Survey Data Processing for Modelling through Artificial Neural Networks (ANNs)

Joaquín Texeira-Quirós , Maria do Rosário Texeira Justino , António José Gonçalves , Marina Godinho Antunes , Pedro Ribeiro Mucharreira *

Posted Date: 15 May 2024

doi: 10.20944/preprints202405.1008.v1

Keywords: survey; data; processing; modelling; neural networks; ANN

 doi:10.20944/preprints202405.1008.v1

Article

Survey Data Processing for Modelling through Artificial Neural Networks (ANNs)

Joaquín Texeira-Quirós [1], Maria do Rosário Texeira Justino [2], António José Gonçalves [2], Marina Godinho Antunes [2] and Pedro Ribeiro Mucharreira [3,*]

[1] Autonomous University of Lisbon
[2] Lisbon Accounting and Business School
[3] University of Lisbon, ISCE-Instituto Superior de Lisboa e Vale do Tejo
* Correspondence: prmucharreira@ie.ulisboa.pt

Abstract: Nowadays, inquiries are one of the most important tasks to be executed in order to get valued information. When an inquiry is done to gather information about a human behavior there are some problems inherent to it. One of the main problems is how the information about many different persons can be processed to give good information about their environment. Most of the information gathered by these inquiries is used to analyze the behavior of the respondents. It can be used to estimate the success of a certain product on the market or just to validate the satisfaction of current costumers. The veracity of this information can be critical to model environments with some acceptable level of precision. Modelling environments through Artificial Neural Networks (ANN) is highly common because ANN's are excellent to model predictable environments using a set of data. ANN's are good in deal with sets of data with some noise, but they are fundamentally surjective mathematical functions and they aren't able to give different results for the same input. So if an ANN is trained using data where the same input has different images which can be the case of survey data, this can be a problem for the success of modelling the environment. Therefore is necessary to process the data in order to adjust and eliminate invalid data.

Keywords: survey; data; processing; modelling; neural networks; ANN

1. Introduction

With the dynamism inherent to some environments, the forecast modelling is a tool than can permit the prediction of the state of the environment after an application of an action. Hoel, P. (1966, p.45) defends that the advantage of the use of mathematical models is to permit the forecast of results. Gonçalves, A. (2016) demonstrated that is possible to model a predictive environment based on data survey about the application level of management strategies and their impact on financial results. Modelling an environment where the dynamism doesn't follow fixed rules, as are the commercial markets, it is necessary an observation of the environments to model and decide by a method that allows modelling with a degree of precision suitable for the purpose for which it is intended.

The modelling of real environments with high dynamism may depend on a large number of variables making it more difficult to achieve. To accurately model an environment it is necessary to set assumptions that define under what conditions that model is valid. That definition should allow an improvement in the accuracy of the model by restricting the universe in which it is valid.

This paper is organized as follow: In Section 2 is described the empirical research with the sample characterization, the research hypotheses and the research methodology used in data processing. At last, in Section 3, are referred the main conclusions and contributions of this research.

 doi:10.20944/preprints202405.1008.v1

2. Empirical Research

In this sense, this article intends to describe and justify the processing and filtering the data that will serve to model an environment. The environment that is intended to be modelled is the impact of the strategies, used by the management of small and medium Portuguese companies, in the respective financial results. The intended modelling method is based on artificial neural networks. Although good at modelling data with some noise, should increase their accuracy if training data are adequate.

The data to be filtered and processed were obtained through a questionnaire made to the managers of 449 small and medium Portuguese companies. The questionnaire consists of the evaluation, of the level of importance and application of high level strategies in the companies, made by managers.

The filtering of the data should allow the choice of the surveys and companies that allow a good modelling through the ANN's, creating assumptions and conditions of use of the models.

Freire, A. (2008, p.511) argues that problems should be identified in advance using forecasting techniques, preparing the organization for future eventualities. Controlling, influencing or acting on the sources of uncertainty can allow mitigation of the impacts caused. The increased flexibility and structural competitiveness reduce exposure to uncertainty.

Santos, A. (2008, p.23) points out that predictability is not the inverse of uncertainty, but rather the degree of "probabilistic certainty with which one can foresee certain events". There is an inverse relationship between predictability and uncertainty. The more dynamic and complex the surrounding environment, the more difficult is to predict events by the organization.

2.1. Research Hypothese

It should be remembered that the hypothesis must be in agreement with the objective and that at this point it is not intended to evaluate the predictions generated through ANNs, but to evaluate the possibility of modelling a given environment.

In this way the hypothesis to be considered in this article, is the evaluation of the model before and after the filtering and processing of the initial data. An important consideration to take into account is the fact that the processing of the data should not modify the knowledge induced by the initial data.

The hypothesis to be studied is to identify a significant improvement in the final model after the initial data are filtered and processed, comparing to the model obtained from the initial data.

The assumptions to be followed in the new model must be identified, such as a constraint in the universe applicable to the model.

In this way, the hypothesis to be verified is defined as the significant improvement in the modelling through processed and filtered data in comparison with the original data obtained through a questionnaire considering a significant degree of subjectivity.

2.2. Research Description and Sample Characterization

The questionnaire is one of the most important pieces in the intended modelling since it is from the data collected through it that will be possible to identify the impact the strategies have on the results.

Table 1 shows the possible values as well as the area of each question in the questionnaire.

Table 1. Numerical representation of the strategies in the questionnaire.

Questions	Strategy	Interval	Observations
Q2	Price Increase /Reduction	[-9,9] ϵ Z	Reduction: Negative Increase: Positive
Q3	Quality Increase /Reduction	[-9,9] ϵ Z	Reduction: Negative Increase: Positive
Q4	Reduced Personnel Cost	[0,9] ϵ Z	

 doi:10.20944/preprints202405.1008.v1

Q5	Investment	[0,9] ϵ Z	
Q6	Decrease Financing	[0,9] ϵ Z	
Q7	Product Diversification/Specialization	[-9,9] ϵ Z	Specialization: Negative Diversification: Positive
Q8	Reduction / Increase of Customers or Markets	[-9,9] ϵ Z	Increase: Negative Reduction: Positive
Q9	Business Synergies	[0,9] ϵ Z	
Q10	Product Disclosure	[0,9] ϵ Z	
Q11	Business Reorganization	[0,9] ϵ Z	
Q12	Renegotiation with Suppliers	[0,9] ϵ Z	

A possible representation of the answers to the questionnaire by the managers could be:

Table 2. Example of business data representation.

	Inc. 1	Inc. 2	Inc. 3	Inc. 4	Inc. 5	Inc. 6	Inc. 7	Inc. 8	Inc. 9
q2a2013	4	0	-4	4	-4	0	0	1	0
q2a2014	5	0	-5	0	-4	0	-5	1	0
q3a2013	7	0	4	8	4	0	0	6	4
q3a2014	7	0	5	8	4	0	0	7	4
q4a2013	0	0	4	0	2	0	0	0	2
q4a2014	0	2	5	0	3	0	5	0	2
q5a2013	0	0	3	5	2	0	0	4	0
q5a2014	0	0	6	3	2	0	0	6	0
q6a2013	0	0	0	9	0	0	0	2	0
q6a2014	0	0	0	9	0	0	5	1	0
q7a2013	-7	0	-4	-4	0	0	9	-8	0
q7a2014	-7	0	-5	-5	0	0	9	-8	0
q8a2013	-4	0	3	-5	-3	0	0	-5	0
q8a2014	-5	0	4	-5	-3	0	0	-7	0
q9a2013	5	0	0	4	0	0	5	3	0
q9a2014	5	0	0	3	0	0	5	4	0
q10a2013	2	0	5	5	6	0	5	0	1
q10a2014	2	0	4	5	6	0	7	8	1
q11a2013	1	0	3	0	4	0	0	3	0
q11a2014	1	1	4	0	4	0	5	3	0
q12a2013	2	0	0	5	0	0	0	8	0
q12a2014	2	0	0	5	0	0	5	8	0
Absolute Sum 2013	32	0	30	49	25	0	19	40	7
Absolute Sum 2014	34	3	38	43	26	0	46	53	7
Total Sum	66	3	68	92	51	0	65	93	14

The values identify the importance and the level of application of each strategy: a strategy with an application of 8 means that it has a significantly more importance than a strategy with an application of 4.

2.3. Research Methodology

The methodology followed to demonstrate the hypothesis is oriented to an empirical study carried out in phases. At the beginning, results of the modelling done through the initial data without any type of processing or filtering should be presented.

As each phase is covered, both processing and filtering of the data, the results of the modelling through neural networks will be presented.

 doi:10.20944/preprints202405.1008.v1

The generated neural networks although similar, may have some different characteristics. In order to identify the success of a model, experiments were made with several topologies and with several training methodologies of the neural networks. In this context, it was chosen the ANNs with the best results.

2.4. Analysis and Discussion Oh the Results

2.4.1. Initial Phase

In order to evaluate the possibility of modelling, through the data generated by the survey and the financial data obtained, it was tried to model the environment without any processing or filtering the original data.

The training methods of the neural networks used were: backpropagation, resilient backpropagation + and resilient backpropagation. For each one of the algorithms, several ANN topologies were tested.

Although it isn't the objective of this article an explanation of how the neural networks work, to understand the graphics, it is necessary to understand a fundamental concept. The modelling of a given environment, through neural networks, must be done by two sets, the training set and the set of tests. Each of these sets must be obtained randomly from a set of observations describing the environment to be modelled. The training set, as the name implies, serves to train the ANN to be modelled and the test set is used to evaluate the quality of the ANN to model the intended environment and whether if this environment is predictable.

In this study it will be presented three graphics for each one of the tests. The first one (optimal) demonstrates what would be ideal. It presents the total set and how the graphic should be if it models the environment perfectly. The second (predicted test set) demonstrates the modelling of the test set. For this article this graphic doesn't, necessarily have to be similar to the graphic designated by optimal. The third (predicted train set) demonstrates the modelling capability for the training set. This graphic allows evaluating the possibility of modelling the desired environment for the data without forecast. The more similar to the graph designated by optimal, the better is the modelling of the training set. It should be noted that this evaluation can be done only by looking at the diagonal line of the graphic: the closer and more points are on this line the higher the quality of the modelling done.

The subtitle of the graphics identifies what method and topology of the neural network were used to generate the modelling. Example, "Total_ABS_BP_VN_12-8-2" means that all data was used, the training algorithm was backpropagation and the topology (hidden layers) of the neural network was 12-8-2. The financial data to be modelled was the revenue (VN).

Revenue – Backpropagation

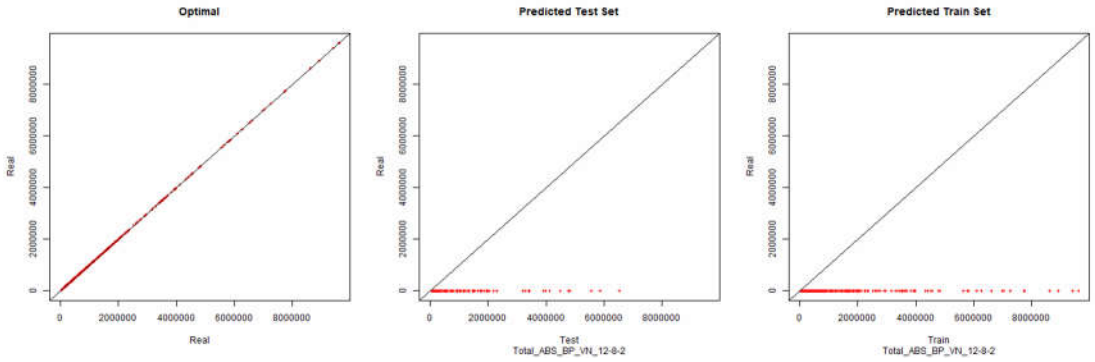


Figure 1. Initial Modelling Charts Revenue (Backpropagation) ANN: 12-8-2.

Economic Performance – Backpropagation

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 May 2024 doi:10.20944/preprints202405.1008.v1

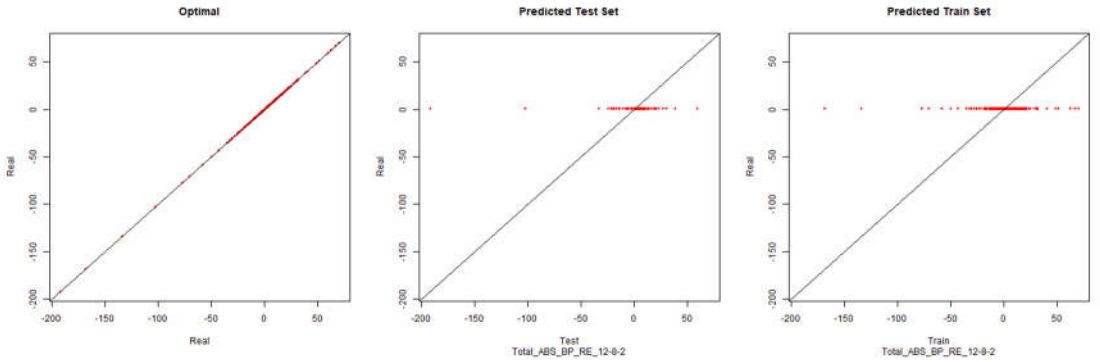


Figure 2. Initial Modelling Charts Economic Performance (Backpropagation) ANN: 12-8-2.

EBIT – Backpropagation

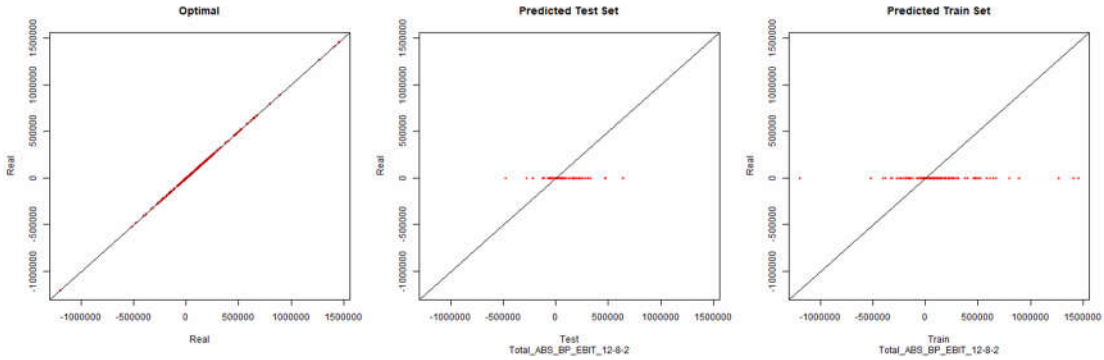


Figure 3. Initial Modelling Charts EBIT (Backpropagation) ANN: 12-8-2.

For the positive or negative resilient backpropagation algorithms no modelling was possible because the neural network training did not converge. Therefore the calculation of the values of the artificial neural network was not possible.

As it can be seen in the presented graphics, it was not possible to model the desired environment through neural networks successfully. Although it is presented here the results for the topology of hidden layers of the neuronal network 12-8-2, several topologies were tested with results similar to the presented topology.

2.4.2.1. st Phase - Strategic Behavior

Analyzing the distribution of the absolute sum of the values answered by the respondents the following graphs were obtained:

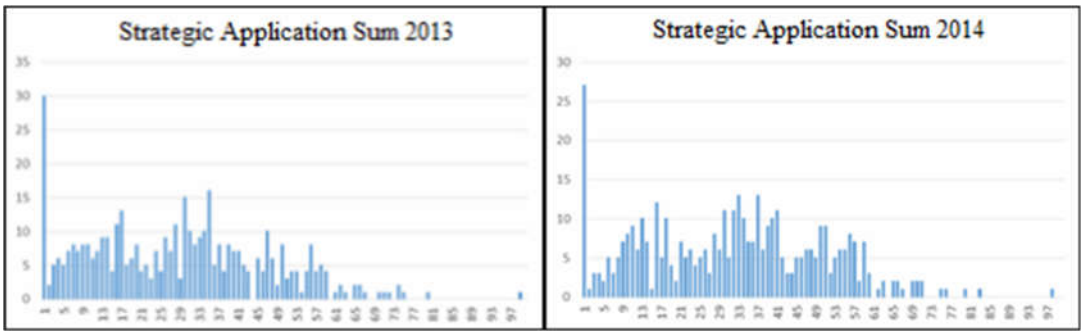


Figure 4. Absolute sum of the application of strategies for the years 2013 and 2014.

 doi:10.20944/preprints202405.1008.v1

It can be seen in this graphics that the responses follow a fairly similar distribution over the two years. The absolute sum is represented on the horizontal axis and the number of companies, with that sum, is represented on the vertical axis. The graph shows that there were 30 organizations that did not implement any of the strategies considered in 2013, and 27 organizations that also did not implement in 2014.

Analyzing the total absolute sums for the two years (2013 and 2014), we obtained the following graph:

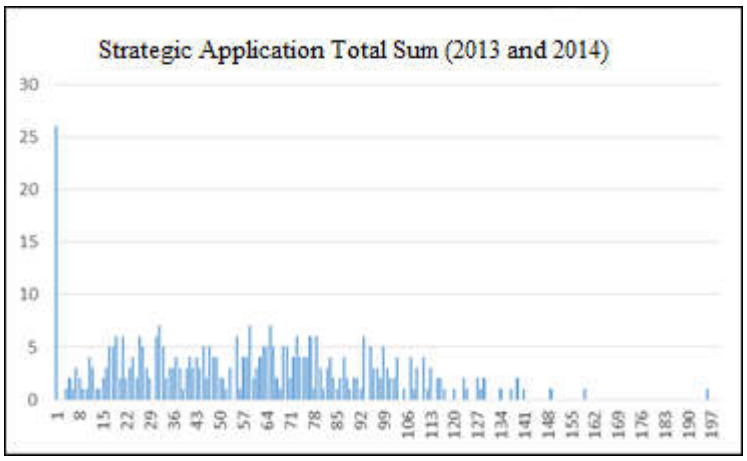


Figure 5. Total sum of the application of strategies in the years 2013 and 2014.

There are 3 points to consider:

(1) There were 26 organizations that did not implement any strategy considered in the years of 2013 and 2014.
(2) It can be seen that organizations have a strategic behavior similar to a normal distribution. For the purposes of this study and since there were found no studies related to this behavior, it will be considered this assumption.
(3) Behaviors that do not fit in most of the organizations can create unwanted noise in the analyses inherent to the study.

The problem, described in section 3, should be minimized by eliminating the data from organizations that do not fit within the distribution of strategic behavior of most of the organizations.

Moore, D. (2003, p58) reports that an appropriate density curve is often adequate to describe a standard behavior of the distribution, although a real data set is usually not possible to describe accurately through a function distribution.

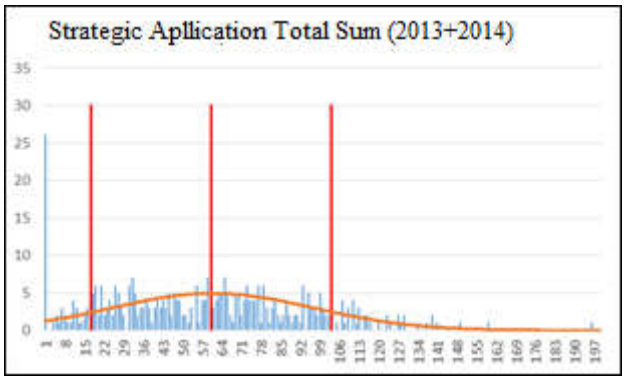


Figure 6. Projected normal distribution in the total absolute sum (2013 and 2014).

From the graphic it can be deduced that the expected behavior of the managers obeys a high degree of similarity to a normal distribution.

 doi:10.20944/preprints202405.1008.v1

Soong, T. (2004, p.200) shows that by the central boundary theorem the normal distribution is often adequate to describe random events. Although the target samples of the study are not random, for filtering the data and by similarity of the distribution of the samples with a distribution, is believed to be adequate for this purpose.

Montgomery, D. and Runger, G. (2002, p.109) argue that the normal distribution is the most commonly used to describe random events, but that there are events that, although not random, can be considered as described by a normal distribution.

The calculated mean of the total absolute sum of respondents' responses is 59,685 and the standard deviation is 36,094. The coefficient of variation is 60.47%. Dancey, C. and Reidy, J. (2017, p.76) define the standard deviation as the measure of the degree to which the sample deviates from the mean. Hoel, P. (1966, p. 101) reports that the normal distribution is entirely defined by its mean and standard deviation, so there is no specific need for further calculations to define the intended distribution. In order to calculate the lower and upper limits, it was decided to use a value greater than the standard deviation of 20% (43,133). Thus the lower limit is 59,685-43,133 = 16,372 and the upper limit is 59,685 + 43,313 = 102,998.

This means that all surveys whose behavior is within normal parameters were accepted at this stage. The normality considered for this purpose was the surveys whose absolute sum of responses in the two years is between 16 and 103.

All data from organizations whose total absolute sum does not belong to the interval [17,102] were eliminated for subsequent analyses. At this stage, data from 110 organizations were eliminated from a total of 449, and for this reason data of 339 organizations were considered for further analysis. In this case the acceptance of the set of samples was of 75.5%.

The process described here can be compared in a very simple way to a questionnaire to assess the market potential of a new ice cream flavor. Let's imagine a survey about various flavors of ice cream, for example, strawberry, banana, chocolate, cream, vanilla, lemon and the new flavor, which we will call flavor A. In the sample responses we have respondents who answered all 1 and others who answered everything 9. It can then be assumed that respondents who answers all 1 do not like ice cream. On the other hand the respondents who answered 9 like ice cream a lot. However does not bring more information about the acceptance of new flavor by potential consumers. This behavior is not typical of a "normal" consumer and therefore these responses should not be taken into account for the purpose of modelling the typical consumer behavior in relation to the new flavor.

A manager, who has answered all 9, should not have his answers taken into consideration. When it comes to small and medium enterprises, the costs of applying strategies can have a significant impact on the final financial results.

After this data filter, it was attempted to model the environment with the same type of ANN that was used to attempt the modelling with the initial data.

Revenue – Backpropagation

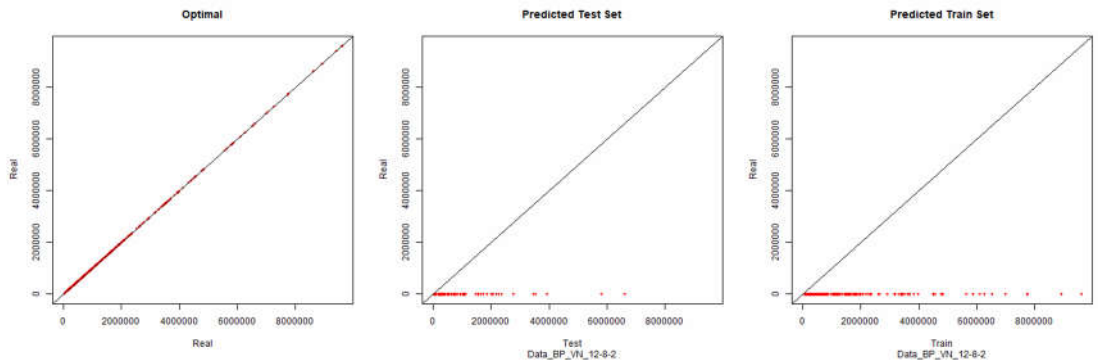


Figure 7. Modelling Charts Phase I Revenue Backpropagation ANN: 12-8-2.

Economic Performance – Backpropagation

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 May 2024 doi:10.20944/preprints202405.1008.v1

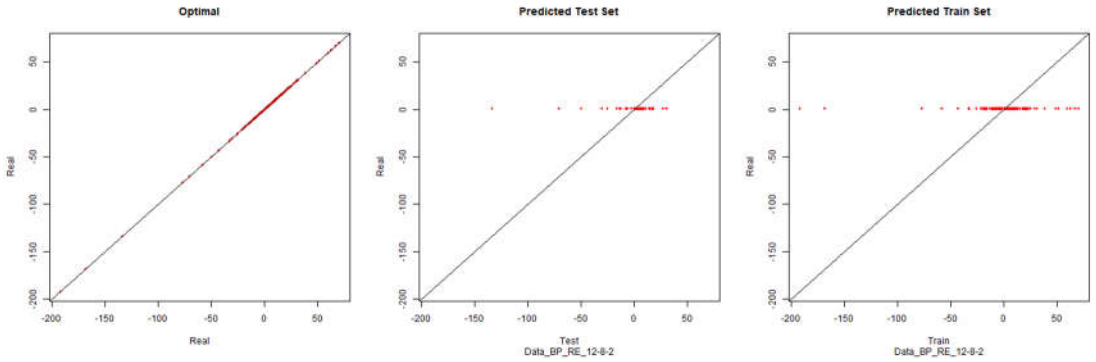


Figure 8. Modelling Charts Phase I Economic Performance Backpropagation ANN: 12-8-2.

EBIT – Backpropagation

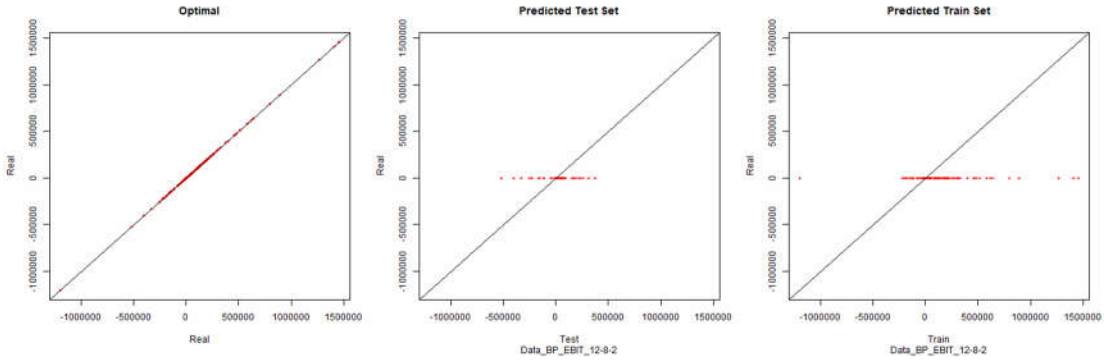


Figure 9. Modelling Charts Phase I EBIT Backpropagation ANN: 12-8-2.

At this stage, as can be seen from the graphs, there was no visible improvement in modelling through artificial neural networks. Despite filtering data excluding questionnaires that might raise suspicions about the quality of their responses, modelling is still not possible.

2.4.3. 2nd Phase - Data Processing

One of the main problems that were found in this type of survey is that the answers may be subjective regarding the comparison between different people. This means that the answers given by the same person are expected to have a significant coherence. For example if a manager gave a response of 5 to Q3 (quality) and 8 to Q6 (funding), it can be assumed that the strategy of reducing financing had a higher priority and importance than the strategy of increasing product quality.

However when comparing the managers responses referenced in the previous paragraph, if one gave a response of 4 to question Q3 and a response of 7 to question Q6, there is no possibility of making a direct comparison of the importance the strategies for both managers, since there is no precise notion of what the values mean to each. However, trusting the answers, it can be assumed that the isolated responses of each manager have important information about the strategies applied to the respective company.

Similarly, considering the example of the inquiry on a new flavor (flavorA) of ice cream if one respondent (A) answers for example 7 to chocolate, 3 to vanilla and 6 to flavorA, then it can be assumed that the coherence of the answers leads us to conclude that the respondent likes chocolate more than flavorA, but likes more flavorA than vanilla. However if another respondent (B) responds 8 to chocolate, 4 to vanilla and 7 to flavorA, it cannot be inferred that the respondent (B) likes more chocolate than the respondent (A) because there is no direct relationship between the meaning of the values for each of the respondents.

 doi:10.20944/preprints202405.1008.v1

In order to emulate a direct relationship between the values of each of the respondents it is necessary to process the results of the surveys. Data processing is based on normalization. It does not give us an exact notion of what each of the values represents between each respondent but approximates the relation of the values of each of the respondents. In the ice cream taste survey this would mean that someone who answered everything 4 compared to someone who answered everything 5 the ratio of the respondents between the tastes should be the same.

As it is not objectively possible to assess the significance of the levels of application for each of the managers responding the survey, the way to overcome this obstacle was to normalize the results of the survey. The assumption underlying this normalization is that responses represent the priorities each organization has when implementing its strategies. The normalization consists in making that for each one of the sums of the survey to each respondent to be equal, in this case, to 1(one). And process the data so that the relative information between the choices made by the respondents is not lost.

Thus an organization that responds that has an application level of 5 to 4 strategies and 0 to all others has an application priority of 0.25 for each of the strategies. Similarly, an organization that responded 3 to a level of application to two strategies and 6 to a level of application to other two strategies, it has actually an application priority of 0.166 to two strategies and of 0.333 to the other two strategies. In percent, the values represent the relative importance of each of the strategies, as well as the level of application of each strategy in the organization.

The normalization allows the relation of the data answered individually to each of the questionnaires. In this way, one can more accurately understand the values obtained in each of the questionnaires and use them to relate the level of application of the strategies to the results.

To achieve a normalization of the data, each strategy is divided by the absolute sum of the year in question. Thus, for example:

Table 3. Sample questionnaire values.

q2 2014	q3 2014	q4 2014	q5 2014	q6 2014	q7 2014	q8 2014	q9 2014	q10 2014	q11 2014	q12 2014	Absolute Sum
-1	-1	4	7	2	-5	-6	0	6	0	4	36
0	5	3	7	0	-3	-5	0	6	4	6	39
5	9	0	6	6	0	0	4	5	4	7	46
-6	0	8	0	0	0	-8	0	5	0	0	27
-6	0	0	5	0	5	0	0	6	0	7	29

Normalizing the table of values of the questionnaire for the respective companies would have:

Table 4. Example standardized questionnaire values.

q2 2014	q3 2014	q4 2014	q5 2014	q6 2014	q7 2014	q8 2014	q9 2014	q10 2014	q11 2014	q12 2014	Absolute Sum
-0,028	-0,028	0,111	0,194	0,056	-0,139	-0,167	0	0,167	0	0,111	1
0	0,128	0,077	0,179	0	-0,077	-0,128	0	0,154	0,103	0,154	1
0,109	0,196	0	0,13	0,13	0	0	0,087	0,109	0,087	0,152	1
-0,222	0	0,296	0	0	0	-0,296	0	0,185	0	0	1
-0,207	0	0	0,172	0	0,172	0	0	0,207	0	0,241	1

These normalized values will be the values to be used for the inputs and outputs of the neural network, together with the actual values of the results for each organization.

Revenue- Backpropagation

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 May 2024 doi:10.20944/preprints202405.1008.v1

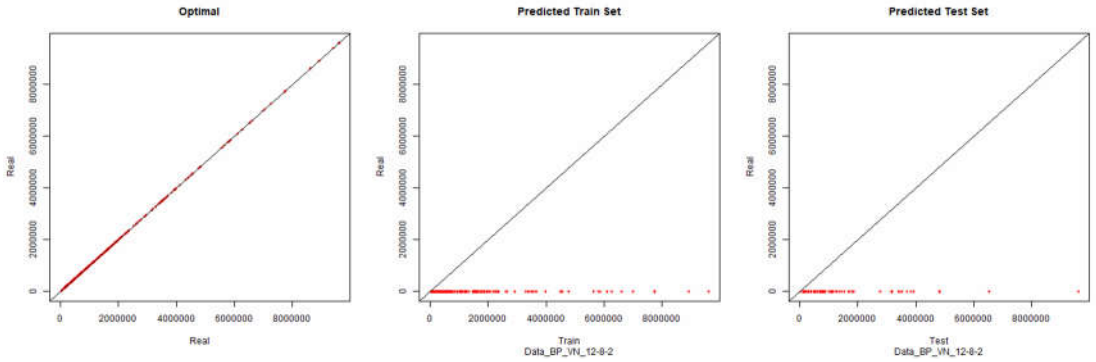


Figure 10. Modelling Charts Phase II Revenue (Backpropagation) ANN: 12-8-2.

Economic Performance – Backpropagation

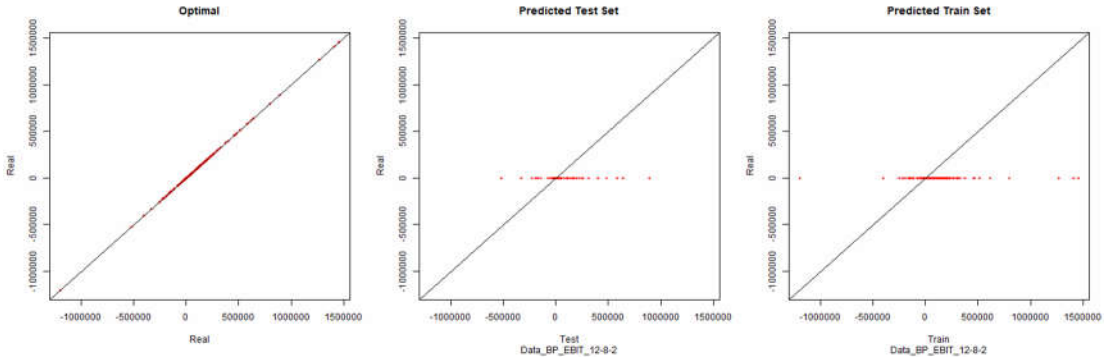


Figure 11. Modelling Charts Phase II Economic Performance (Backpropagation) ANN:12-8-2.

EBIT – Backpropagation

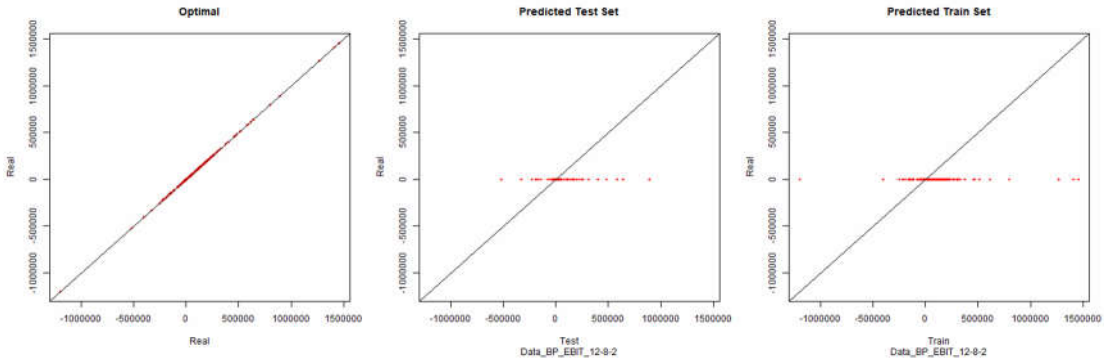


Figure 12. Modelling Charts Phase II EBIT (Backpropagation) ANN:12-8-2.

Once more, if we look at the graphics we can conclude that the processing and normalization of the questionnaire data did not have a significant impact on the modelling of the respective environment.

2.4.4. 3rd Phase - Processing and Analysis of Organizational Results

Data from the SABI database (http://sabi.bdvinfo.com) were collected for the year 2012, 2013 and 2014 for the organizations being analyzed. The information considered relevant for the study was:

(1) Revenue (VN)
(2) Operating Income (RO)

 doi:10.20944/preprints202405.1008.v1

(3) Net Results (RL)
(4) Economic Performance (RE)
(5) Financial Performance (RF)
(6) EBIT
(7) EBITDA
(8) Solvency Ratio (RSol)

An analysis of each of these points for the organizations referenced in the study allowed the perception of some problems that could arise in modelling the strategic vs. results.

According to Porter, M. (1996) the concept of strategy is inherent in the need to create an advantageous position against its competitors, through a set of actions. Porter also points out that the essence of strategy is mainly to choose different paths from its competitors.

The diversity in the data induced some evaluation criteria regarding the modelling attempt. It allowed one to induce that some of these results should have nothing to do with the strategies used and that the introduction of these organizations in later phases of the study could have significant and negative impact in the attempt of strategic modelling.

For example, if an organization increased its revenue by 300% in 2014 compared to 2013 or if a company went from significantly positive net results in 2013 to negative net results in 2014, one would normally not be able to associate these changes with a strategy, but with an extraordinary factor. Therefore, in the same way that companies that did not fit the desired profile regarding the application of strategies were eliminated in the first phase, it was opted to eliminate data from organizations that did not fit into a financial stability profile in the years 2013 and 2014.

According to Hitt, Irreland and Hoskisson (2011, p.6) strategic competitiveness can be achieved by formulating and implementing a strategy that creates value. The strategy should be used to gain a competitive advantage that allows exploring the core competencies of the organization, through a set of commitments and actions previously outlined. Not all results achieved may be inherent to the strategies applied or may depend on strategies whose impact can be delayed.

Some operational concepts were defined as rules for acceptance / deletion of the data referring to the organizations for the study, and "delta" (δ) is defined as the difference between the result of 2014 and the result of 2013 to be divided by the result of 2013, ie δ = (I2014-I2013) / I2013 * 100 (in%):

(1) Revenue:
    a. The δ should not be less than -10% or greater than 30%.
(2) Operating results:
    a. Operating results for 2013 and 2014 should be positive.
    b. The δ should not be less than -20% or greater than 50%.
(3) Net Income:
    a. The net result for 2013 and 2014 should be positive.
    b. The δ should not be less than -40% or greater than 600%.
(4) Economic Performance:
    a. Economic Performance should be positive in 2013 and 2014.
    b. The δ should not be less than -10% or greater than 50%.
(5) Financial Performance:
    a. Financial Performance should be positive in 2013 and 2014.
    b. The δ should not be less than -50% or greater 200%.
(6) EBIT:
    a. EBIT should be positive in 2013 and 2014.
    b. The δ should not be less than -10% or greater than 50%.
(7) EBITDA:
    a. EBITDA should be positive in 2013 and 2014.
    b. The δ should not be less than -10% or greater than 50%.

 doi:10.20944/preprints202405.1008.v1

(8) Solvency Ratio:

a. The solvency ratio should be positive in 2013 and 2014.

It should be noted that the above filters are useful only to filter companies whose strategies may not have the impact generated on the results. This means that there may be eliminated companies in which the strategies used were actually responsible for the impacts of the results or that companies whose strategies did not have a significant impact on their results were accepted. The creation of these assumptions induces a reduction in the probability of being accepted in the modelling, but it does not absolutely prevent it to happen.

In order to assess the eligibility of strategic data of organizations, a point system was created, where each infraction described above is equivalent to 1 penalty point. In this way we can use the data of the organizations in different parts of the modelling, both in the training of the artificial neural network and in the evaluation of the performance of the same one:

1) 0 (Zero) or 1(One) penalty points: The organization's responses to the questionnaire, as well as its results, will serve to model the environment, since the responses/results are those that should induce less noise in the model.
2) More than 1 penalty point: The organization's responses, as well as its results, will be eliminated from modelling.

This method excludes data from organizations that may have been subject to non-strategic situations and may have had an impact on results. In this way, companies whose results may not be directly related to the strategies applied are excluded. After analyzing the data of the 339 organizations we have the following table:

Table 5. Penalty points vs. Number of organizations accepted.

Points	0	1
#Organizations	26	22

It will be 36 data sets of organizations that will serve to model the environment and 12 sets of data that can be used to evaluate the performance of the model.

There is no problem in excluding samples that we consider invalid according to the assumptions that were introduced. However, one should be aware if the final number of samples is sufficient to model the desired environment. In this case we consider that the sample size is sufficient and we still have the margin to provide some samples that allow the evaluation of the model.

Revenue – Backpropagation

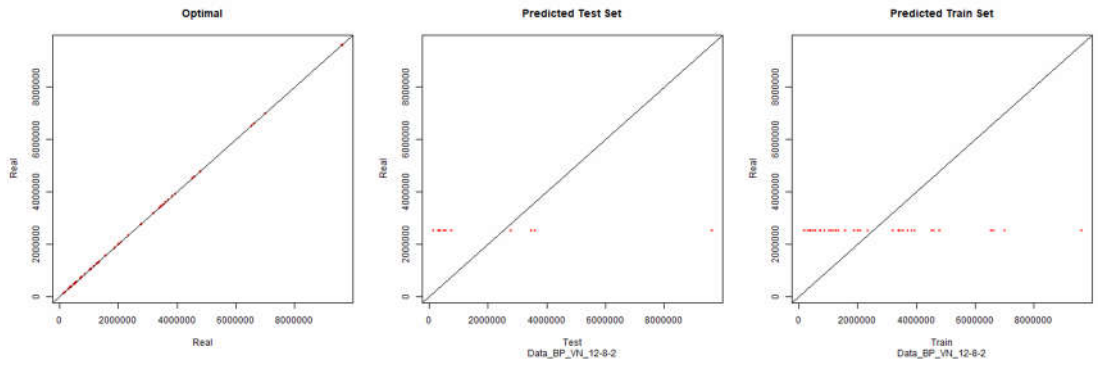


Figure 13. Modelling Charts Phase III Revenue (Backpropagation) ANN: 12-8-2.

Economic Performance – Backpropagation

 doi:10.20944/preprints202405.1008.v1

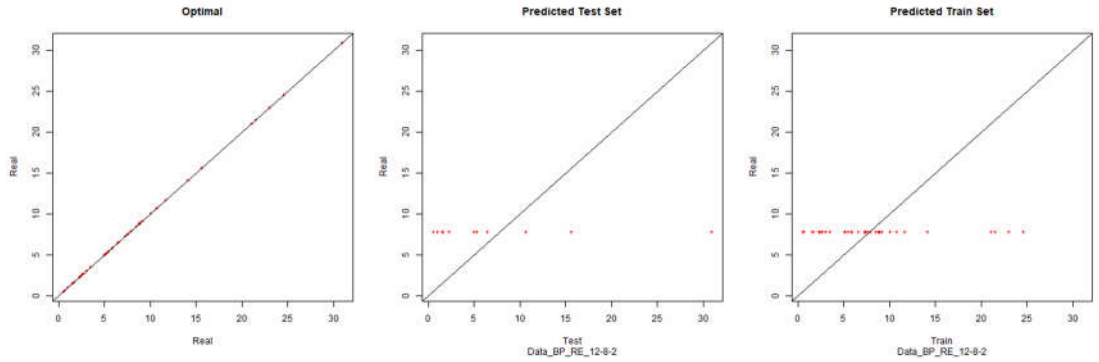


Figure 14. Modelling Charts Phase III Economic Performance (Backpropagation) ANN: 12-8-2.

EBIT – Retropropagação

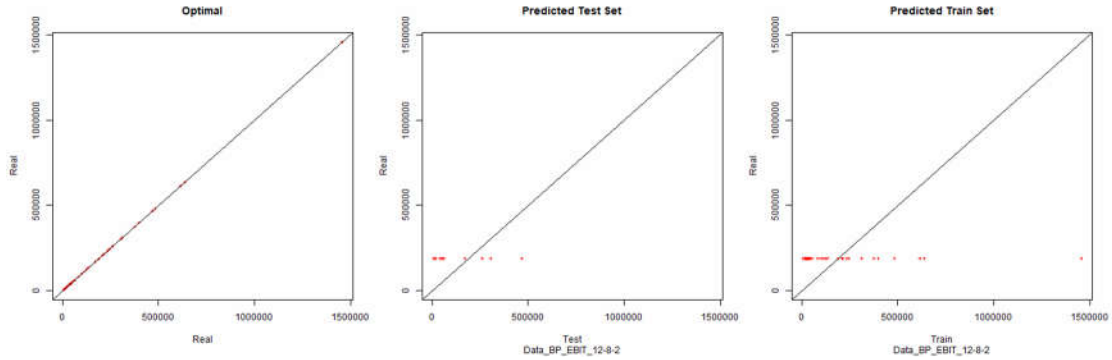


Figure 15. Modelling Charts Phase III EBIT (Backpropagation) ANN: 12-8-2.

Economic Performance–Resilient Positive

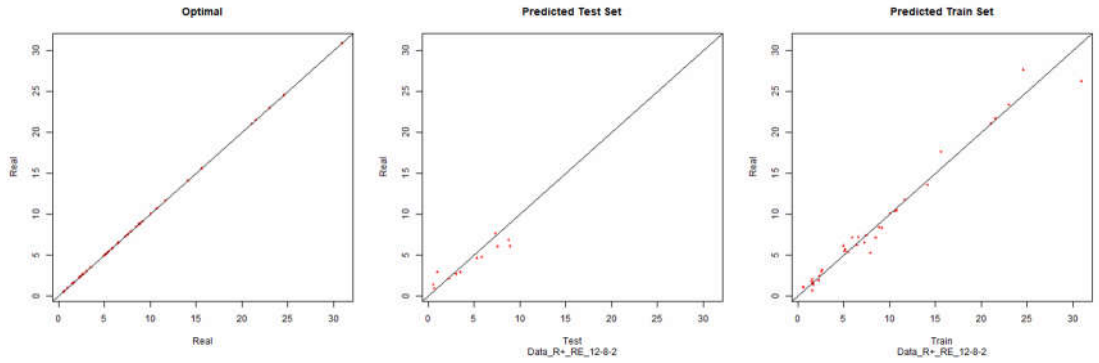


Figure 16. Modelling Charts Phase III Economic Performance (Backpropagation Resilient Positive) ANN: 12-8-2.

Economic Performance– Positive Resilient

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 May 2024 doi:10.20944/preprints202405.1008.v1

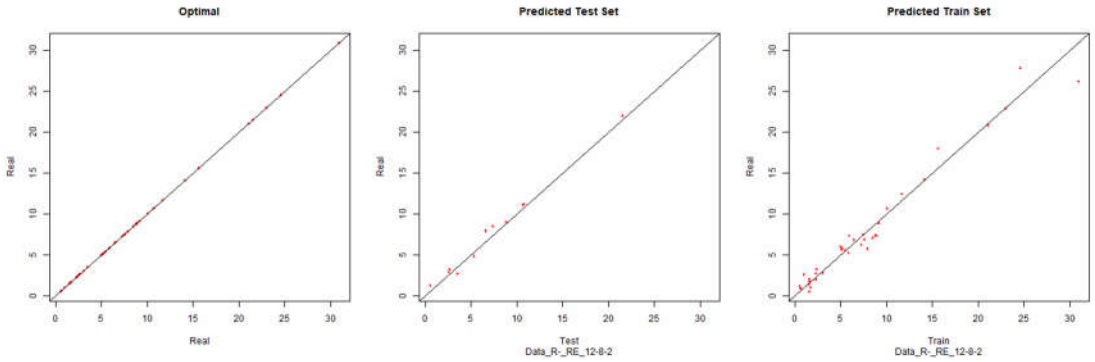


Figure 17. Modelling Charts Phase III Economic Performance (Backpropagation Resilient Negative) ANN:12-8-2.

As it can be seen from the graphics, while using the pure backpropagation algorithm, the desired environment cannot be modelled. With the processing and analysis of the financial data of the companies and filtering the companies that do not fit the defined assumptions, it is possible to model the desired environment for the economic yield with some degree of precision, using the algorithms of negative and positive resilient backpropagation.

2.4.5. 4th Phase – Processing Financial Data

The next data processing will use the delta (δ). The delta is a growth indicator for comparison between the results of 2013 and 2014. In this way for the inputs of ANNs it will be used the normalized data from the survey and for the outputs it will be used the delta of the component of the financial results to be modelled.

Revenue – Backpropagation

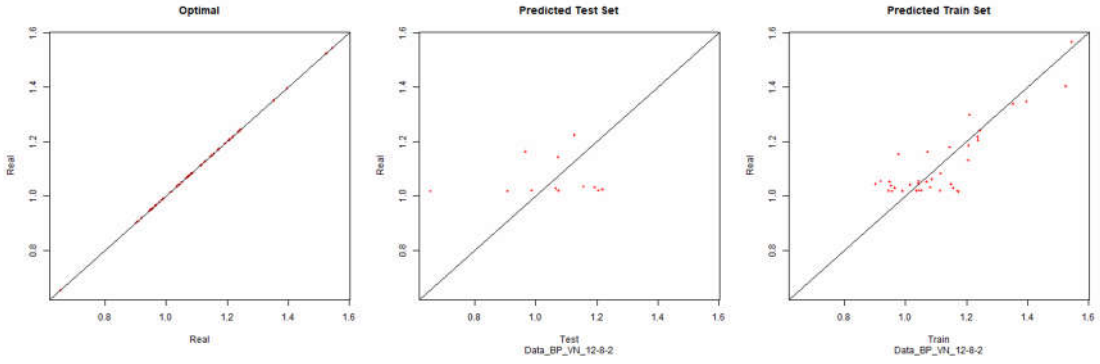
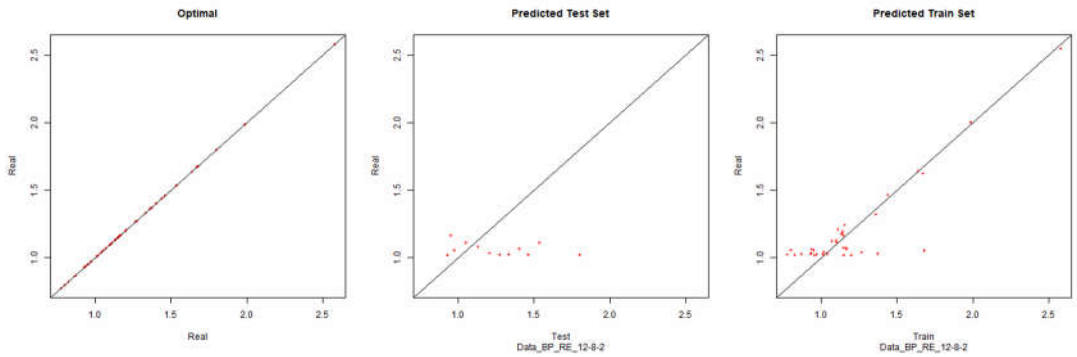


Figure 18. Modelling Charts Phase IV Revenue (Backpropagation) ANN:12-8-2.

Economic performance – Backpropagation



 doi:10.20944/preprints202405.1008.v1

Figure 19. Modelling Charts Phase IV Economic Performance (Backpropagation) ANN:12-8-2.

EBIT – Backpropagation

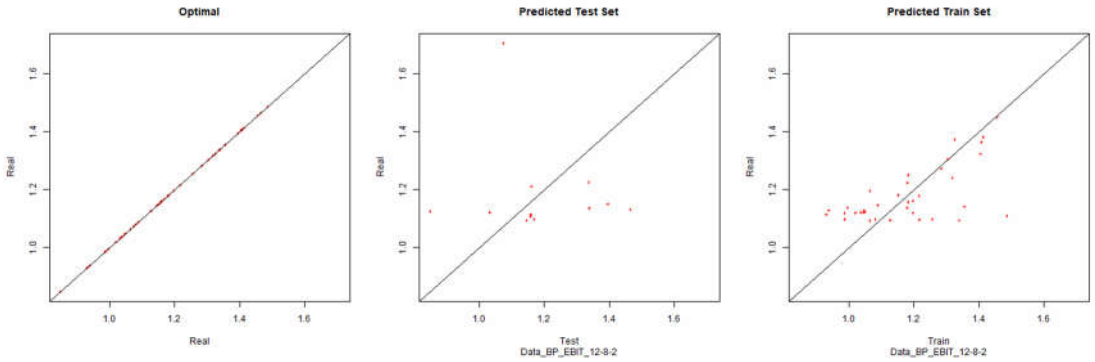


Figure 20. Modelling Charts Phase IV EBIT (Backpropagation) ANN:12-8-2.

Revenue – Resilient Positive

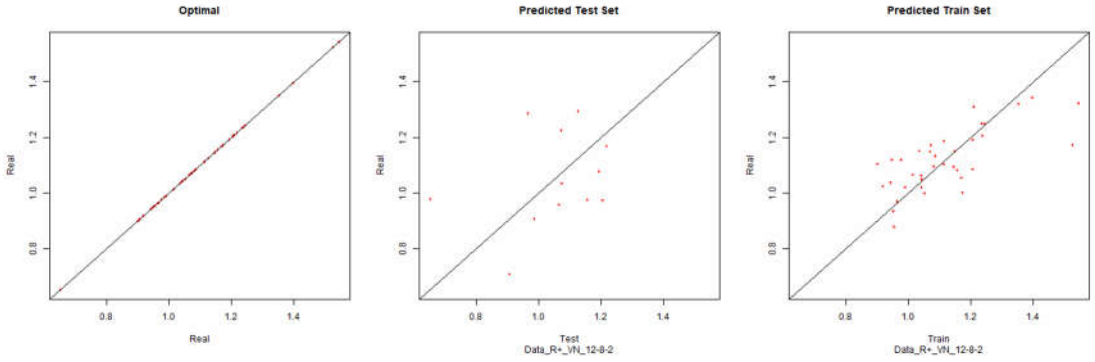


Figure 21. Modelling Charts Phase IV Revenue (Backpropagation Resilient Positive) ANN:12-8-2.

Economic Performance – Resilient Positive

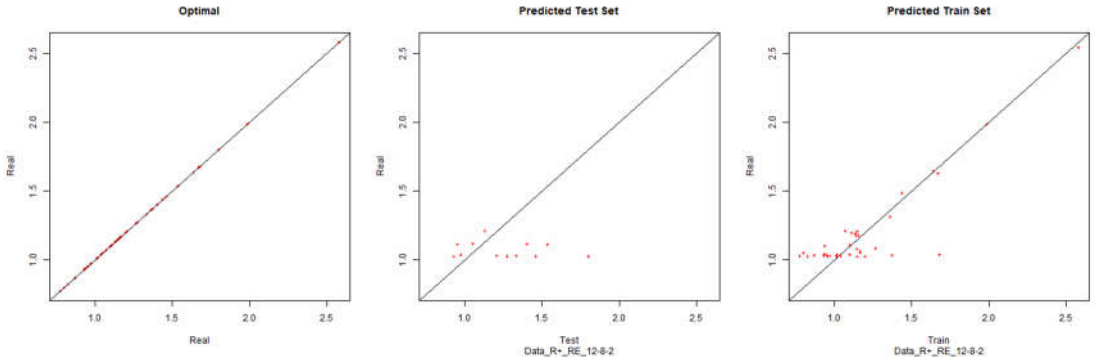


Figure 22. Modelling Charts Phase IV Economic Performance (Backpropagation Resilient Positive) ANN:12-8-2.

EBIT – Resilient Positive

 doi:10.20944/preprints202405.1008.v1

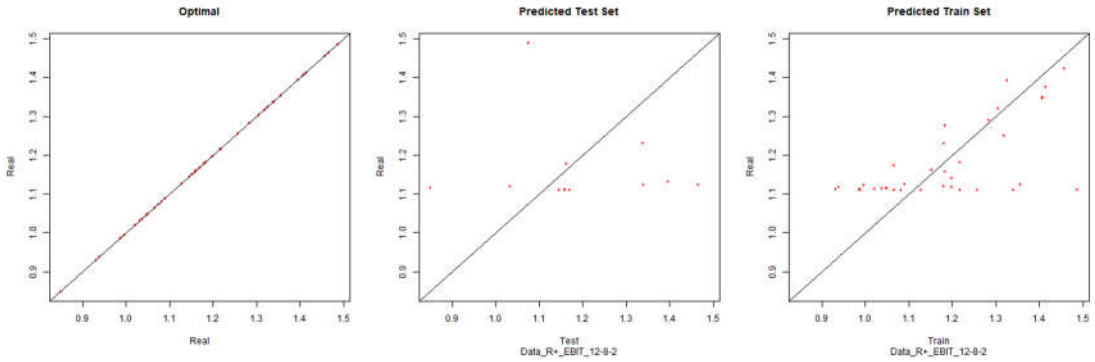


Figure 23. Modelling Charts Phase IV EBIT (Backpropagation Resilient Positive) ANN:12-8-2.

Revenue - Resilient Negative

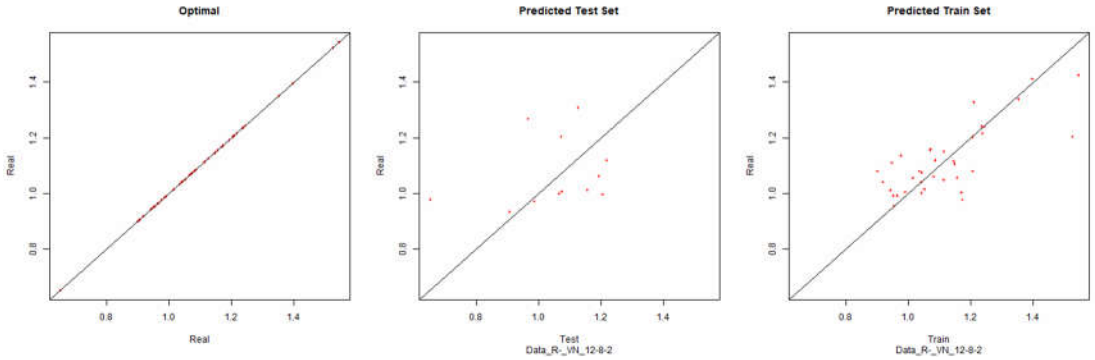


Figure 24. Modelling Charts Phase IV Revenue (Backpropagation Resilient Negative) ANN:12-8-2.

Economic Performance – Resilient Negative

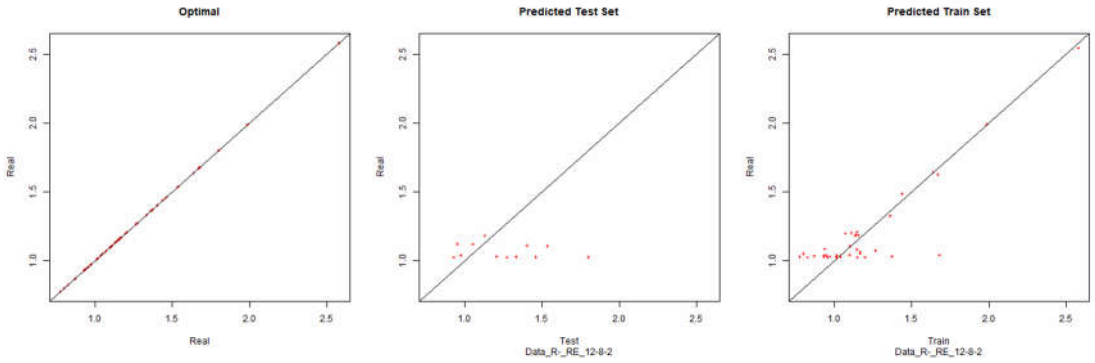


Figure 25. Modelling Charts Phase IV Economic Performance (Backpropagation Resilient Negative) ANN:12-8-2.

EBIT – Resilient Negative

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 May 2024 doi:10.20944/preprints202405.1008.v1

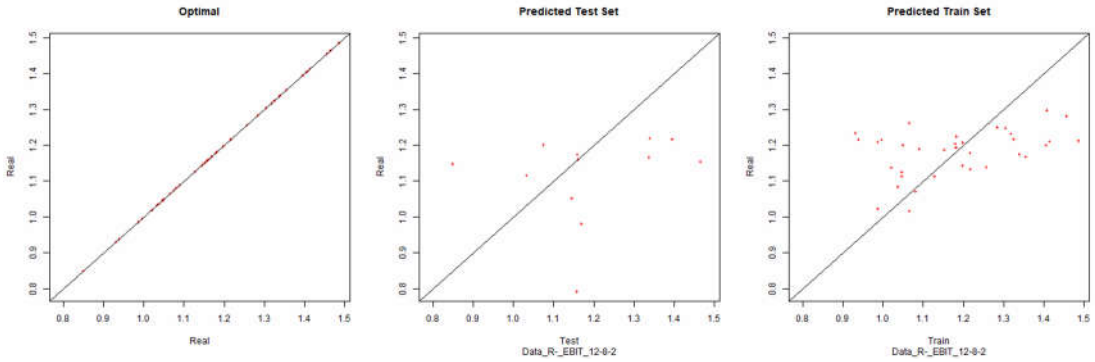


Figure 26. Modelling Charts Phase IV EBIT (Backpropagation Resilient Negative) ANN:12-8-2.

EBITDA – Resilient Negative

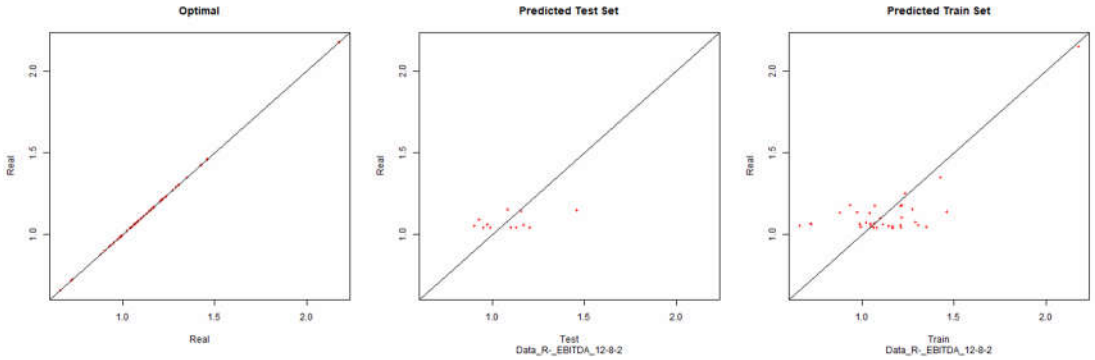


Figure 27. Modelling Charts Phase IV EBITDA (Backpropagation Resilient Negative) ANN:12-8-2.

As it can be seen from the graphics it was possible to model the environments with some degree of precision. However the modelling of net results hasn't considered successful. Although the degree of precision is not high in the other financial results, it is necessary to remember that the objective of the study is to analyze the impact of the data processing in the modelling by neuronal networks. Therefore an exhaustive study about the best topology for modelling was not made. It should be noted that the data processing enabled the modelling with all the algorithms used.

3. Conclusions

3.1. Main Conclusions of the Research

Filtering the companies accepted to integrate the modelling allowed the reduction of noise in the model of the proposed environment. One should be aware that all the phases of data elimination may restrict the environment where the model is valid. This means that when not considering certain companies of the survey, the circumstances that lead to the exclusion of those companies should be taken into account when applying an organization to the model.

The number of samples should allow not only the calculation of the model but also the verification of the performance of the model. Although some organizations could have been eliminated from the creation of the model, these can be used to verify the performance of the model. This only applies if the reason of exclusion doesn't have a direct impact in the result that is being calculated. For example, it can be a good test to the model limits if what is being calculated is a forecast of the revenue for a company that wasn't accepted because it financial performance was greater than 200%.

It was verified that the data processing and filtering had a significant impact in modelling the desired environment. Some experiments didn't bring a satisfactory result. These less satisfactory

 doi:10.20944/preprints202405.1008.v1

results could exist because the environment couldn't be model through ANNs, but no further studies were made. Studying the possibility of modelling through another topologies or algorithms could induce information inadequate to the study.

In the initial phase the modelling wasn't possible in any of the studied cases, inclusive in backpropagation resilient algorithms, the ANN could not be calculated. In phase IV not only it was possible to calculate de ANNs with all tested algorithms, but in some cases a satisfactory modelling could be achieved.

Although the evaluation of the test sets generated through the ANN was not one of the purposes of the study, it was possible to verify, in some cases, a good prevision of the financial results, such as the case of phase IV of revenue with backpropagation resilient negative.

3.2. Major Contributions of the Research

The importance of this study is directly related with the necessity of modelling subjective environments. Subjective data could make the modelling of environments without any kind of data processing, impossible. The method evaluated in this study should allow a better modelling after processing the data or even allow to modelling where previous it wasn't possible. This way, a subjective data set from an inquiry can be used to model an environment through ANNs, that otherwise should be impossible to model.

Although it can reduce the scope where the model can be applied, processing data before training the ANNs, can open new opportunities to model environments based in behaviors. If the same behavior can have different results it will be extremely difficult to model predictive behavior environments based in mathematical models. So, this method has a significant importance in improving the changes on the creation of successful predictive models based on surveys data.

References

1. Dancey, C.; Reidy, J. (2017). *Statistics without Maths for Psychology*, 7th ed. Pearson.
2. Freire, A. (2008). *Estratégia: Sucesso em Portugal*, 12.ª ed. Lisboa: Editorial Verbo.
3. Gonçalves, A. J. (2016). *Master Thesis: Método para modelação do impacto de estratégias nos resultados através de Redes Neuronais*, Instituto Superior de Contabilidade e Administração de Lisboa/Instituto Politécnico de Lisboa.
4. Hitt, Michael; Ireland, R. Duane; Hoskisson, Robert (2011). *Concepts Strategic Management: Competitiveness & Globalization*, 9.ª Ed. Canada: Cengage South-Western.
5. Hoel, P. (1966). *Introduction to Mathematical Statistics*, 3rd ed. John Wiley and Sons, Inc.
6. Montgomery, D.; Runger, G. (2002). *Applied Statistics and Probability for Engineers*, 3rd ed. John Wiley and Sons, Inc.
7. Moore, D. (2003). *The Basic Practice of Statistics*, 3rd ed. Freeman Publishers.
8. Porter, M. (1996). What is strategy?, Harvard Business Review, Product Number 4134
9. Santos, A. (2008). Gestão Estratégica: Conceitos, modelos e instrumentos, Lisboa: Escolar Editora.
10. Soong, T. (2004). Fundamental of Probability and Statistics for Engineers, John Wiley and Sons, Inc.