Journal of Infrastructure, Policy and Development 2024, 8(9), 7108.
https://doi.org/10.24294/jipd.v8i9.7108

Article

Survey data preprocessing for optimal modelling through ANNs applied to management environments

Joaquín Texeira-Quirós[1], Maria do Rosário Texeira Justino[2], António José Gonçalves[1], Marina Godinho Antunes[3], Pedro Ribeiro Mucharreira[4,*]

[1] Department of Economics and Business Sciences, Autonomous University of Lisbon, 1169-023 Lisboa, Portugal
[2] Department of Management, Lisbon Accounting and Business School, Polytechnic University of Lisbon, 1069-035 Lisboa, Portugal
[3] Department of Accounting and Auditing, Lisbon Accounting and Business School, Polytechnic University of Lisbon, 1069-035 Lisboa, Portugal
[4] Department of Education, CI-ISCE, ISCE—Instituto Superior de Lisboa e Vale do Tejo, 2620-379 Odivelas, Portugal
*** Corresponding author:** Pedro Ribeiro Mucharreira, pedro.mucharreira@isce.pt

CITATION

Texeira-Quirós J, Justino MdRT, José Gonçalves AJ, et al. (2024). Survey data preprocessing for optimal modelling through ANNs applied to management environments. Journal of Infrastructure, Policy and Development. 8(9): 7108. https://doi.org/10.24294/jipd.v8i9.7108

ARTICLE INFO

Received: 12 June 2024
Accepted: 10 July 2024
Available online: 6 September 2024

Abstract: Surveys are one of the most important tasks to be executed to get valued information. One of the main problems is how the data about many different persons can be processed to give good information about their environment. Modelling environments through Artificial Neural Networks (ANNs) is highly common because ANN's are excellent to model predictable environments using a set of data. ANN's are good in dealing 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 samples with the same input configuration has different outputs, which can be the case of survey data, it can be a major problem for the success of modelling the environment. The environment used to demonstrate the study is a strategic environment that is used to predict the impact of the applied strategies to an organization financial result, but the conclusions are not limited to this type of environment. Therefore, is necessary to adjust, eliminate invalid and inconsistent data. This permits one to maximize the probability of success and precision in modeling the desired environment. This study demonstrates, describes and evaluates each step of a process to prepare data for use, to improve the performance and precision of the ANNs used to obtain the model. This is, to improve the model quality. As a result of the studied process, it is possible to see a significant improvement both in the possibility of building a model as in its accuracy.

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

1. Introduction

With the dynamism inherent to some environments, forecast modelling is a tool that can permit the prediction of the state of the environment after an application of an action. Hoel (1966) defends that the advantage of the use of mathematical models is to permit the forecast of results. Gonçalves (2020) 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.

The modelling of real environments with high dynamism may depend on many variables making it 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 describes 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.

2. Data processing to improve ANN's models

Gonzalez Zelaya (2019) defends that data preprocessing is a generally accepted concept as a process that has a significant impact on final results. Supporting this affirmation, Maharana and Nemade (2022) argues that one of most important steps to influence performance of supervised machine learning algorithms is data preprocessing.

Mumuni and Mumuni (2024) refers that although an emergent importance has driven by some small number of studies in data processing methods, with exception of few works, who dedicated a small of their research to automated data processing, the majority of the published articles in automated machine learning doesn't focus on processing data. Gonzalez Zelaya (2019) support this affirmation in a general way, defending that there is not much work done to measure the impact of data preprocessing in machine learning activities.

Cai et al. (2018) discuss some of the most common methods for evaluating indicators for automatically selecting resources using supervised, unsupervised and semi-supervised methods.

Additionally, Garcia-Carrasco et al. (2023) defends that the decision about which preprocessing techniques should be applied depends on the experience of experts. The correct application of the techniques may depend on the data model to be used, this can lead to a process exposed to errors that requires whoever uses it to know many of the combinations that may appear.

In a study developed to predicting student academic performance in which the ANNs needed to be used in combination with data preprocessing to allow greater effectiveness of their findings in evaluating academic achievement. ANNs have indicated high accuracy in predicting the outcome of academic achievements (Baashar et al., 2022).

The difference between this study and others in same area, it is that this work will demonstrate step by step the data preprocessing with detail and measure the impact of each step on the model's quality. The importance of the study is directly related, firstly, to the difficulty of applying data pre-processing in order to improve the quality of the model and secondly, to the difficulty of finding scientific articles that detail each step of the process. The importance in the study is inherent to the necessity of improving models' quality.

3. Empirical research

In this sense, this article intends to describe and justify the processing and filtering of 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 technology 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 (Appendix) 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 data used to describe the process are from 2012, 2013 and 2014. The age of the data is due to two critical factors. The first is that these are data that require confidentiality due to the fact that they are strategies used by real companies and which are unlikely to make current data available. The second reason is that it was necessary to test whether the data could be used to obtain a model with adequate performance. The study inherent to the creation and evaluation of this model has been published (Justino et al., 2024).

The importance of these studies is inherent to the fact that we may be working with data that could never allow a model, based on neural networks, to be created. In that case this study could never give satisfactory results. Even so, an impact on the importance of the study or even on the performance of the model is not expected due to the age of the data.

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

3.1. Research hypotheses

It should be remembered that the hypothesis must agree with the objective and 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 through ANNs before and after the filtering and processing of the initial data. An important consideration it is that the data processing 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 have been filtered and processed, comparing to the model obtained from the initial data obtained through a questionnaire considering a significant degree of subjectivity.

Therefore, the main objective of the study is to analyze and evaluate a process for obtaining processed data that makes it possible to model an environment through ANNs even if it is not possible with the initial data.

3.2. 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 scope of each question in the questionnaire.

Answers range to each question.

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

Questions	Strategy	Interval	Observations
Q1	Price Increase /Reduction	[−9, 9] ϵ Z	Reduction: Negative Increase: Positive
Q2	Quality Increase /Reduction	[−9, 9] ϵ Z	Reduction: Negative Increase: Positive
Q3	Reduced Personnel Cost	[0, 9] ϵ Z	
Q4	Investment	[0, 9] ϵ Z	
Q5	Decrease Financing	[0, 9] ϵ Z	
Q6	Product Diversification/Specialization	[−9, 9] ϵ Z	Specialization: Negative Diversification: Positive
Q7	Reduction / Increase of Customers or Markets	[-9, 9] ϵ Z	Increase: Negative Reduction: Positive
Q8	Business Synergies	[0, 9] ϵ Z	
Q9	Product Disclosure	[0, 9] ϵ Z	
Q10	Business Reorganization	[0, 9] ϵ Z	
Q11	Renegotiation with Suppliers	[0, 9] ϵ Z	

The first column [Questions] refers to the reference of the question, the second column [Strategy] refers to the strategy inherent to the question and the third column [Interval] indicates which values the answer to the question can be inserted in. The fourth column [Observations] indicates whether the strategy is bidirectional and how the values should be interpreted.

A possible representation of the answers to the questionnaire by the managers are presented in **Table 2**:

Example of questionnaire answers internal representation.

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

Table 2. (*Continued*).

	Inc. 1	Inc. 2	Inc. 3	Inc. 4	Inc. 5	Inc. 6	Inc. 7	Inc. 8	Inc. 9
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 table represents an example of how the responses from 8 incorporations could be represented. 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. Each answer can have integer values between 0 and 9 (−9 and 9, in the case of bidirectional strategies). An answer of 0 it means that the strategy was not applicable in the organization and 9 or −9 it means that the strategy is highly applicable (or critical) in the organization.

3.3. 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.

In the first phase, the data is analyzed in general, and a certain distribution is identified. The data is filtered by excluding all organizations with outliners data. Each organization that doesn't have the expected behavior is eliminated.

In the second phase, survey data is normalized. Because each manager can have their own beliefs in each level of applicability represents, responses to the questionnaires are standardized so that they have a value inherent to the strategic priorities of each organization.

In the third phase, the financial state of organizations is analyzed and all organizations that do not fit within normal parameters are excluded from the study. For example, an organization that has had a 300% increase in turnover is excluded from the study as this increase is possibly not a consequence of a short-term strategic implementation. This phase differs from the first phase since in the first phase strategic behavior is evaluated and, in this phase, financial behavior is evaluated.

In the fourth phase, the financial results used as output from the neural network are normalized. The range of financial results can be enormous. Instead of evaluating raw financial results, you will evaluate the rate at which those financial results change. For example, if an organization increased its turnover from 10000 monetary units (m.u)

to 20000 m.u., it will considered as 2.00 (200%) as a result.

The experiments focus on modeling three financial indicators of different natures: revenue, economic performance and EBIT. Because the nature of these indicators differs mainly in the entropy generated by variables that are not considered in the model, and are difficult to evaluate, the conclusions will be more precise and coherent with the reality of the study, since there are different levels of difficulty in modeling the environments considered.

The generated neural networks although similar, may have some different characteristics. 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.

Although the choice of topology and specific parameters of the ANN is important for the quality of the model (Lopez-Ramirez et al., 2023), in this study, once the topology and parameters of the ANN were chosen, these criteria did not change throughout the study, to better evaluate the impact exclusively data processing.

In the end, it must be possible to understand the importance of processing data in order to increase the possibility of success in modeling an environment through ANNs.

3.4. Analysis and discussion OH the results

3.4.1. Initial phase

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, positive resilient backpropagation and negative 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 test set. 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. For the training set, 75% of the samples were used and the remaining 25% of the samples were associated with the test set.

In this study, it will be presented three graphics for each one of the tests, as can be seen in **Figures 1–3**. 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 must 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).

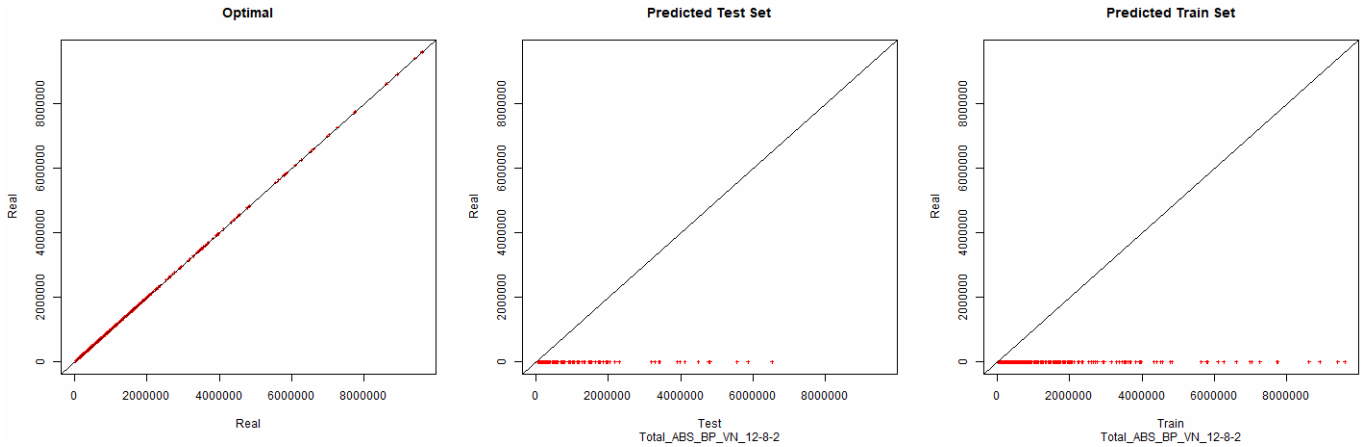

Figure 1. Initial modelling charts revenue (backpropagation) ANN: 12-8-2.

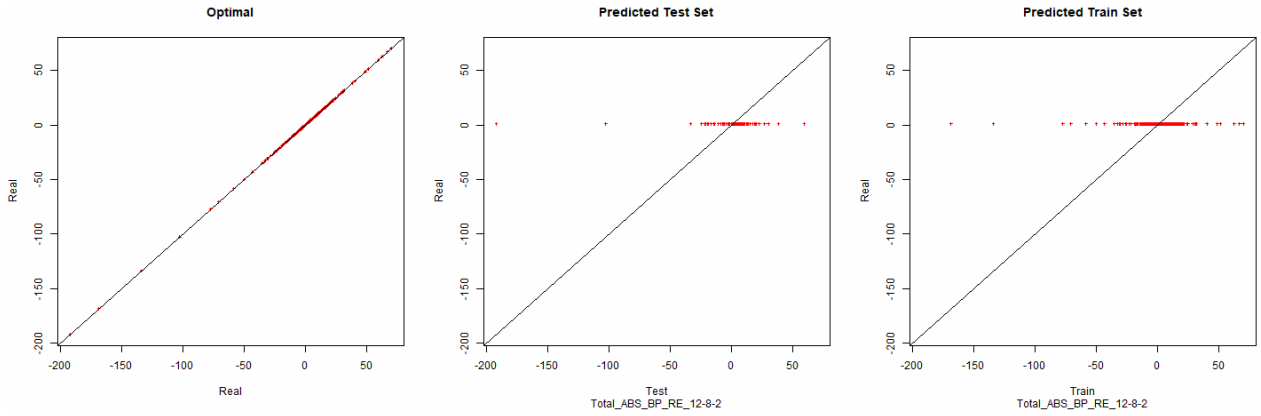

Figure 2. Initial modelling charts economic performance (backpropagation) ANN: 12-8-2.

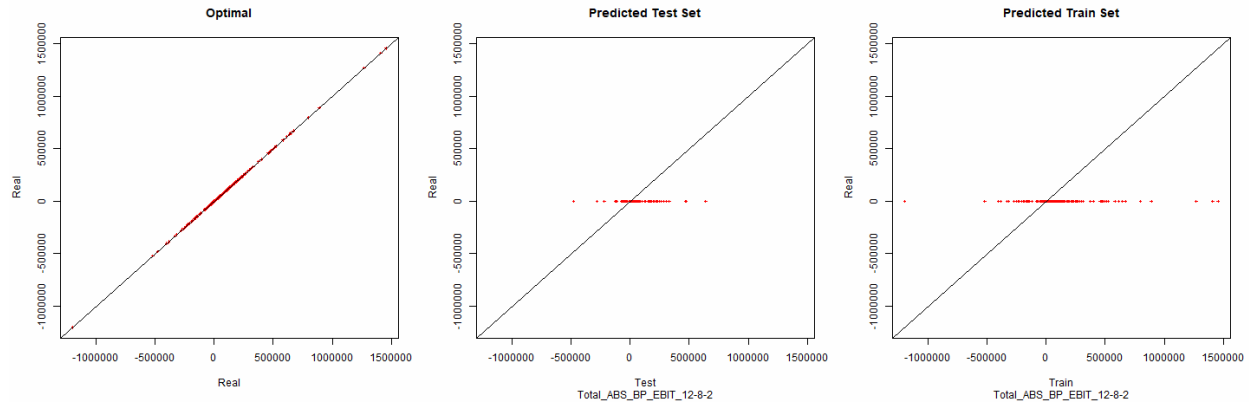

Figure 3. Initial modelling charts EBIT (backpropagation) ANN: 12-8-2.

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.

3.4.2. 1st phase—Strategic behavior

The problem that we try to mitigate at this stage is the fact that there are answers to the questionnaire that do not bring veracity to the model. This means an increase in entropy that generates noise in the training of the neural network and therefore an inadequate precision for the model.

Analyzing the distribution of the absolute sum of the values answered by the respondents the following graphs—**Figure 4**—were obtained:

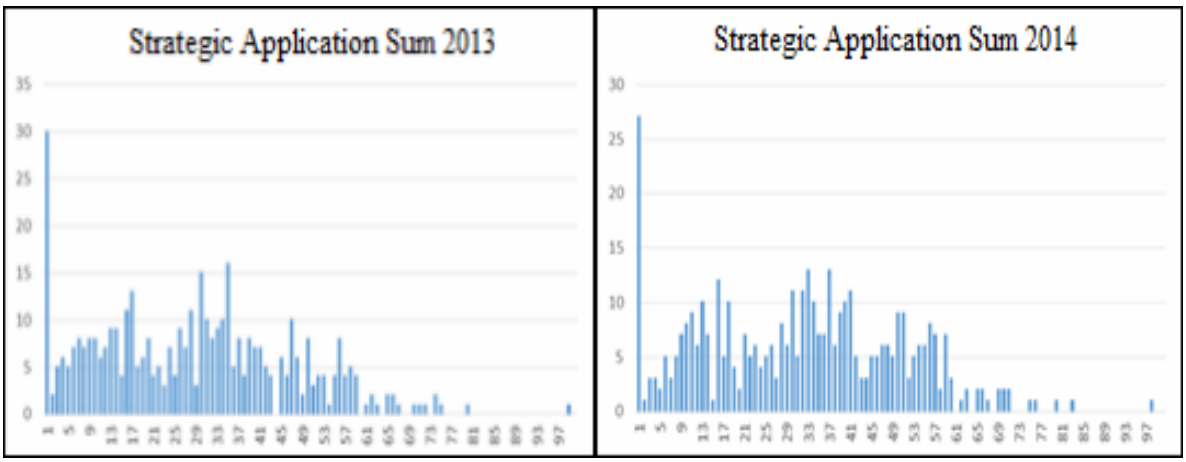

Figure 4. Absolute sum of the application of strategies for the years 2013 and 2014.

It can be seen in these 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 **Figure 5**:

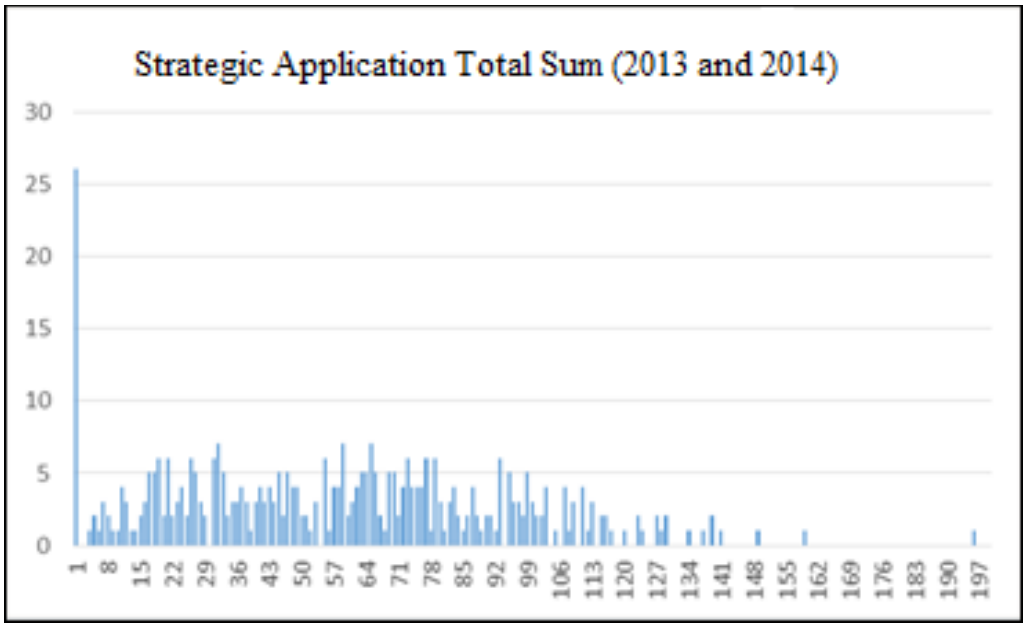

Figure 5. Total sum of the application of strategies in the years 2013 and 2014.

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) 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.

David Moore (2003) 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, as seen in **Figure 6**.

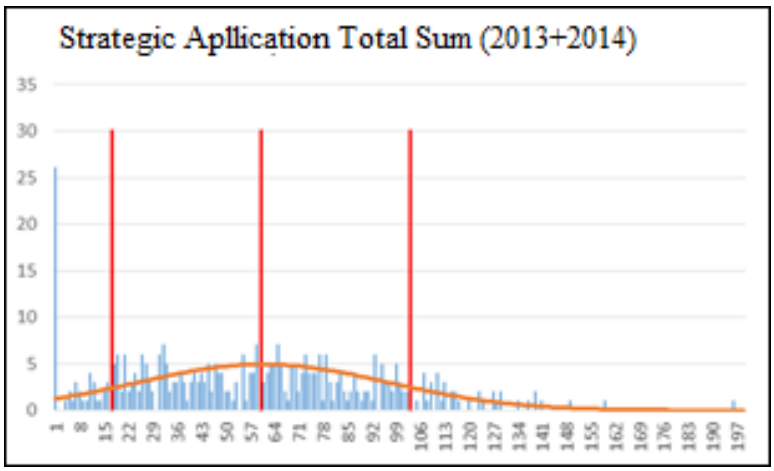

Figure 6. Projected normal distribution in the total absolute sum (2013 and 2014).

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

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 (Soong, 2004, p. 200).

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 and Reidy (2017) define the standard deviation as the measure of the degree to which the sample deviates from the mean. 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 (Hoel, 1966, p. 101). 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, as can be seen in **Figures 7–9**.

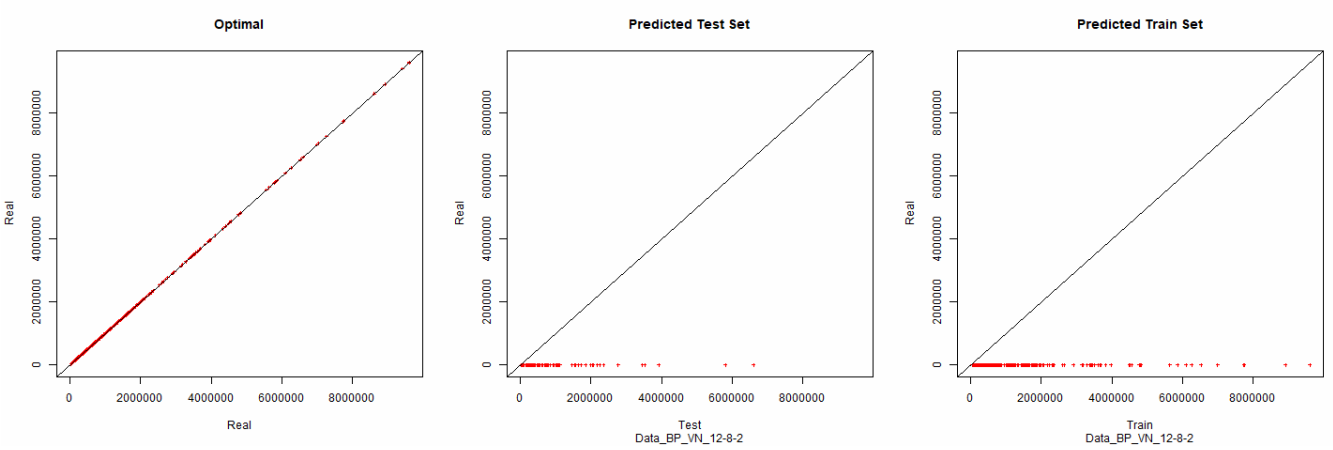

Figure 7. Modelling Charts Phase I Revenue Backpropagation ANN: 12-8-2.

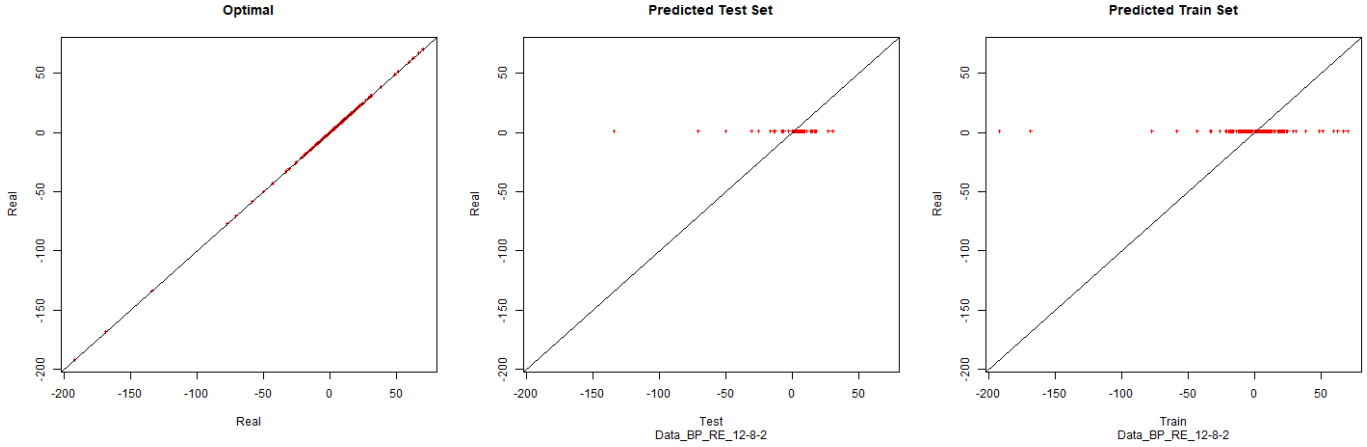

Figure 8. Modelling charts phase I economic performance backpropagation ANN: 12-8-2.

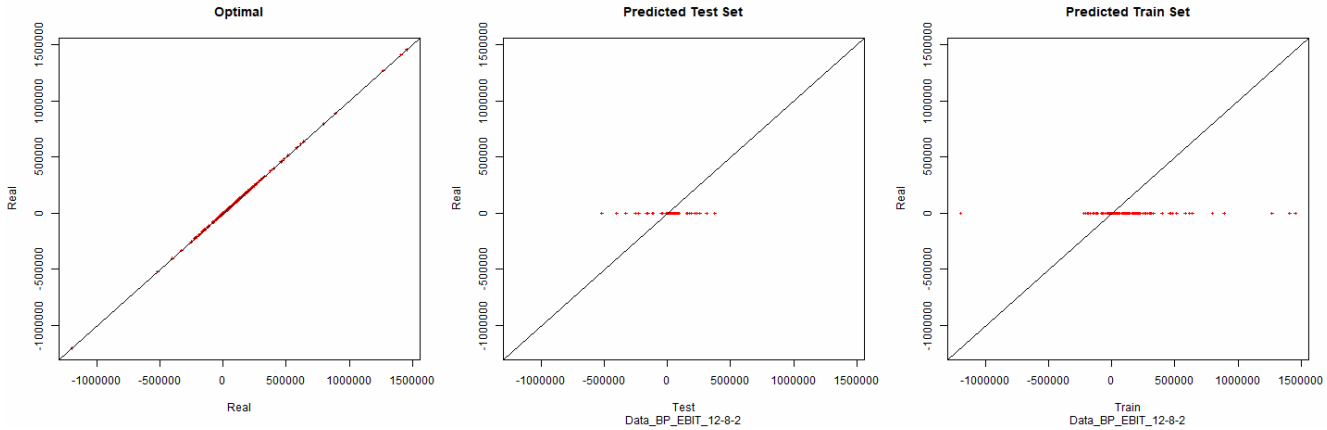

Figure 9. Modelling charts phase I ebit backpropagation ANN: 12-8-2.

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.

3.4.3. 2nd phase—Data processing

The problem associated with a survey on behavior, whether personal or strategic,

is the fact that it is difficult to accurately relate the responses from different individuals. The perception of different individuals may be different depending on their values, experiences and beliefs. At this stage, an attempt is made to mitigate the error in attempting this relationship.

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 of 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 organization.

Similarly, considering the example of the inquiry on a new flavor (flavor A) of ice cream if one respondent (A) answers for example 7 to chocolate, 3 to vanilla and 6 to flavor A, and a respondent (B) responds 8 to chocolate, 4 to vanilla and 7 to flavor A. It can be assumed that the coherence of the answers leads us to conclude that the respondent (A) and respondent (B) likes chocolate more than flavor A, but likes more flavor A than vanilla. However, 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. If there is no process to create this relationship, it is very difficult to relate the input values (questionnaire) with the output values (financial indicators) for most organizations.

To emulate a direct relationship between the values of each of the respondents it is necessary to process the results of the surveys. This 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:

Example of questionnaire answers in raw values, as can be seen in **Table 3**.

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:

Example of questionnaire answers already normalized (see **Table 4**).

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, as can be seen in **Figures 10–12**.

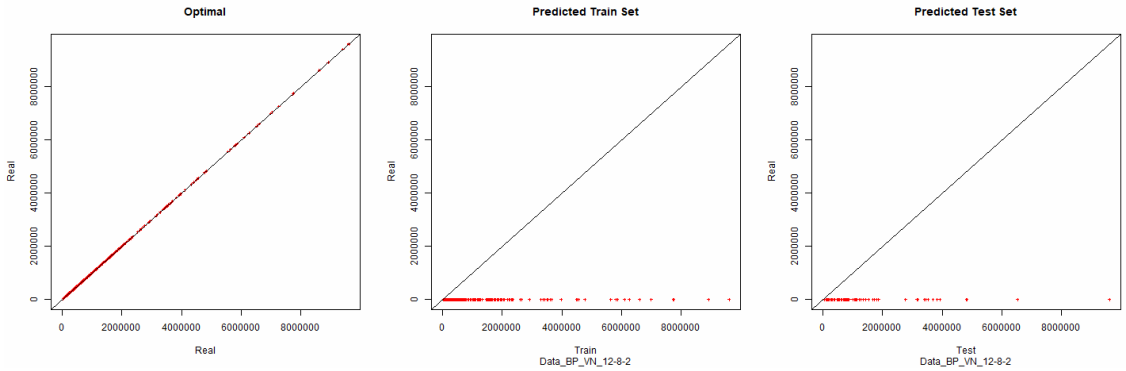

Figure 10. Modelling charts phase II revenue (backpropagation) ANN: 12-8-2.

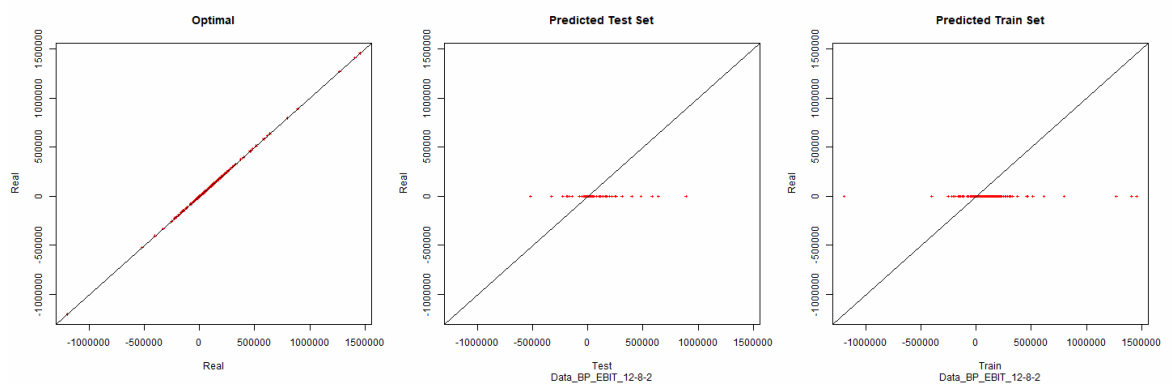

Figure 11. Modelling charts phase II economic performance (backpropagation) ANN:12-8-2.

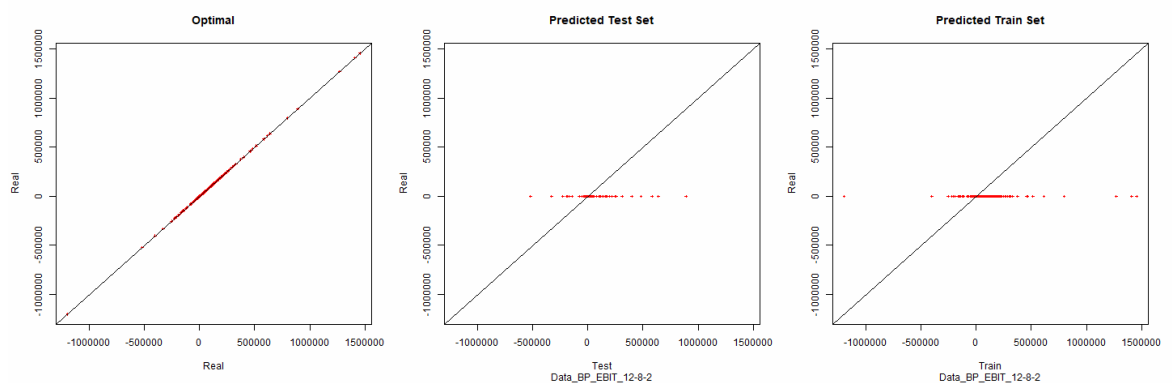

Figure 12. Modelling charts phase II EBIT (backpropagation) ANN:12-8-2.

Once more, if we look at these figures, 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.

3.4.4. 3rd phase—Processing and analysis of organizational results

The third phase is necessary to mitigate the existing problem regarding questionable financial behavior. That is, if the behavior of financial indicators from one year to the next is not a consequence of the strategies applied, it is inducing noise in the model. When introducing the compromised sample into the training or testing of the ANN, one will be introducing data that is not part of the environment to model.

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)
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.

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 (Porter, 1996).

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.

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 (Hitt et al., 2011, p. 6). 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, i.e., $\delta = (I2014 - I2013)/I2013 \times 100$ (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%.

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: That particular organization's sample 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 5**:

Organizations versus Penalty Points.

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, as can be seen in **Figures 13–17**.

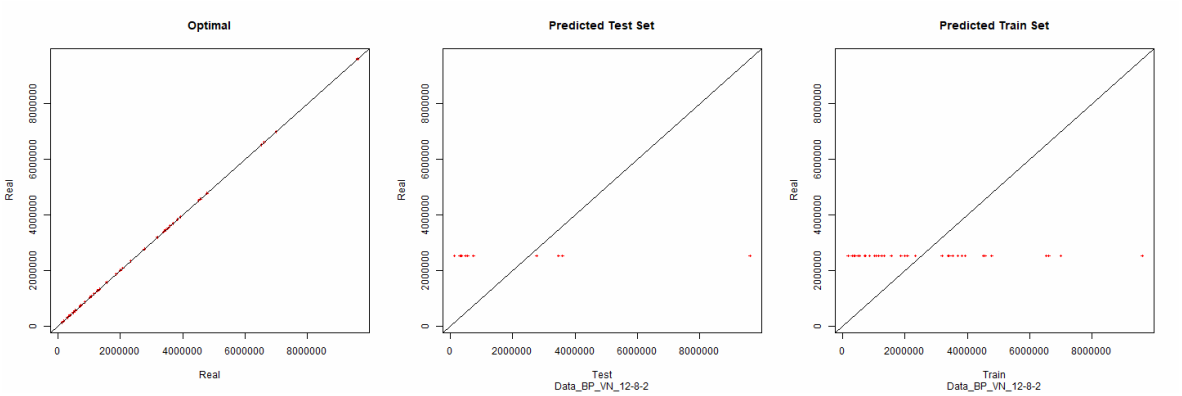

Figure 13. Modelling charts phase III revenue (backpropagation) ANN: 12-8-2.

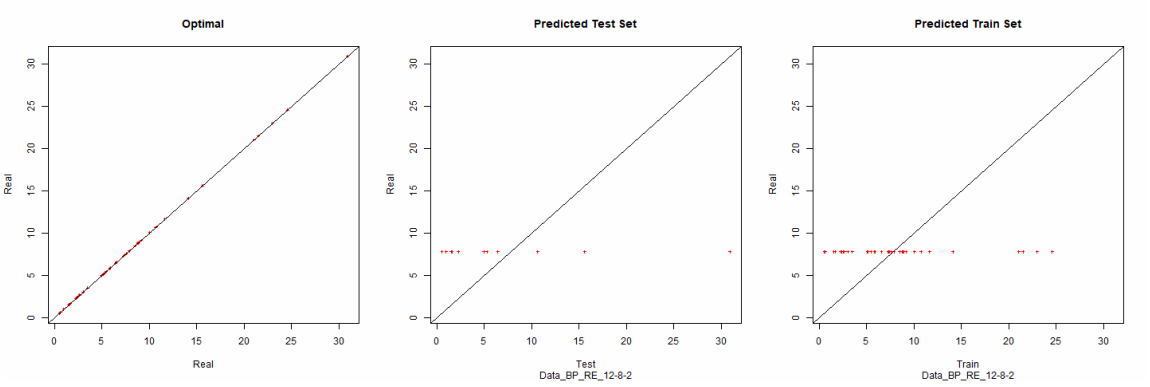

Figure 14. Modelling charts phase III economic performance (backpropagation) ANN: 12-8-2.

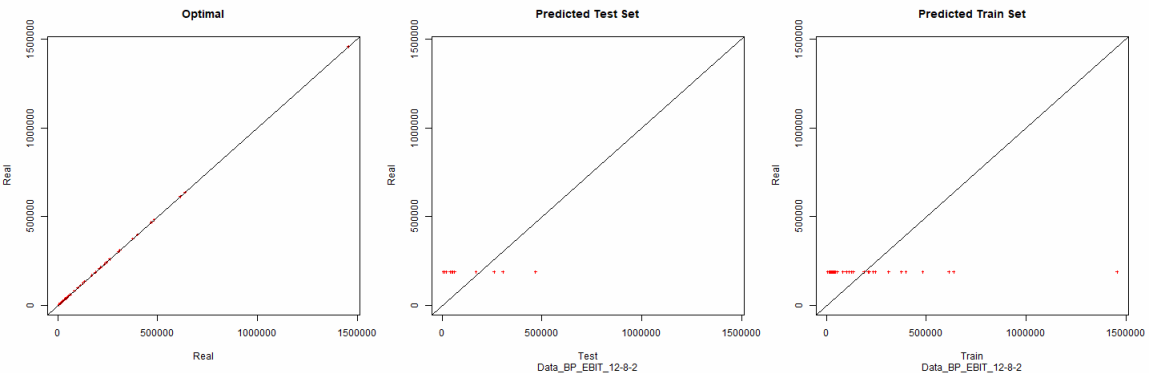

Figure 15. Modelling charts phase III EBIT (backpropagation) ANN: 12-8-2.

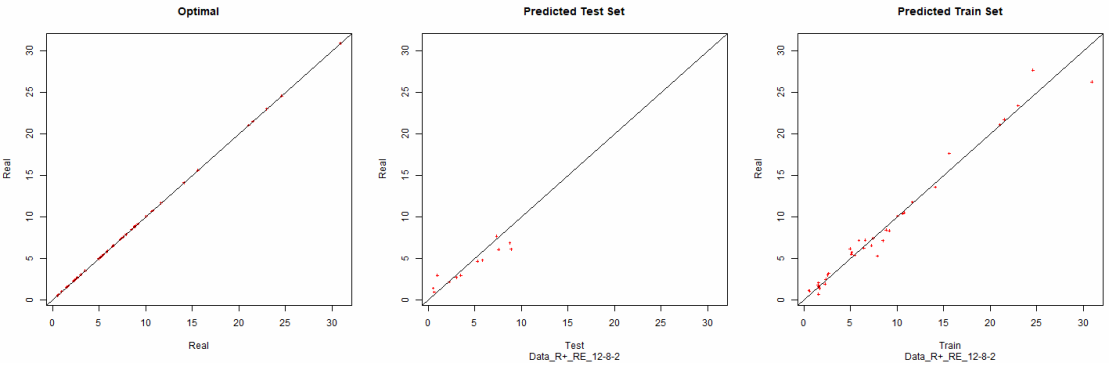

Figure 16. Modelling charts phase III economic performance (backpropagation resilient positive) ANN: 12-8-2.

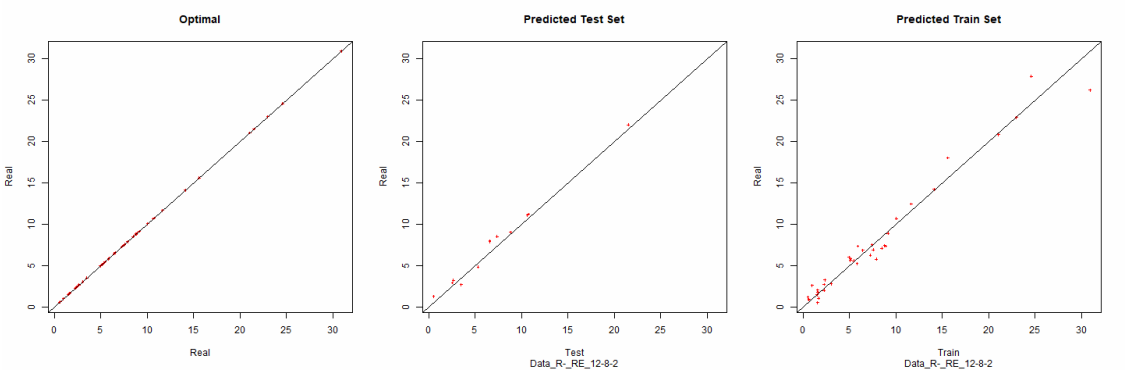

Figure 17. modelling charts phase III economic performance (backpropagation resilient negative) ANN:12-8-2.

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.

3.4.5. 4th phase—Processing financial data

The fourth phase tries to minimize the problem of the lack of relationship between the financial results of different organizations. Just as in the second phase, there was a problem with the lack of relationship between the responses to the questionnaire from different participants, the same problem also exists in the financial indicators of different organizations.

This problem is associated with the different sizes of organizations and therefore their financial indicators. Larger organizations are expected to have larger financial results depending on their size. The way to mitigate this problem is to use ratios that have information regarding increases or decreases in indicators from one year to the next.

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, as can be seen in **Figures 18–27**.

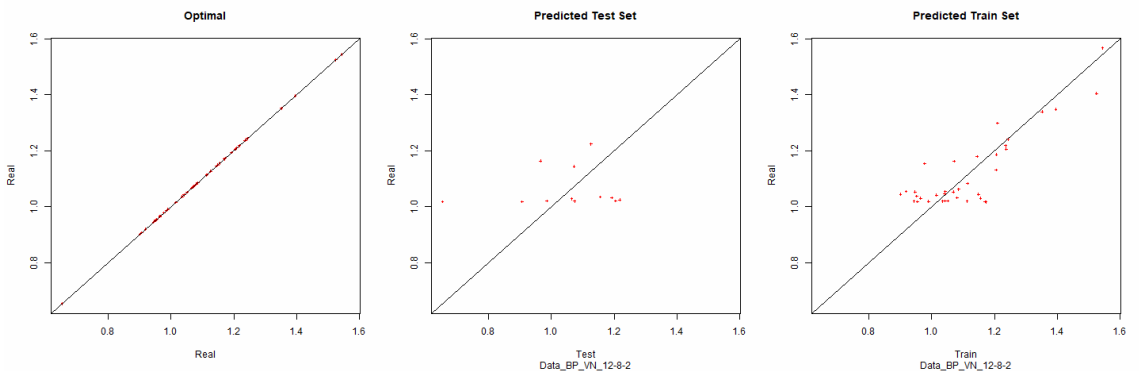

Figure 18. Modelling charts phase IV revenue (backpropagation) ANN:12-8-2.

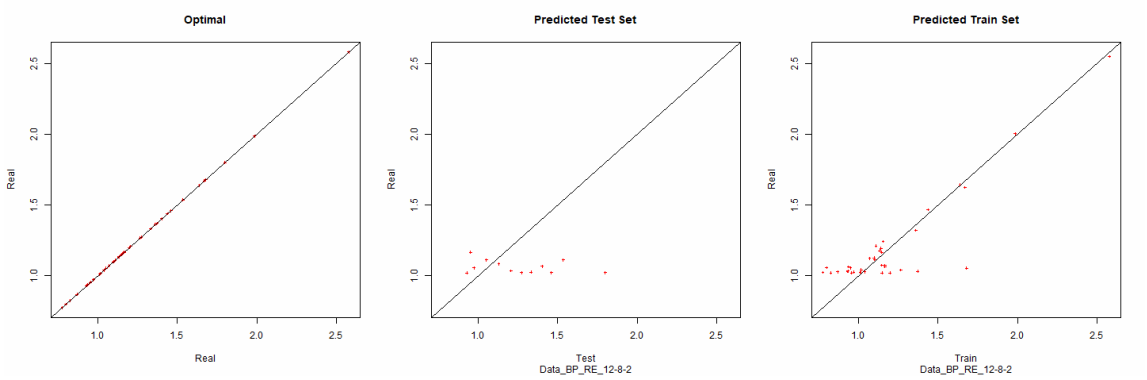

Figure 19. Modelling charts phase IV economic performance (backpropagation) ANN:12-8-2.

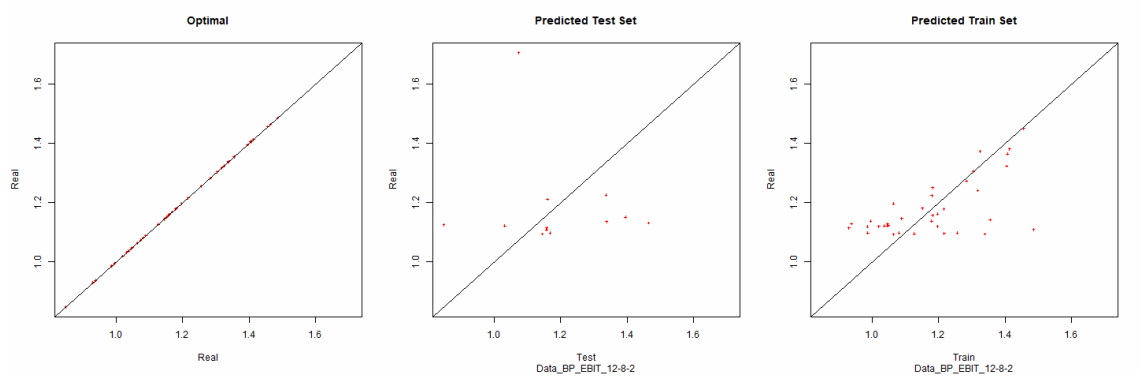

Figure 20. Modelling charts phase IV EBIT (backpropagation) ANN:12-8-2.

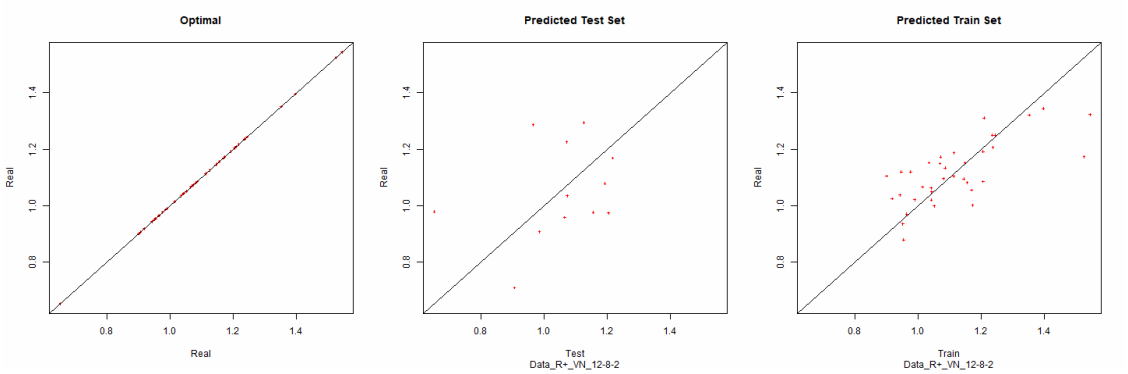

Figure 21. Modelling charts phase IV revenue (backpropagation resilient positive) ANN:12-8-2.

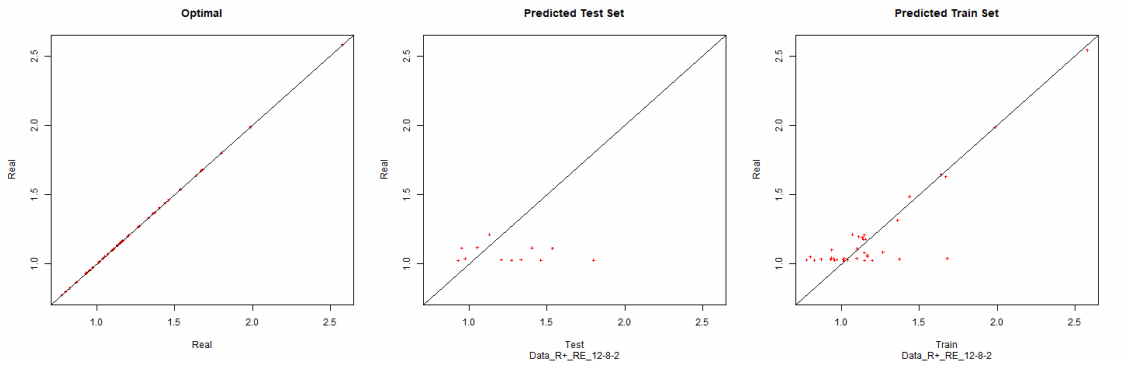

Figure 22. Modelling charts phase IV economic performance (backpropagation resilient positive) ANN:12-8-2.

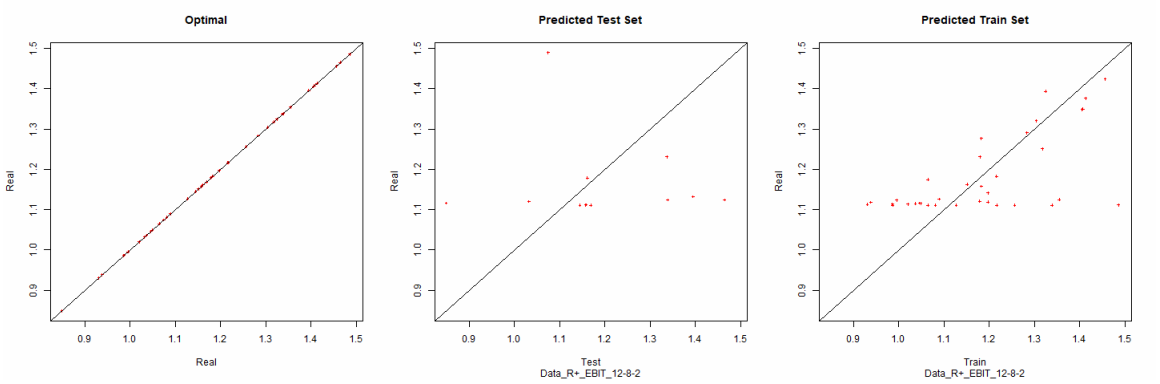

Figure 23. Modelling charts phase IV EBIT (backpropagation resilient positive) ANN:12-8-2.

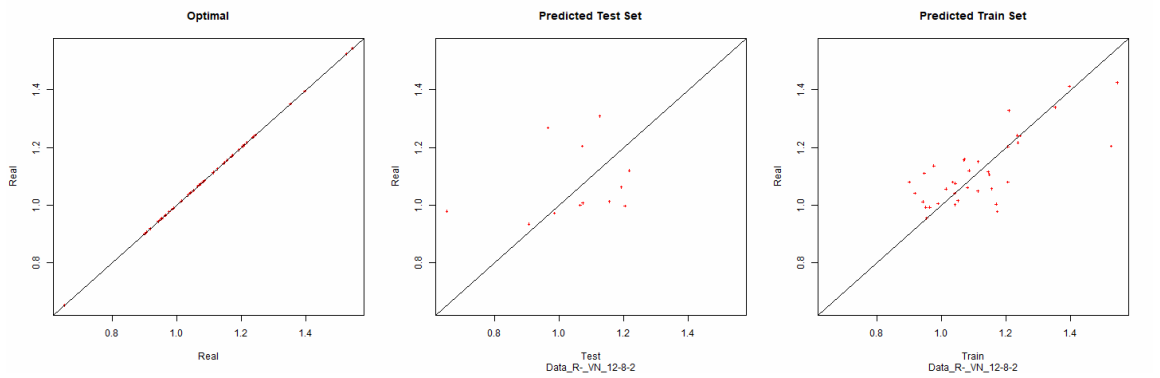

Figure 24. Modelling charts phase IV revenue (backpropagation resilient negative) ANN:12-8-2.

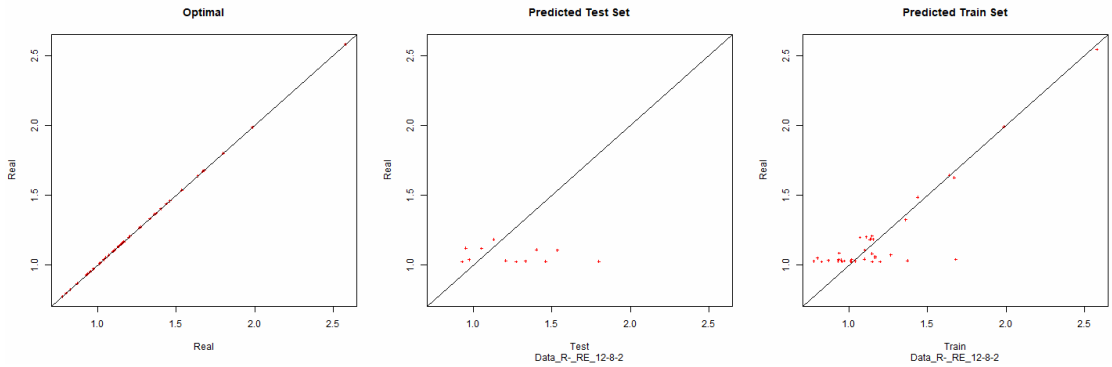

Figure 25. Modelling charts phase IV economic performance (backpropagation resilient negative) ANN:12-8-2.

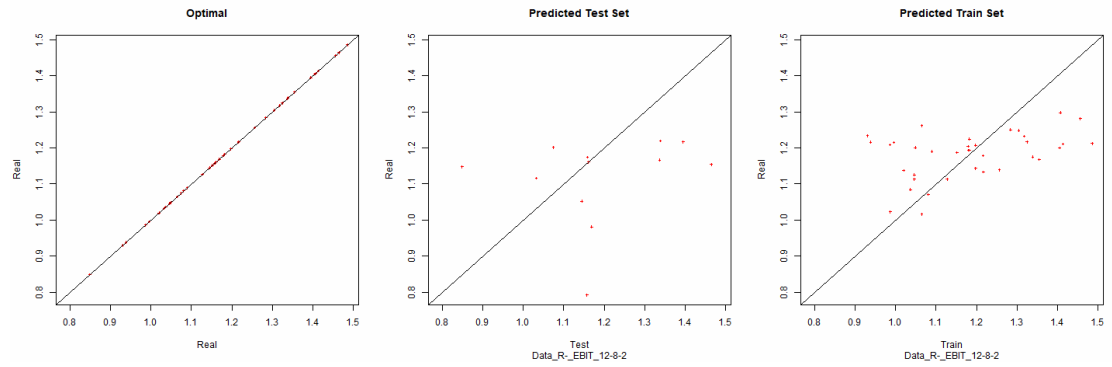

Figure 26. Modelling charts phase IV EBIT (backpropagation resilient negative) ANN:12-8-2.

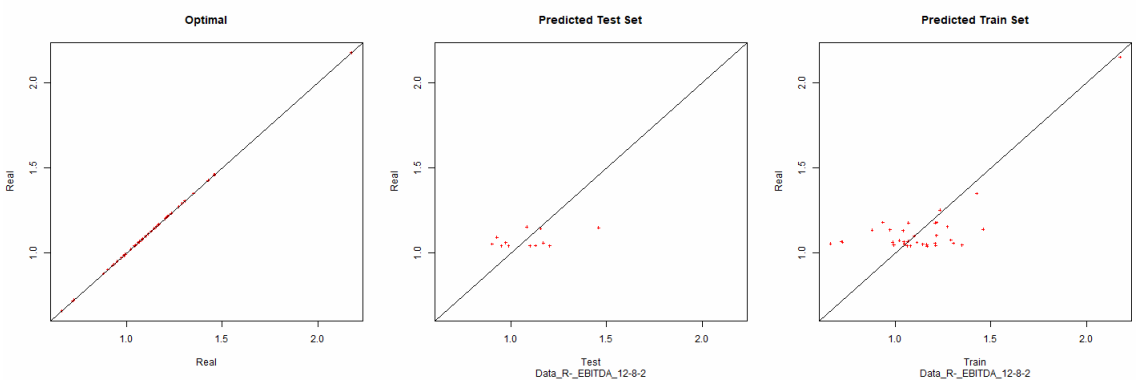

Figure 27. Modelling charts phase IV EBITDA (backpropagation resilient negative) ANN:12-8-2.

As it can be seen from the graphics it was possible to model the environments with some degree of precision. A study on revenue forecasting through applied strategies that used this data pre-processing methodology managed to obtain very good results, but in this case the analysis of the topology used and the ANN parameters were extensive (Justino et al., 2024). Although the degree of precision is not high, it is necessary to remember that the objective of the study is to analyze the impact of the data processing in the modelling by ANN. Therefore, an exhaustive study about the best topology for modelling was not carried out. It should be noted that the data processing enabled the modelling with all the algorithms used.

4. Conclusion

4.1. Main conclusions of the research

Filtering the companies and use the ones accepted to integrate the modelling samples, allow 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 considered 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.

It was verified that the data processing and filtering had a significant impact in the quality of the desired environment model. Some experiments didn't bring a satisfactory result. These less satisfactory 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 bias to the study.

In the initial phase the modelling wasn't possible in any of the studied cases, inclusive in backpropagation resilient algorithms. 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.

4.2. Major contributions of the research

The importance of this study is directly related with the necessity of modelling subjective environments. Subjective data could make impossible the modelling of environments without any kind of data processing. 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 ANNs can open up new opportunities for modeling behavior-based environments. If the same behavior can have different results, it will be extremely difficult to model predictive behavior environments based on mathematical models. Therefore, this method has significant impact in improving the quality of models based on questionnaire data.

Author contributions: Conceptualization, JTQ, MdRTJ and AJG; methodology, JTQ, MdRTJ and AJG; software, JTQ, MdRTJ and AJG; validation, JTQ, MdRTJ and AJG; formal analysis, JTQ, MdRTJ and AJG; investigation, JTQ, MdRTJ and AJG; resources, JTQ, MdRTJ and AJG; data curation, JTQ, MdRTJ and MGA; writing—original draft preparation, MdRTJ, AJG and MGA; writing—review and editing, MdRTJ, MGA and PRM; visualization, MdRTJ, MGA and PRM; funding acquisition, JTQ, MdRTJ and MGA. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The publication of this paper was partially funded by the Autonomous University of Lisbon.

Conflict of interest: The authors declare no conflict of interest.

References

Baashar, Y., Alkawsi, G., Mustafa, A., et al. (2022). Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs). Applied Sciences, 12(3), 1289. https://doi.org/10.3390/app12031289

Cai, J., Luo, J., Wang, S., et al. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Dancey, C., & Reidy, J. (2017). Statistics without Maths for Psychology, 7th ed. Pearson.

García-Carrasco, J., Maté, A., & Trujillo, J. (2023). A Data-Driven Methodology for Guiding the Selection of Preprocessing Techniques in a Machine Learning Pipeline. In: Proceedings of International Conference on Advanced Information Systems Engineering; Springer, Cham.

Gonçalves, A. J. (2020). Strategic variables for forecasting financial results in small companies through neural networks and decision trees (Spanish) [PhD thesis]. Universidad de Extremadura, Badajoz, Spain.

Gonzalez Zelaya, C. V. (2019). Towards Explaining the Effects of Data Preprocessing on Machine Learning. In: Proceeding of the 2019 IEEE 35th International Conference on Data Engineering (ICDE). https://doi.org/10.1109/icde.2019.00245

Hitt, M., Ireland, R., & Hoskisson, R. (2011). Concepts Strategic Management: Competitiveness & Globalization, 9th ed. Canada : Cengage South-Western.

Hoel, P. (1966). Introduction to Mathematical Statistics. New York, London & Sydney: John Wiley & Sons, Inc.

Justino, M. do R. T. F., Texeira-Quirós, J., Gonçalves, A. J., et al. (2024). The Role of Artificial Neural Networks (ANNs) in Supporting Strategic Management Decisions. Journal of Risk and Financial Management, 17(4), 164. https://doi.org/10.3390/jrfm17040164

Lopez-Ramirez, E., Lopez-Zamora, S., Escobedo, S., et al. (2023). Artificial Neural Networks (ANNs) for Vapor-Liquid-Liquid Equilibrium (VLLE) Predictions in N-Octane/Water Blends. Processes, 11(7), 2026. https://doi.org/10.3390/pr11072026

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Moore, D. (2003). The Basic Practice of Statistics 3rd ed. Freeman Publishers.

Mumuni, A., & Mumuni, F. (2024). Automated data processing and feature engineering for deep learning and big data applications: A survey. Journal of Information and Intelligence. https://doi.org/10.1016/j.jiixd.2024.01.002

Porter, M. (1996). What is strategy? Harvard Business Review.

Soong, T. (2004). Fundamental of Probability and Statistics for Engineers. Jonh Wiley & Sons, Inc.

Appendix

Questionnaire:

1) The price increase or decrease strategy must provide optimization of business volume results. That is, with the price increase, although there may be a drop in the number of sales, it can still compensate in the total turnover or, at least, in the operational results. Generally, by reducing the price, the objective of the strategy is to focus on increasing the number of sales and therefore increasing the turnover.
   a. Price increase?
   b. Price decrease?
2) The application of strategies such as increasing or decreasing the quality of services/products can give you a competitive advantage over your competitors. These strategies are usually accompanied by a price increase or reduction. Even a decrease in quality can be important in the organization's survival, since excessive quality may not have an impact on customer satisfaction, but have a significant impact on costs, making them excessive.
   a. Increase in service/product quality?
   b. Decrease in service/product quality?
3) Reducing personnel costs is one of the most used strategies today to reduce operational or administrative costs. This category includes staff reduction, layoffs, non-renewal of contracts, or the non-attribution of bonuses or reduction of benefits, such as the use of a company car for personal reasons.
   a. Personnel costs reduction?
4) The definition of strategies that imply an increase in investment may be inherent to the acquisition of production machines, as well as technological tools, which aim to improve processes and/or products that allow for the optimization of production, as well as an increase in production capacity.
   a. Investment increase?
5) Strategies that aim to reduce financing are inherent in the intention of reducing financing expenses, such as reducing debts with extraordinary amortizations, either from capital increases or from the organization's net results.
   a. Decrease in financing?
6) Diversification or specialization of products/services is a strategy used by organizations to provide specialized services, that is, they no longer offer non-profitable services or products. Or on the other hand, to diversify their products and services to reach a wider market and thus obtain a greater number of customers.
   a. Products/Services diversification?
   b. Products/Services specialization?
7) The strategies inherent to reducing or increasing customers or markets aim, in the case of reduction, to no longer be present in markets or available to a certain type of customer that do not provide the desired profitability. This strategy can be associated with market segments, but also with market demographics. Likewise, an increase in customers or markets can bring added value to the organization, if these markets are profitable.
   a. Client/market increase?
   b. Client/Market decrease?
8) Strategies allow synergies with commercial partners and are based on strategic partnerships with other organizations that allow an increase in customers derived from the need for complementary services.
   a. Synergies with commercial partners?
9) The application of marketing and advertising strategies has an important role in attracting customers, as well as creating value through the organization's image. With this topic we intend to know whether a careful analysis was carried out to promote the services/products and how successful this strategy was, based on the objectives. This type of strategy includes promotions or publicity on social media.
   a. Promoting products/services?

10) Business reorganization is a strategy that involves organizational changes to improve processes or reduce personnel costs. This category includes the change of personnel functions such as the transition from one department to another, in order to transfer employees where there is little work to where there is a shortage of labor. Another point that may be inherent to this strategy is the creation or removal of leadership positions depending on the organization's needs.
    a. Business reorganization?
11) Renegotiation with suppliers can have a significant impact on operational results. Some of the strategies inherent to this topic are the renegotiation of prices, delivery times or even the quality of products/services.
    a. Renegotiation with suppliers?

The answers given represent the level of application/importance of each of the strategies used by the organization, from 0 to 9, with zero for strategies not applied and 9 for strategies considered critical or extremely important.