Estimation of the Global Amount of Mandatory Investments for Distribution Network Expansion Planning

Pedro Miguel Macedo
Centre for Power and Energy Systems
INESC TEC
Porto, Portugal
pedro.m.macedo@inesctec.pt

José Nuno Fidalgo, João Tomé Saraiva
Centre for Power and Energy Systems of INESC TEC and
Department of Electrical Engineering and Computers
Faculty of Engineering of University of Porto
Porto, Portugal
jfidalgo@inesctec.pt, jsaraiva@fe.up.pt

***Abstract*— The financial planning of distribution systems usually includes the prediction of annual mandatory investments, concerning the resources that the DSO is compelled to allocate as a result of new network connections, required by new consumers or new energy producers. This paper presents a methodology to estimate the mandatory investments that the DSO should do in the distribution network. These estimations are based on historical data, load growth expectations and various socioeconomic indices. However, the available database contains very few annual investment examples (one aggregated value per year since 2002) compared to the large number of variables (potential inputs), which is a factor of regression overfitting. Thus, the applicable regression techniques are restrained to simple but efficient models. This paper describes a new methodology to identify the most suitable estimation models. The implemented application automatically builds, selects, and tests estimation models resulting from combinations of input variables. The final forecast is provided by a committee of models. Results obtained so far confirm the feasibility of the adopted methodology.**

Keywords*— *Mandatory Investments, Distribution Network Planning, Cost Estimation, Linear and Non-linear Regression.

I. NOMENCLATURE

CPI	Construction Production Index
DN	Distribution Network
DSO	Distribution System Operators
EC	Energy Consumption
EM	Estimation Model
GDP	Gross Domestic Product
HMVEC	High and Medium Voltage Energy Consumption
HV	High Voltage
IR	Inflation rate
LNOL	Total length of new overhead lines
LNUC	Total length of new underground cables
MAPE	Mean Absolute Percentage Error
MI	Mandatory investments
MV	Medium Voltage
NCP	Number of consumers and producers
NNC	Number of new constructions
NRB	Number of new residential buildings
PDIDN	Plan for Development and Investment in the Distribution Network
PISS	Power installed in secondary substations
RMSE	Root Mean Square Error
UR	Unemployment rate

II. INTRODUCTION

According to the current Portuguese legislation and regulations regarding the electricity sector (Decree-Law 215-A/2012) [1], Distribution System Operators (DSO) are required to prepare a Plan for Development and Investment in the Distribution Network (PDIDN) [2]. This plan should be designed for a time frame of five years, updated every two years and submitted in even years, based on the technical characteristics of the network and its current and predicted supply and demand [3].

In this plan, the DSO must specify the future investments to be made in the Distribution Network (DN). These investments have to be approved by the Regulatory Agency for the Energy Services and once approved and implemented over the years they have an impact on the determination of the regulated revenue of the Distribution activity to be recovered by the Tariffs for the Use of Distribution Networks. These tariffs are organized in HV, MV and LV terms, as a part of the Access Tariff and are paid by the end consumers according to their voltage connection level. This explains the relevance for regulatory and tariff setting purposes of the definition and approval of distribution network investments, as distribution network tariffs, for LV consumers, correspond to about 25% of final end-user tariffs and about 45% of corresponding access tariffs.

In terms of the distribution expansion planning procedure, a part of the network investments corresponds to Mandatory Investments (MI), that is, investments that DSO has to conduct resulting of the terms of the concession contract established with the Portuguese state and they include the installation of connections to new consumers or new distributed generation plants or even subsequent network reinforcements to assure the security of supply. In the case of a novel requested connection, the applicant (consumer or producer) is also compelled to pay part of the corresponding connection costs.

The main goal of this research work was to develop a methodology and a tool to estimate the MI. The expected MI amount is a function of historical data, load and consumers' number growth, and socioeconomic indicators such as the inflation rate, unemployment rate, construction production index, among others.

Given the large combination of potential inputs and pre-processing variants, a MATLAB application was developed to automate the estimations process. This program combines inputs and defines Estimation Models (EM), evaluates these

978-1-6654-3597-0/21/$31.00 ©2021 European Union

EM and selects the best performers to build a committee of predictors that lead to the final predictions.

Having in mind the objectives stated above, this paper is organized as follows. After this Introduction, Section III details the developed methodology, Section IV presents the results that were obtained using realistic data from a large Portuguese DSO and Section V includes the final conclusions and comments.

III. Methodology

The authors did not find scientific literature or project reports on research works similar to the one described in this paper. Therefore, it was necessary to test different alternatives, to understand the best way to define an effective methodology to address the problem of estimating the amount of Mandatory Investments to be made by a DSO. In the present study, a wide range of variables that could influence the MI was identified. Moreover, the number of samples is very limited (only sixteen, since for each year from 2002 to 2017 it was available the global amount of the MI). EM built with many parameters and a small number of examples usually show a reduced generalization capacity. Even though these EM allow a good adaptation to the historic data, they are likely to generate predictions with large volatility – overfitting [4]. That is the reason why we chose to define a methodology based on models with a small number of parameters.

Other factors of complexity of the estimation process to be conducted are as follows:

- Some amendments in the regulation directives occurred in the last decade, with impact on the conditions to establish new connections to the grids. In 2007 there was a change in the legislation [5] that establishes that the costs paid by the requesters would be based on the physical distance to the grid instead of the electrical distance. Hence, the costs supported by the requester decreased, making the connections' requests more appealing. However, in 2013, the regulation changed again [6], reverting to a situation similar to the rules before 2007;
- Strong changes in the Portuguese socioeconomic environment namely in view of the economic and financial crisis of 2010 -2015;
- Some investment peaks in the historical MI evolution, virtually unexplained by the input variables.

A. Data Selection

The first step of EM development started with the identification of factors that potentially affect MI, such as:

- The social-economic welfare (e.g. construction index, unemployment and inflation rates);
- DN status – capacity use (utilization index), demand and number of consumers;
- Local network conditions: population density, terrain orography, zone type (rural, industrial, etc.);
- Changes in legislation (e.g. creation of barriers or opportunities for the establishment of new connections).

The selection of the inputs aims at explaining and characterize the evolution of the mentioned factors, considering the credibility of available sources.

The available input data were organized as follows:

- DN status:
  - Total DN energy consumption (EC);
  - High and medium voltage energy consumption (HMVEC);
  - Number of consumers and producers (NCP);
- Socioeconomic welfare:
  - Gross domestic product (GDP);
  - Inflation rate (IR);
  - Unemployment rate (UR);
- Construction sector:
  - Construction production index (CPI);
  - Number of new residential buildings (NRB);
  - Number of new constructions (NNC);
- DN new assets:
  - Power installed in secondary substations (PISS);
  - Total length of new overhead lines (LNOL);
  - Total length of new underground cables (LNUC).

The data related to DN status (historical and projections) were provided by the DSO. The historical of social-economic indexes were extracted from the databases of the National Institute of Statistics - INE [7]. Their projections were obtained by the average projections of several entities, such as the Portuguese Federation of Construction, Bank of Portugal, International Monetary Fund, Ministry of Economy, and Organization for Economic Co-operation and Development, available in a document prepared by the Portuguese Public Finance Council [8]. Beyond the historical data-driven from these sources, the projections for the coming years are also important. In some cases, such as the projection of the CPI index, they were obtained using the estimation approach described in section D. Nevertheless, the first year of CPI projection was based on indications provided by the Portuguese Construction Federation, [9]. The historical data for DN assets were provided by the DSO, and its projections were estimated based on the CPI expected evolution (see section D).

Fig 1 shows the historical and projected data of the explanatory variables, provided by certified entities at the start date of the project (beginning of 2018).

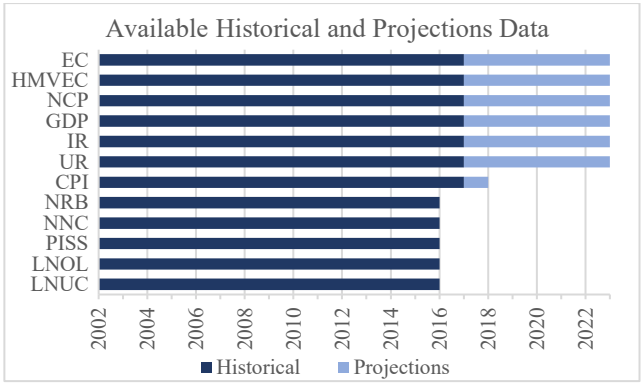

Fig. 1: Data availability of the explanatory variables.

It is fair to assume that the value of MI each year will depend generally on the inputs' values in the same year. As the goal is to provide MI estimations for a time horizon of five years ahead, it was necessary to obtain forecasts of these inputs up to 2020. For some variables, there are reliable and official sources that provided predictions for a few years. The variables with no available predictions are estimated similarly

to the procedure applied for the MI, as explained in sections D and E.

Fig. 2 characterizes the linear correlation (determination coefficient, R^2) between the inputs listed above and the MI [4].

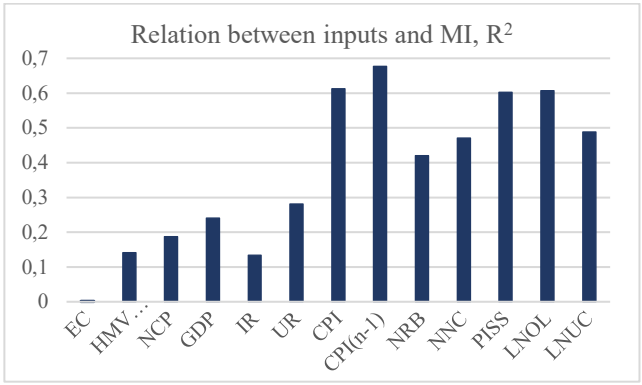

Fig. 2: Correlation, given by R^2, between the available inputs and MI.

This graph shows that the individual relationship between each of the inputs and the MI is relatively weak (<0.7). Nonetheless, it is important to note that this indicator only detects linear correlations, and therefore, the conclusion from this analysis must be conservative: if the coefficient is high, there is a relationship between the variables; however, if it is low, it is not possible to get any final indication. In the last case, the variables can be related by a non-linear function, or may be part of a complex function with other variables or may also not be related at all. Despite this uncertainty, section D describes how the grouping of variables allowed to construct EM with good performance. Given the larger correlation between MI and the CPI of the previous year, this variable is inputted with a shift of one year.

B. Data Pre-Processing

In this research, we considered several alternatives for inputting each variable into the models: historical value, normalization, annual variation and desensitization. The next paragraphs provide details on these alternatives.

Normalization – The Min-Max [4] was adopted to normalize the data, given that it allows setting an equal scale for all input variables. In an estimation process based on linear regressions, this operation is redundant, as it just involves parameters adaptation. However, the coefficient optimizing process is harder to work with if inputs have very different scales. Typically, the defined interval for this type of approaches is [-1 to 1]. However, in this work, the chosen range was [1 to 3], due to the difficulty of some regression processes in treating negative variables or variables close to 0 (e.g. logarithmic regression).

Annual variation – it intends to estimate the outputs based on the inputs' variations from one year to another. This approach, even if in some cases allows a good adaptation to the historical data, sometimes results in very volatile estimates. This was the reason why it was only applied for the estimation of some intercalary variables (e.g. new assets in the DN) - since one intends to obtain projections within a trend without large oscillations over the years.

Desensitization – the goal is to soften the importance of some input variables, making the output less dependent (less sensitive) to their variations. The application of this technique assumes that some explanatory variables occasionally exhibit inexplicable or inconsistent oscillations when compared to the output or with other inputs. Such a technique involves replicating each example twice, resulting in three patterns: the original, the input negatively and positively affected by a small amount. The variation of the input is determined by a desensitization coefficient which, in this case, was based on the coefficient of variation or relative standard deviation [4]. However, it was decided to consider only one-half and one-third of this coefficient, so as not to over-soften the values and thereby pervert the available information.

Fig. 3 shows three estimation approaches referring to EM containing the same variables, but for different pre-processing treatments: normalized values, variation and desensitization.

The variable LNOL was chosen to illustrate these procedures because of its volatility and difficulty in identifying an EM that matches the original historical values. In fact, the historical values exhibit peaks with no clear interpretation. Besides, no strong relation was found of this behaviour with the other input variables. However, as it is observed, the desensitization of the inputs allowed to soften the influence of these peaks during the training process.

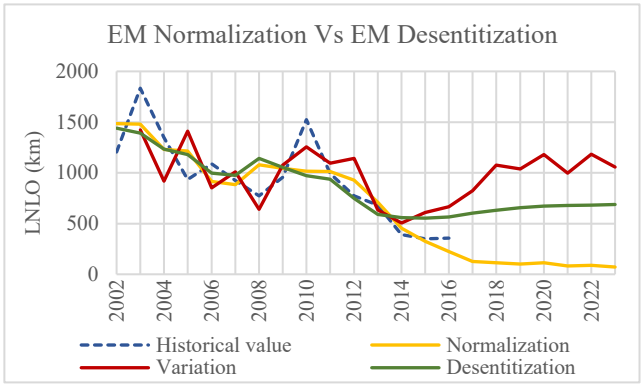

Fig. 3: Estimates of LNOL obtained by EM (UR, NRB, NNC) with normalization, variation and desensitization.

In the EM case based on data variation, the estimates are rather unstable (high peaks; high standard deviation). It is important to note that the EM with a high standard deviation will be discarded by the filtering process described in Section E. Note that, in the case of normalized values, the curve decreases to a level that is also discarded by the filtering process. In summary, each input variable might be inputted to the EM by four different processes. The particular process to be adopted for each variable will be selected according to its contribution to the EM performance.

C. Accuracy Measures

As a way to evaluate the EM, it was decided to use just one or two cases (years) for testing purposes and to keep all the rest for training. This split of data was adopted due to the low number of historical examples, aiming at providing the maximum information possible to determine the coefficients of the model. Due to the same fact, a weighted mean of the error measures was performed, between the training error, with 70%, and the test error, with 30%. The error measures used for performance diagnosis were the Mean Absolute Percentage Error (MAPE) [4] and a normalization of the Root Mean Square Error (RMSE) [4], designated by NRMSE, which consists of dividing the RMSE by the mean of the historical values.

As mentioned, the goal of this work is to forecast the MI for the next five years with a special focus for the first two

years of this period. As so, the performance diagnosis is complemented by the analysis of the estimates for this time horizon, namely by checking if the forecasts are compatible with the standards of each variable. In effect, it was observed that some EM were well adapted to the training and test periods (low errors), but the predictions for the five years were unfeasible (e.g. negative MI). These EM were discarded.

D. Estimation Processes

As mentioned before, there were no official projections for 2017-2023 for the explanatory variables concerning the construction sector evolution and the new assets in the DN. These will be estimated using the relations with variables for which projections are already available.

The order of variables to be estimated is dictated by the highest correlation (R^2) with those already available, as shown in Fig. 4. Based on this assumption and on the fact that CPI already has one more historical year and the indication of a year of projection, the estimation of this indicator was the first to be carried out.

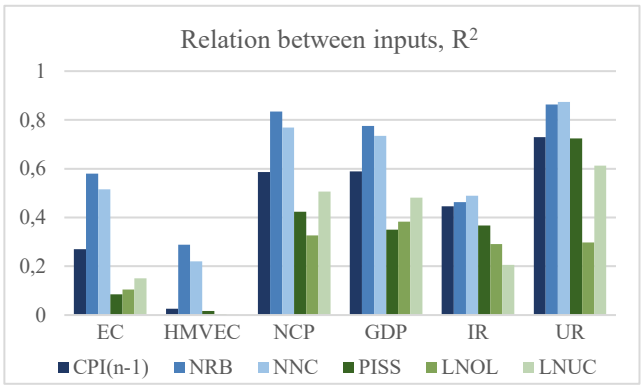

Fig. 4: Correlation, given by R^2, between the inputs

Several approaches were tested to find combinations of variables that produce performant EM. This study involved testing as inputs the value of the variable itself in the last year, as well as other regressions models such as the exponential, logarithmic, 2nd degree polynomial and power types.

However, it was found that the inclusion of data for the previous year did not improve the estimation performance. On the contrary, it was concluded that delayed data tends to excessively reinforce the historical trend and as disregard the effects of the current year factors.

The linear regression showed a good adaptation to the real curve, superior to the other types, except for the polynomial. Despite that, this regression requires a larger number of model parameters, which can impair the generalization performance. Therefore, it was decided to use a linear regression, due to its simplicity, its good performance and lower parameter requirements. The process for obtaining the final estimates of the CPI followed the approach described in Section E.

Regarding NRB and NNC estimations, it was found these variables are quite related to CPI. Thus, several tests were developed to estimate NRB and NNC as a function of CPI. In this phase, we consider the following regression types: linear, exponential, logarithmic, polynomial, and potential. Again, the unfeasible solutions (e.g. negative values) were discarded. The N feasible solutions were then combined (average) in groups from 2 to N, and the case with the lowest NRMSE was selected.

The integration of new assets in the DN comes from the need to connect clients and/or from the need to increase the grid power flow capacity. These actions are mostly associated with network reinforcement and expansion. The historic data shows these variables (LNOL, LNUC and PISS) have big volatility. Moreover, the correlation between these variables and the other available variables is generally weak as displayed in Fig. 4. Therefore, as in the estimation of CPI, it was decided to adopt the process described in Section E. It should be noted that the bank of variables available to create the EM is added to the indicators regarding the civil construction indices, since this activity displays a considerable relation of causality to the new distribution assets. Having acquired all estimates of these intermediate variables, MI estimation is conducted according to the approach outlined in Section E.

Although the individual correlation between inputs and MI is relatively weak (as illustrated in the graph of Fig. 2), it is noticeable that the grouping sets of variables allows developing estimation models with good performance.

The EM were defined using just one or two samples of the last years for testing. The final estimates attained by the average of the 10 best EM of each case. Fig. 5 illustrates the attained results. In this figure and along this paper, to preserve data confidentiality, a linear transformation was applied to the real MI values provided by the DSO.

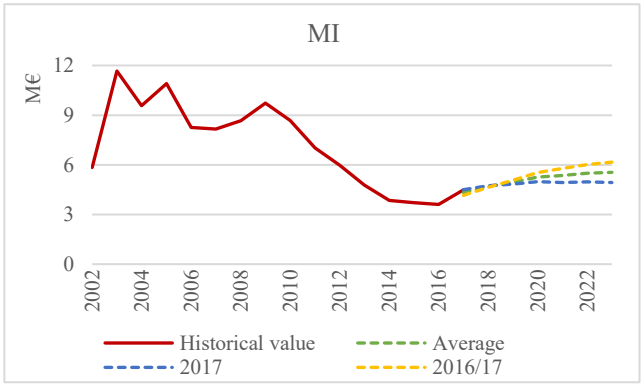

Fig. 5: Evolution of the estimates that consider 2017 and 2016/17 as a test and its average.

E. Automation of the EM Estimation

The high number of possible EM that could result from combinations of the available input variables and different types of data pre-processing/normalization strategies led to the development of a MATLAB application to automate the process. This application considers all possible combinations of simple EM (up to 4 variables), evaluates their response, and selects the best 10 EM. The final estimation will result from the collective output of this ensemble of estimation models.

The flowchart on Fig. 6 describes the process sequence as deployed by the developed application. The application starts by importing the previously prepared input and target data. Next, it successively creates EM, determines its parameters, performs estimations, evaluates each model and records the results into a database. This stage is repeated until all combinations of 2, 3, and 4 input variables are tested.

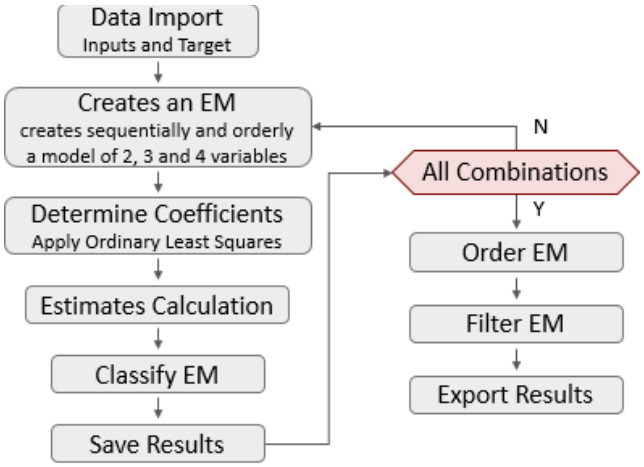

Fig. 6: Flowchart of the developed algorithm.

Afterwards, the tested EM are ranked by merit, according to calculated error measures (MAPE, RMSE) for the training and test years [4]. Then, a filter is used to eliminate the EM that present values outside a user-defined limit (margin defined around the average MI of the last years and the coefficient of variation). The goal is to eliminate EM that contain projections outside the defined threshold - even if they have a small training and testing error up to 2017.

Finally, the algorithm selects the ten best ME, computes the ensemble response, and calculates the performance indicators.

IV. Results

The results presented in this section were obtained by INESC TEC, in the scope of a collaboration project with EDP Distribuição, the largest Portuguese DSO.

As mentioning before, the final MI estimates were based on two cases, considering one or two years for testing as shown in Fig. 5. In this study, we chose to consider both cases, once there was a significant increase in MI from 2016 to 2017. The MI in 2017 was clearly above the expected, once this is not justified by the generality of the explanatory variables/inputs, that in this year show a lower variation. This was probably caused by the high number of new network connections requested by distributed generation promoters.

The following two figures, Fig. 7 and Fig. 8, show the top 10 performing EM obtained with one year for testing (Case 1) and two years for testing (Case 2), respectively.

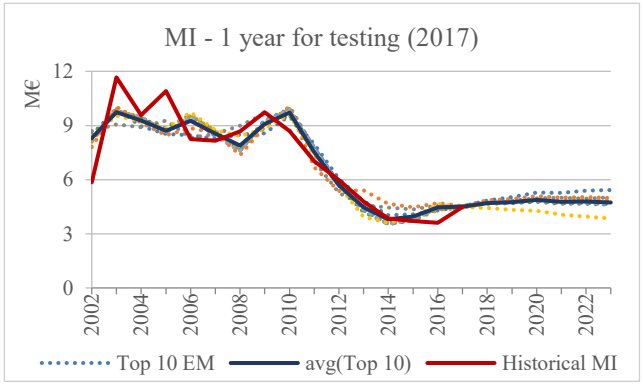

Fig. 7: Case 1 - historical MI and estimations attained by the top 10 EM.

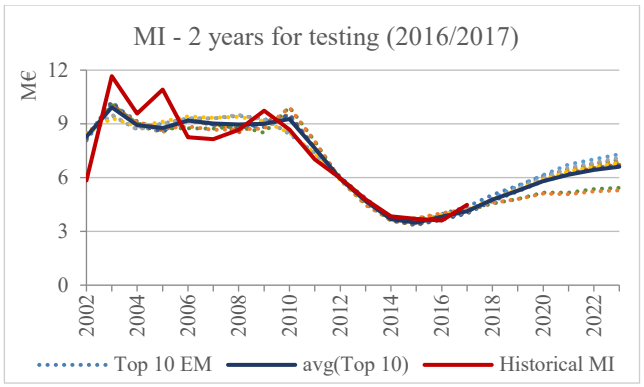

Fig. 8: Case 2 - historical MI and estimations attained by the top 10 EM.

Fig. 7 illustrates the difficulty that the 10 best performing EM have in following the increase of MI from 2016 to 2017, revealing more conservative estimates. The EM presented in Fig. 8, which considers the component of test error in the two years, are able to better follow this increase, continuing the growth trend.

These two figures also show that the amount of MI tends to increase after 2017, reflecting the end of the impacts of the economic crisis that occurred in Portugal from 2010 to 2015.

This crisis started in 2010 and impacted the amount of the MI through its continuous decrease, nearly up to 2014 when it tended to stabilize, only inverting its behaviour in 2016. Similarly, it was expected that coming out of this crisis would be gradual and moderate – yet a slower recovery after the downfall. Also noteworthy, the evolution of the MI is according to the predictions for the social-economic indexes.

The MAPE error for the final estimates of the MI for training, test, as well as the weighted mean, were 9.0%, 7.0% and 8.4% respectively. It should be noted that the weighted average of the errors returned by the estimates obtained by the set of the ten best EM is lower than the error obtained by the EM with better performance.

The final part of this research aimed at assessing which variables are more relevant to the MI estimation process. Fig. 9 shows the number of cases that the inputs were requested by the ten selected EM and their influence in the final estimates of the amount of MI.

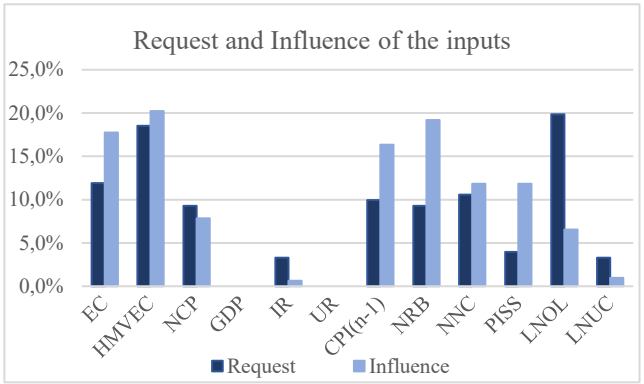

Fig. 9: Request and influence the inputs in the final estimates of MI.

This figure shows that all the inputs were requested, confirming the importance and the correlation with the MI (even if it was quite weak in some cases). It can be observed that the most influential variables are the energy consumed

(EC and HMVEC), $CPI_{(n-1)}$ and NRB. Moreover, it is important to mention that the energy consumption and GDP, as well as CPI, NNC and NRB, have similar historic evolutions, meaning that strong correlations exist among them.

The presence of correlations among several input variables explains the differences between these results and the ones of Fig. 2. On one side, in Fig. 9 the results concern the contextual importance i.e. the global weight of each variable in the multi-linear regression formulas. On the other hand, in Fig 2 the correlation is assessed on an individual basis and just reflecting historical data. For the final ensemble of multi-linear formulas, some of the inputs have strong inter-correlations, signifying that the individual importance of a given variable can be spread by several inter-correlated inputs.

This study had its first edition in 2015, providing estimates for the MI between 2015 and 2021 to the DSO. Therefore, it was possible to evaluate the real error obtained in the previous edition in the years 2015 to 2017 - which was, respectively, 1.6%, 0.7% and 16.9%. The small estimation deviations testify the effectiveness of the proposed methodology. However, Fig. 6 shows that the relatively higher error in 2017 may be associated with an unusual rise in the investments, which was not reflected so rapidly in the available inputs after the end of the economic crisis (2015).

V. Conclusions

This paper proposed and described a new approach to estimate long-term mandatory investments in distribution networks, able to deal with data scarcity and the complexity of the process. This work was developed in close connection and as a request of the largest Portuguese DSO that provided real data regarding network investments and values for other indices that were used in the estimation process. It should be mentioned that the estimation of the Mandatory Investments to be conducted by the DSO is a very relevant step when preparing the 5 years expansion plan of the distribution network and it is also important for tariff purposes because as grid investments take place several variables describing the behaviour of the distribution network are updated and are used by the Regulatory Agency for the Energy Services to set the regulated revenue of the Distribution Activity that ultimately leads to the setting of the corresponding Tariffs for the Use of Distribution Networks to be paid by end consumers according to their voltage connection level. In this sense, it is important to mention that this project took place in 2015 and the resulting models and estimates where used by the DSO when preparing subsequent expansion plans.

Finally, the next paragraphs detail some relevant pieces of evidence inferred from the analysis of the results and the experience gained by the authors during the development of this project:

- The estimates of the MI are aligned with the predictions available for the Portuguese economy;
- The use of inputs with low correlation with the output does not prevent the possibility of obtaining EM capable of generating estimates with a relatively low error;
- There is a high consistency among the estimates attained from the pool of selected models;
- The use of the annual variation instead of the variable itself only compensates for variables with a very stable trend;
- The developed methodology includes an input desensitization module, in order to limit the influence of the volatility of some input variables;
- The application is able to diagnose and register all the defined EM, resulting from the input variables combinations, in under one minute, in a PC with an Intel i7 4710HQ CPU @2.5 GHz and 8GB RAM.
- The estimated MI were very closed to the actual values, despite the large uncertainty usually found in long-term network planning. It is important to emphasize that this plan is updated every two years, the reason why is given a stronger relevance to the estimates for the first two years.
- The development of the methodology and the corresponding application was time-consuming, due to the requirement of exploring several alternatives and techniques, to identify the ones that best meet the purpose of the study. However, the use of the application in new studies is quite efficient and able to provide useful results very rapidly.

Acknowledgment

This project was financed by the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, through national funds, and co-funded by the FEDER, where applicable.

References

[1] Decree-Law nr. 215-A/2012, from 8 of October. Republic Diary, nr. 194, Supplement, Series I. Ministry of Economy and Employment. Lisbon.

[2] EDP Distribuição, "PDIRD-E 2018 Plan for Development and Investment in the Distribution Network", July version, July 2018

[3] Decree-Law nr. 215-B/2012, from 8 of October. Republic Diary, nr. 194, Supplement, Series I. Ministry of Economy and Employment. Lisbon.

[4] S. Makridakis, S. C. Wheelwright and R. J. Hyndman, "Forecasting methods and applications", Third Edition, John Wiley Sons, 1998.

[5] ERSE - Regulatory Agency for the Energy Services. Order No. 12 741/2007, on June 21 of 2007. Republic Diary, 2nd series — No. 118. ERSE, 2007.

[6] ERSE - Regulatory Agency for the Energy Services. Order No. 18/2012, on November 8 of 2012. Republic Diary, 2nd serie — N.º 216. ERSE, 2012

[7] National Institute of Statistic. Statistics Portugal. [http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_base_dados &contexto=bd&selTab=tab2, accessed between January and June 2018]

[8] Public Finance Council. *Summary of Macroeconomic Projections for the Portuguese Economy*. Last update: 28 November 2017. [http://www.cfp.pt/public-finances-information/proje%C3%A7%C3%B5es-macroecon%C3%B3micas/]

[9] Portuguese Federation of the Construction and Public Works Industry. Press releases [http://www.fepicop.pt/index.php?id=22, 2018]