high intermediate
29	high	high	high	Autumn	1	high

30	low		low		0.6	low
31	high			Autumn	0.35	high

The rules are pretended to be divided into equal amounts related to the output types, the number of variables and the use of labels for each variable. The output values from each rule are based on the relation to the input variables as shown in the plots in chapter 2 from [1].

- **Rules 1 - 13**: gives an output based on one variable of each combination. Since only one variable is used, the weights are lower than those with more variables. Also, the weights are higher for the variables with larger importance. It is defined that in the case with no previously sold books, the output is low intermediate.

  According to Figure 2-7, the publishing period of Autumn represents the largest sale, Spring the second largest, Summer the third largest and Winter the lowest sale.

- **Rule 14 - 17:** Since some of the variables have three different states while the output has four, the medium of each variable is split into low intermediate and high intermediate. This is distinguished by the NOT operator. For instance, if author visibility is medium and the system knows that none of the other variables has their highest values, the output is lower intermediate. On the other hand, if none of the other variables has lower values, the output is higher intermediate instead.
- **Rule 18 -30:** The output labels are given an equal amount of inputs, with a combination of 2, 3 and 4 inputs. The weights are assigned according to the number of inputs used; two variables are defined as equal to 0.5, three variables to 0.75 and all variables to 1. In general, when all inputs are low, the output is low, while, when all outputs are high, the values are high. Furthermore, the lower intermediate is based on a combination of low and medium, while the higher intermediate output is based on a combination of medium and high inputs.
- **Rule 31:** Is used as an adjustment for a missing combination that gave a conflicting result. This way, no conflict results emerge.

4. TASK 4: IMPLEMENTATION IN MATLAB.

The implementation in Matlab follows the variables and rules as shown in chapters 2 and 3. In order to analyse the output, the scales follow the power of 10 instead of real values. As the outputs can't be displayed with logarithmic scales, the scales are transformed. For instance, a value of 3 in the plot corresponds to 10^3. To display the behaviour of the system, three 3d plots are displayed:

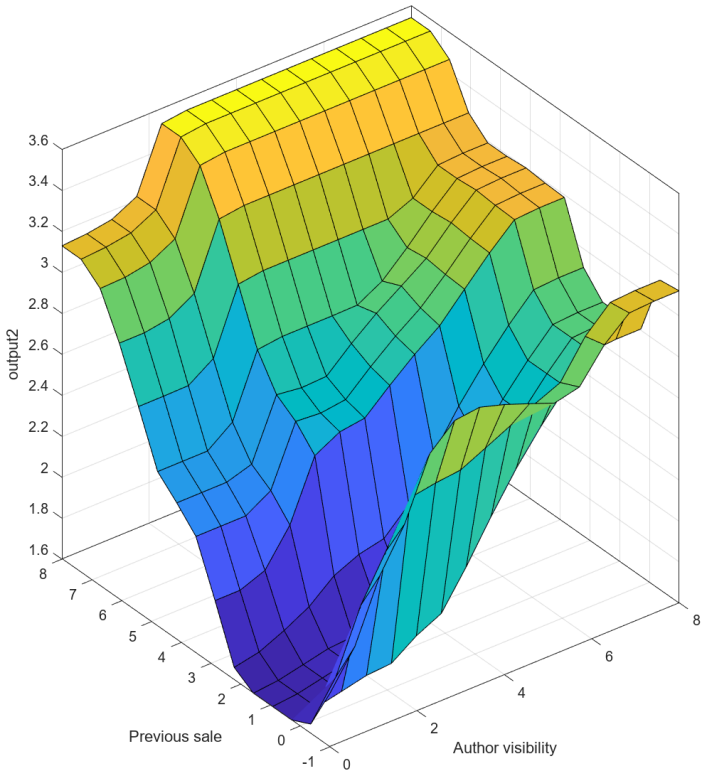


Figure 4-1: Output as a function of the Previous sale and Author visibility

As shown in figure 4-1, a combination of the high previous sale and author visibility gives the highest predicted sale. Additionally, the previous sale is more important than the author's visibility, reflecting figure 3-1.

Furthermore, a value of -1 $(10^{-1} \approx 0)$ represents zero previously sold books and indicates that the number of books sold can be higher for a new author entering the market than in the situation for an existing author that has proved a low book sale. This behaviour is expected from figure 2-5.

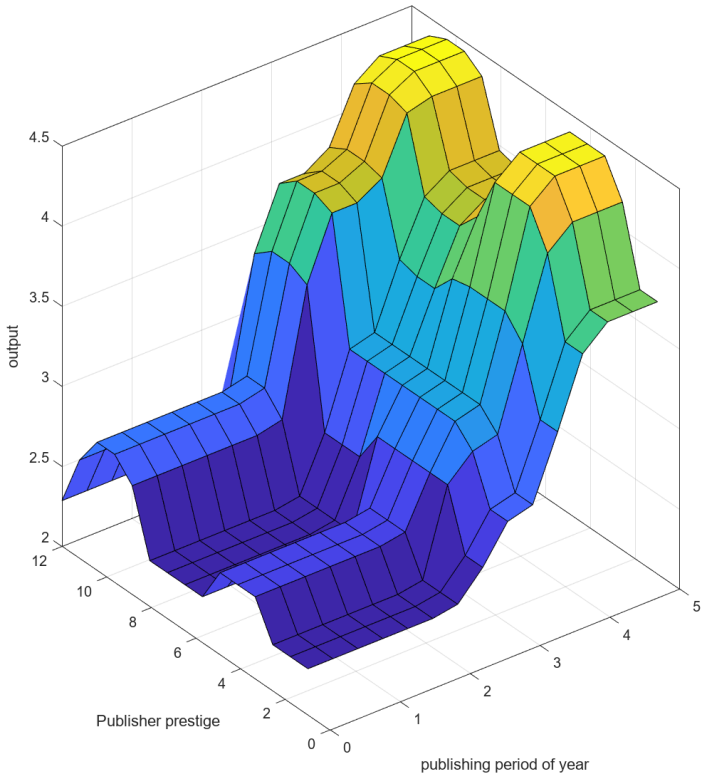


Figure 4-2: Output as a function of Publisher prestige and Publishing period of the year.

The axis labels are opposite, however, the values are correct.

Figure 4-2 shows that with a higher publisher prestige the predicted sale is larger. Also, the fluctuating sale throughout the year is following the sales numbers in figure 2-7. In this plot, it is displayed that the publisher's prestige is much more important for the sale than the period of the year as expected.

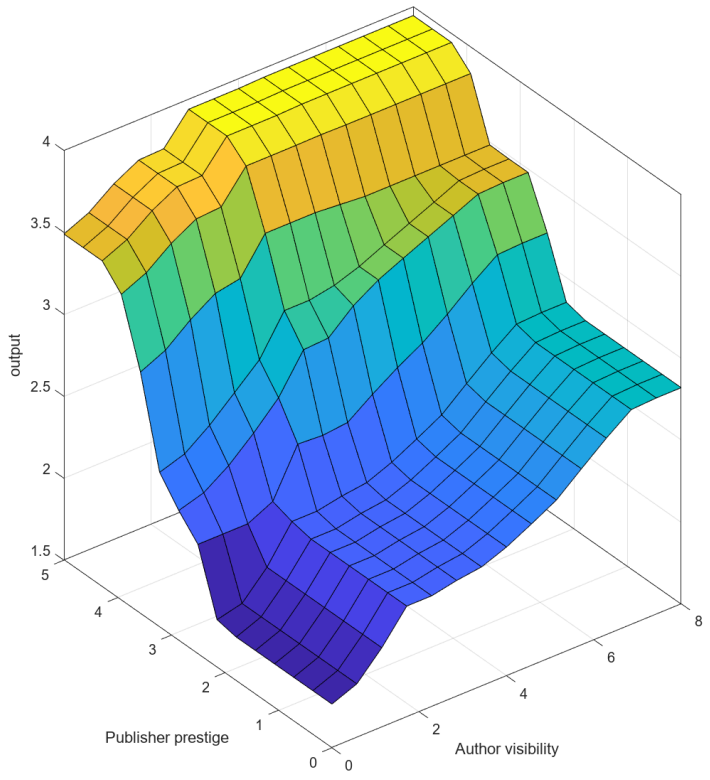


Figure 4-3: Output as a function of Publisher prestige and Author visibility.

Figure 4-3 shows that the output increases with a high publisher prestige and author visibility. The publisher's prestige is more important for the result than the visibility as mentioned in chapter 3.

In general, these 3D plots don't show inconsistencies in the data, but the differences between the levels are not equal and there are some straight parts. This should intuitively be more continuous.

5. TASK 5: TEST CASES.

Four different tests are performed to ensure the quality of the system. Table 5-1 explains the input parameters and output for each test. The cells contain the input/output numerical value and its linguistic representation.

Table 5-1: Test cases with input values and output values.

	Autor visibility	Publisher prestige	Previous sale	Publishing period	Output
1	10^7 High	1000 Medium - high	10^7 high	10 Autumn	30902 High intermediate
2	5 Low	10 Low	2 000 000 High	2 Jan - February	316 low intermediate
3	10^5 Medium-high	3000 Medium - high	0 No	3 Spring	2398 High intermediate
4	70 Medium - low	300 Medium - low	1000 Medium	4 Spring	257 Low intermediate

As the table shows, the outputs have a large correlation to the values of the inputs. Case 3 and 4 demonstrate that when author visibility and publisher prestige is medium but close to high and low, it changes the outputs accordingly. It is also observable that all the outputs in those examples are only low intermediate or high intermediate. In order to get other output values, it is required that all the inputs are high or low. This is probably due to an averaging effect of multiple rules applying.

Figure 5-1 to 5-4 shows the outputs of the cases from MATLAB. As described earlier, the input values in this system are the power of 10, meaning the input value in MATLAB = $log(x)$, where x is the value of the input.

Figure 5-1: Result of test case 1 from MATLAB. Rule 2, 6, 11, 17 and 31 is fired. The output is $10^{4.49} =$ 30902 books.

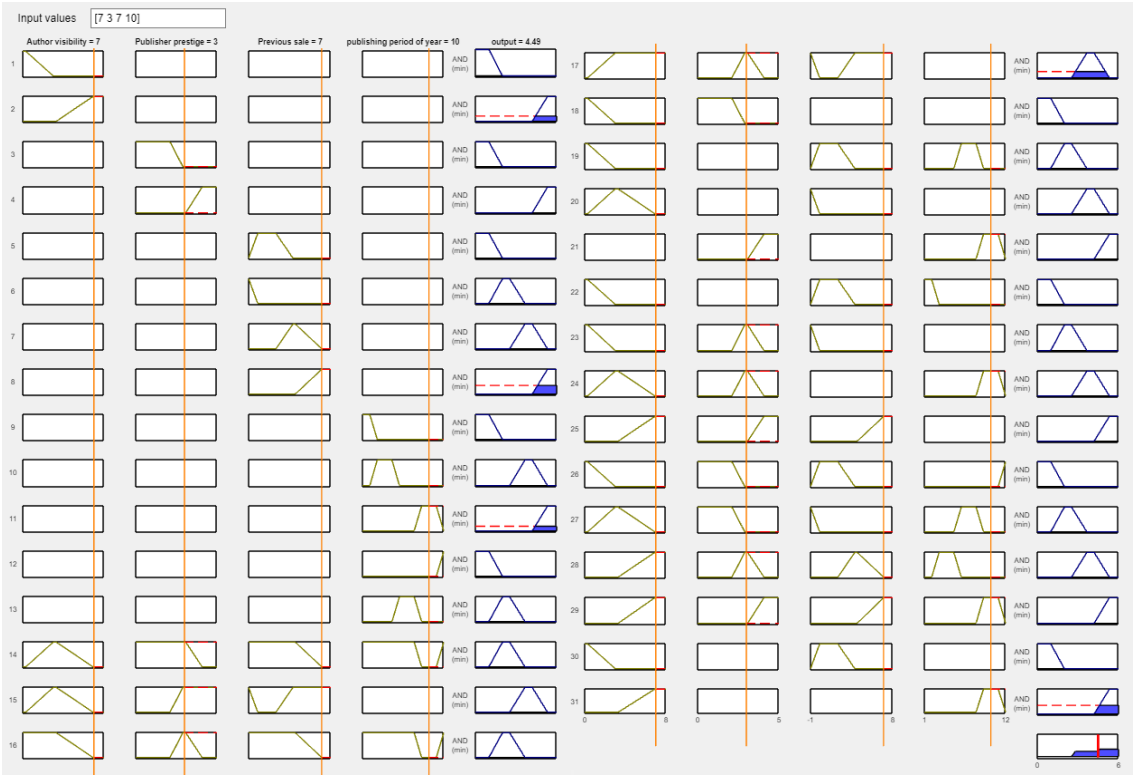


Figure 5-2: Result of test case 2 from MATLAB. Rule 1, 3, 8, 9 and 18 is fired. The output is $10^{2.5} =$ 316 books.

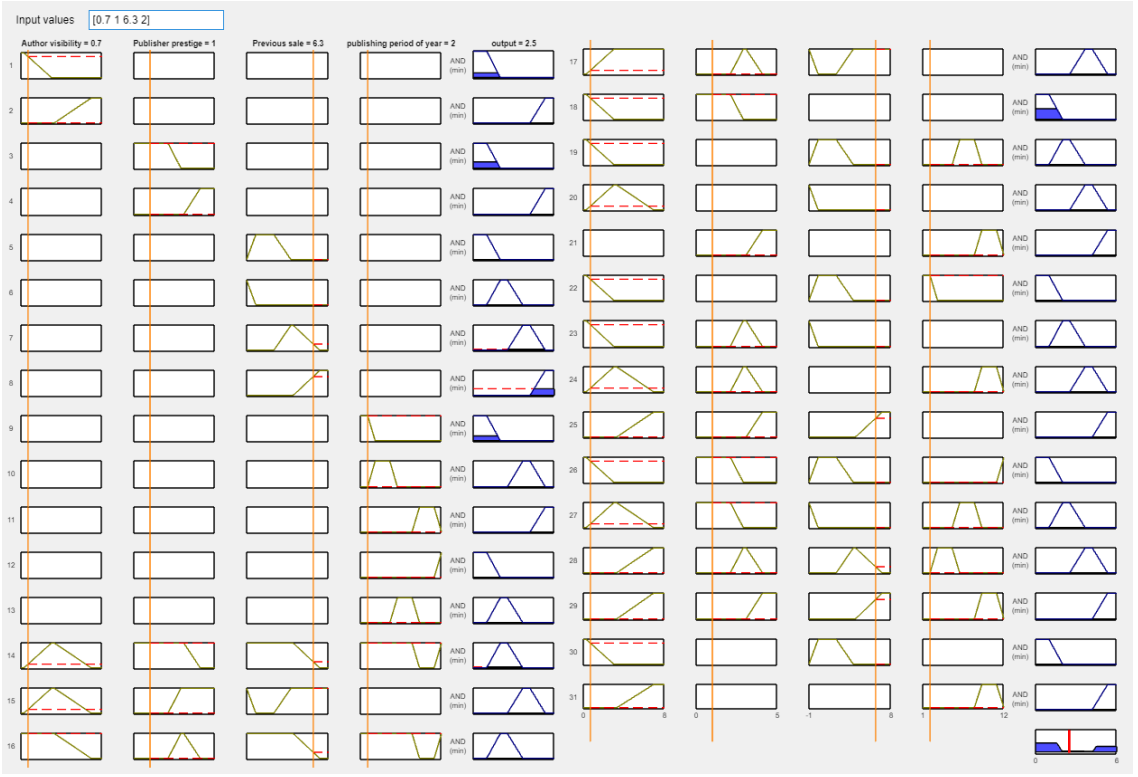


Figure 5-3: Result of test case 3 from MATLAB. Rule 2, 8, 10, 14, 15, 16, 17 and 20 is fired. The output is $10^{3.38}$ =2398 books.

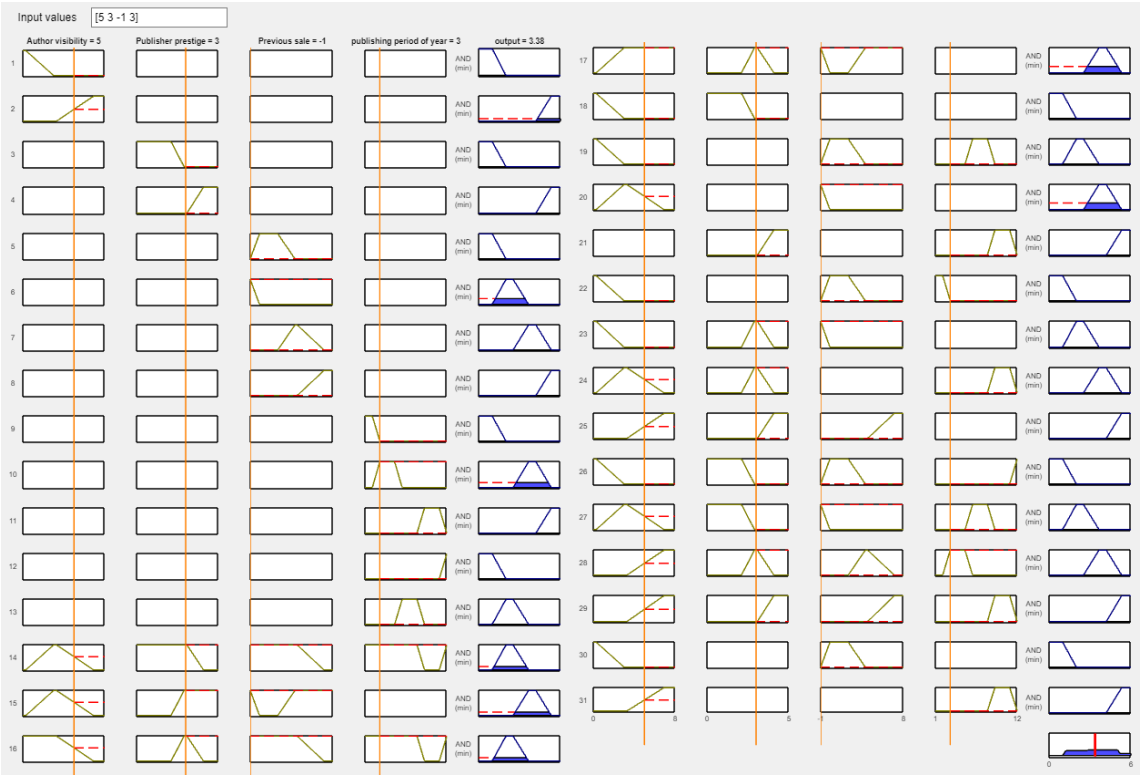
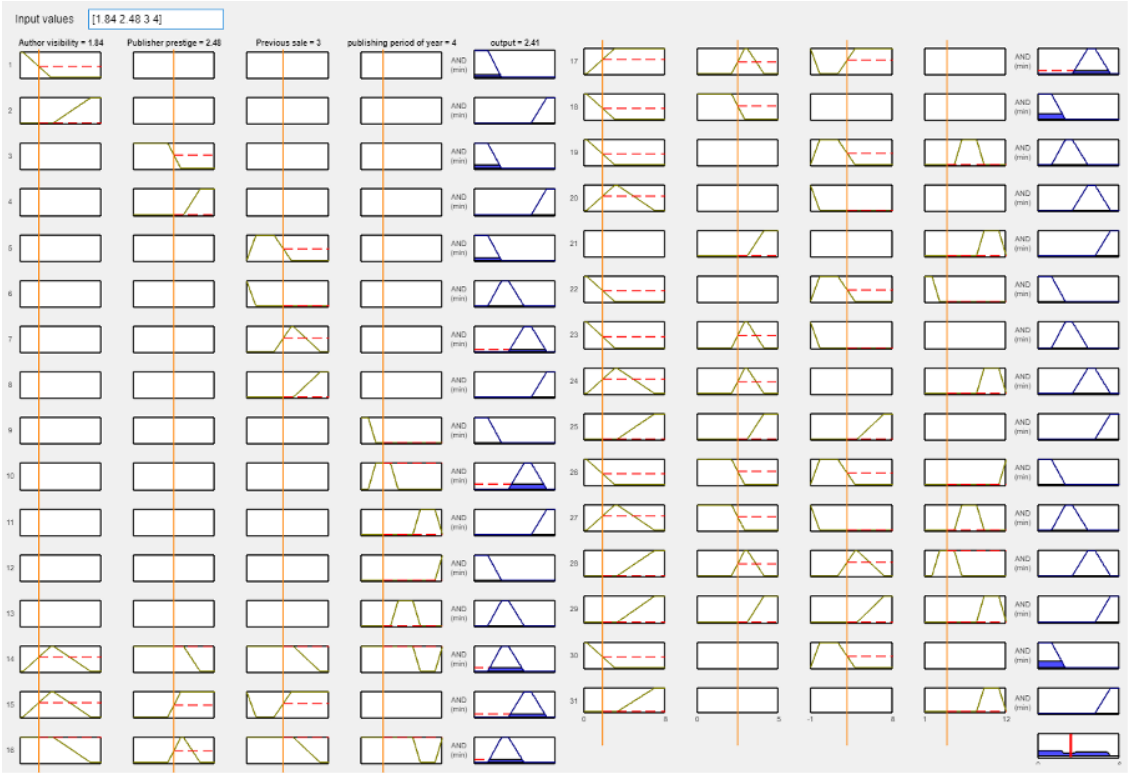


Figure 5-4: Result of test case 4 from MATLAB. Rule 1, 3, 5, 10, 14, 15, 16,17, 18 and 30 is fired. The output is $10^{2.41}$ = 257 books.



6. TASK 6: FINAL QUESTIONS

a. Looking at your fuzzy expert system, what is the influence of the publishing period on the number of books if the author has great visibility?

As shown in figure 6-1, when the author visibility is great, and also in the other cases, the output relies partially on the seasons of the publishing period. Most books will be sold with a release in Autumn, second most for spring, second lowest for summer and the lowest sale would be for books sold during the winter months.

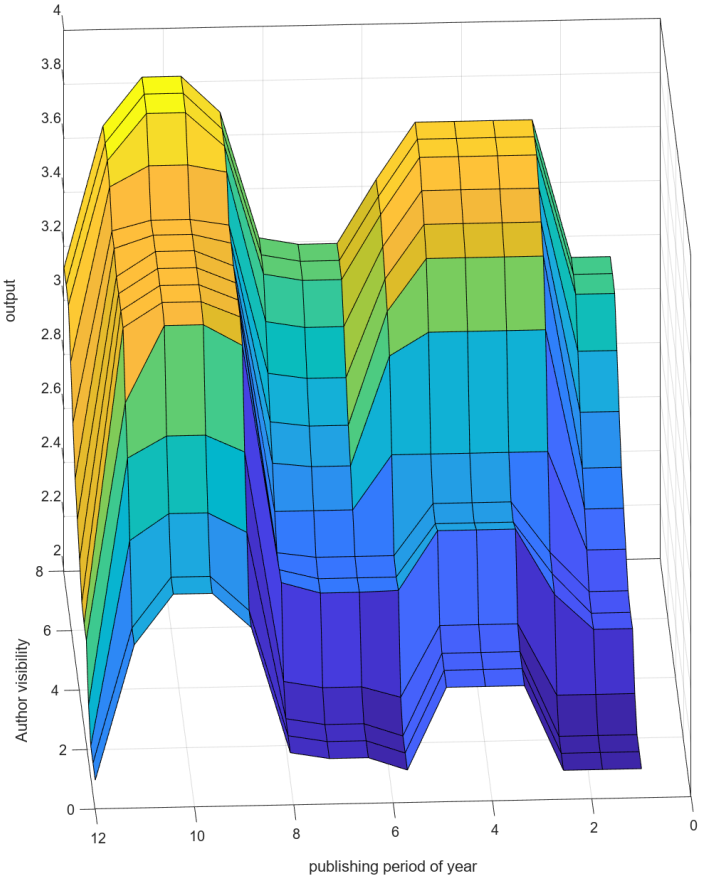


Figure 6-1: Output as a function of Author visibility and publishing period

b. What changes should you make to the system and which are the implications in the system if you want to also consider the age of the readers to whom the book is addressed?

If it is desired to predict the future sale based on the age of the addressed readers, it is necessary to add a new variable for the age. Books with recommended ages often operate with intervals, which can be used as labels for the function. These intervals should also be expanded to the older readers for the model.

In order to make rules for this function, it would be required to get data about the sale related to the ages. Preferably also combined with information about the other measures, so it is possible to make rules based on different situations. For instance, it might be the case that younger people buy books at different periods of the year than the average person.

Some implications introduced is that when more variables are added, it is more difficult to visualise and control the behaviour of the system. Besides that, the increment of age does not necessarily follow an increasing function, instead, it registers a fluctuating pattern, causing a more complex model.

c. If the author's visibility depends on many criteria, how can we model it in this expert system?

In order to distinguish the different criteria for the author visibility, it is possible to split the visibility function into multiple functions, each with fewer criteria. The problem with this approach is that the system increases its complexity.

If the effect of the visibility depends on criteria from the other variables within the system, it is possible to add more rules that take this into account.

Furthermore, If the author's visibility has large uncertainty because it depends on many variables, it is possible to decrease the weights in rules where the variables are used.

7. CONCLUSION

The fuzzy system to predict Book Sale based on the four input variables are constructed as described in the introduction.

The input variables are based on the scale and clustering from [1] while the output labels are based on the observations from [2].

The rules take into account the importance of the variables in [1] and the relation between the input value and the book sale, but it is important to denote that the papers don't describe the effect of combining different variables.

The model is analysed using 3d planes of pairs of variables and some experiments. The outputs show consistency for expected output and inputs in the final system.

It is experienced during the testing that when there are multiple input states with different characteristics and rule combinations, the system gets more complex and might give unexpected results for some combinations of values. It is important to do the validation of the model, to avoid those problems, and get a consistent model.

The result from the machine learning model in the papers shows that the result is spread and difficult to predict accurately. In a domain like this with complex underlying factors, it can be advantageous to use a fuzzy system with linguistic inputs and outputs.

8. BIBLIOGRAPHY

[1] Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad & Albert-László Barabási (2019). *Success in books: predicting book sales before publication.*

[2] Jessie C. Martín Sujo , Elisabet Golobardes i Ribé and Xavier Vilasís Cardona (2022). CAIT: A Predictive Tool for Supporting the Book Market Operation Using Social Networks.