ANALES de Edificación

Received: 25/03/2023
Accepted:10/04/2023

Anales de Edificación
Vol. 9, Nº1, 33-37 (2023)
ISSN: 2444-1309
DOI: 10.20868/ade.2024.5307

Aplicabilidad de la metodología de reducción de la dimensionalidad multifactorial al análisis de variables en edificios sanitarios
Applicability of the multifactor dimensionality reduction methodology to the analysis of variables in sanitary buildings.

Alejandro Prieto-Fernández[a]; Álvaro Carmona-Baltasar[a]; Jaime González-Domínguez[a]; Manuel Botejara-Antúnez[a]; Gonzalo Sánchez-Barroso[a]; Justo García-Sanz-Calcedo[a]

[a] Departamento de Expresión Gráfica, Escuela de Ingenierías Industriales, Universidad de Extremadura, 06006 Badajoz (España) aprietofc@alumnos.unex.es; acarmonaj@alumnos.unex.es: jaimegd@unex.es; manuelba@unex.es; gsm@unex.es; jgsanz@unex.es.

Resumen-- **Para optimizar las actividades ingenieriles es necesario realizar un análisis de una gran cantidad de datos y variables. El objetivo es implementar MDR para abordar mejor el estudio y proponer mejoras para reducir el consumo energético en los centros de salud extremeños. Estas numerosas variables no tienen unas interacciones directas y cuantificables sobre el consumo energético. Para solventar este inconveniente es posible aplicar el método de reducción de dimensionalidad multifactorial (MDR). MDR emplea Machine Learning para buscar las mejores combinaciones entre las variables. Esto permite crear un modelo que simplifica el análisis de los datos estudiados. De todo el conjunto se selecciona la combinación de variables que mejor describe el estudio, agrupando en alto o bajo riesgo. Se ha comprobado que de esta forma es posible comprender mejor y optimizar las actividades ingenieriles. MDR puede ser empleado en numerosos análisis ingenieriles: consumo energético, mantenimiento de equipos, generación de residuos, etc.**

Palabras clave— Reducción de la dimensionalidad multifactorial; Ingeniería; Construcción sanitaria; Aprendizaje automático; Minería de datos.

Abstract— **In order to optimise engineering activities, it is necessary to analyse a large amount of data and variables. The objective is to implement MDR to better address the study and propose improvements to reduce energy consumption in Extremadura's health centres. These numerous variables do not have direct and quantifiable interactions on energy consumption. To overcome this drawback, it is possible to apply the multifactor dimensionality reduction (MDR) method. MDR uses Machine Learning to search for the best combinations of variables. This makes it possible to create a model that simplifies the analysis of the data studied. From the whole set, the combination of variables that best describes the study is selected, grouping them into high or low risk. It has been proven that in this way it is possible to better understand and optimise engineering activities. MDR can be used in numerous engineering analyses: energy consumption, equipment maintenance, waste generation, etc.**

Index Terms— Multifactor Dimensionality Reduction; Engineering; Healthcare Building; Machine Learning; Data Mining..

I. Introduction

ON On many occasions, situations arise where decision-making plays a key role during the design phase of healthcare buildings. A large number of variables need to be considered (González-Domínguez et al., 2022), which makes it difficult to obtain a clear and concise result. An analysis of the variables that best describe the excess energy consumption of health centres in Extremadura has been carried out in order to better address the study and propose improvements for lower energy consumption.

The MDR (Multifactor Dimensionality Reduction) methodology was developed with the aim of being applied to the field of genetics, for the study of different diseases and the response of the organism to different drugs (Motsinger et al., 2006). This method makes use of Machine Learning to identify the best model that presents the maximum coherence with the data studied. MDR has been developed as a non-parametric method without a predefined model, i.e., the results obtained depend solely on the data analysed (Ritchie et al., 2003). As a result, the combination of variables that best describe the problem is obtained. In addition, a categorisation into high risk or low risk of excess energy consumption is obtained according to the number of health centres that exceed or do not exceed the average consumption of the data studied.

This methodology was originally used to determine cancer risk based on the interaction between genes and environmental factors (Ritchie et al., 2001). This methodology has been tested in many research studies in the field of genetics.

This methodology has also been used to determine gene-to-gene interactions that may be associated with the risk of type 2 diabetes. Using this methodology, it was possible to show the interaction of two loci out of 23 candidate loci (Cho et al., 2004). MDR was also used to detect susceptibility to pulmonary tuberculosis. Through the application of this methodology a strong interaction between 3 polymorphisms out of the initial 19 was detected (Collins et al., 2013). In studies of diseases such as autism, which have a strong genetic dependence, this methodology has been used to detect risk based on gene-gene interaction (Ma et al., 2005).

Over time, a number of modifications and extensions of the original methodology have been developed depending on the characteristics of the starting data and applications of the methodology (Gola et al., 2016).

An extension of this methodology is GMDR (Generalized Multifactor Dimensionality Reduction) which allows the adjustment of discrete and quantitative covariates and the application to dichotomous and continuous phenotypes. This extension has been used to study nicotine dependence (Lou et al., 2007).

Another possible extension is the implementation of Cox-MDR (MDR-based Cox analysis) for the detection of gene-gene interactions in leukaemia patients. This extension allows for better covariate adjustment (Lee et al., 2012).

The aim of this work is the implementation of MDR to evaluate the variables that best describe the excess energy consumption in health centres in Extremadura, in order to improve the energy management of the building and propose an action plan to optimise energy consumption. In the same way, MDR can be applied to other engineering activities that are affected by many variables in order to address the study (González-Domínguez et al., 2021; Sánchez-Barroso et al., 2021).

II. Methodology

MDR was used to analyse the energy consumption of health centres in Extremadura (Spain), because there are many variables that affect energy consumption and as many dimensions as there are variables (Martínez de Salazar et al., 2019). These variables can be the surface area, locality, year of construction, population in the locality, basic health area, power of the air conditioning equipment, etc. All the variables considered for this analysis have an interaction with energy consumption, however, this interaction is neither direct nor linear with energy consumption in health centres (González González et al., 2018).

As the direct relationship between these variables and between these variables and energy consumption was not known, the importance of each of them and the real effect on the analysis is unknown. This makes it difficult to study and propose improvements to reduce energy consumption. Therefore, it has been decided to use MDR for the data analysis.

The use of MDR in this analysis has made it possible to reduce the problem variables used to generate the most accurate model that best describes the study. The model provides a classification of all data into high or low risk of excess energy consumption. For the use of MDR it is necessary to dichotomise the variables, for this purpose, groups were made for each of the variables in the most convenient way for the study. The energy consumption of each health centre was used as a control variable. Once the variables have been dichotomised, MDR can be applied to the data.

In this way, with two variables affecting the analysis, with three groups of each variable, nine combinations are generated, and for each of these nine combinations a set of health centres that exceed their energy consumption and a set that do not exceed their energy consumption are obtained. Thus, these combinations can be classified as high risk in terms of excess energy consumption if the ratio between those that exceed the average consumption and those that do not exceeds a threshold. They would be low risk in excess energy consumption if this threshold is not exceeded. For this research, a unit-valued threshold has been considered, so that if the number of health facilities with higher-than-average energy consumption exceeds the number of those that do not exceed the average, that combination will be high risk and vice versa. Figure 1 shows an example of the combination of two variables, X10 and X7, with those at high risk represented in dark and those at low risk in light.

In the same way, the analysis has been carried out to quantify the difference in average consumption of the health centres with respect to the average. In this case, for each combination, the average energy consumption of the health centres in this

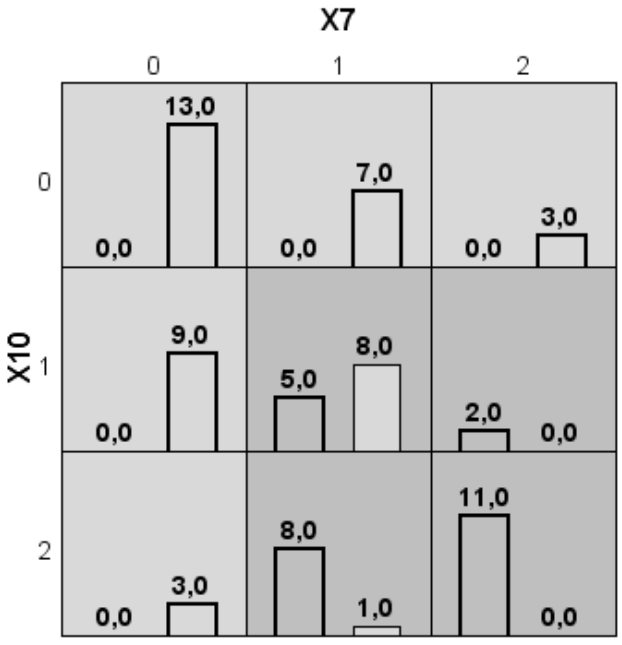

Fig. 1. Example of combination of variables with count

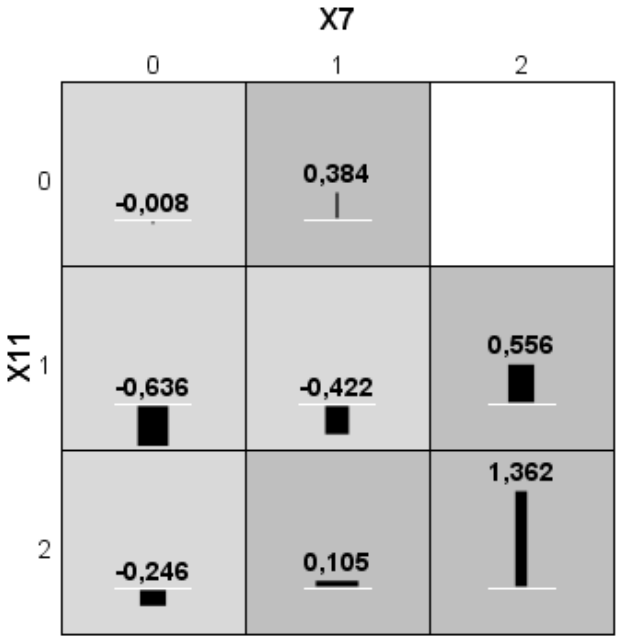

Fig. 2. Example of combination of variables with difference from the mean

combination is calculated and the average consumption of all the health centres is subtracted. The combinations are then classified according to whether the difference is negative or positive as low or high risk in excess energy consumption, respectively. An example can be seen in Figure 2 with the high-risk groups with a positive difference and the low-risk groups with a negative difference, in white the groups with no data.

MDR thus produces a reduction of the dimensionality of the n-dimensional analysis to a high or low risk dimension, thus facilitating the case study. This one-dimensional variable is used for training and testing the Machine Learning algorithm, allowing the selection of the best combination of variables. Cross-validation is used to find the best combination of variables in a supervised learning situation, i.e., the inputs and outputs of the data set are known. With the inputs, the values of the variables, the algorithm learns to predict the outputs, the control variable. To select the best combination of variables, the balanced accuracy is used as an indicator, which is calculated for all combinations and MDR keeps the best combination of them.

The application of cross-validation divides the data set into ten parts, generally, to use nine of these parts for the training set of the algorithm and the remaining part for the testing and validation set (Motsinger et al., 2006). The algorithm is run ten times, using each of the ten parts as the validation set, and the cross-validation is determined. The cross-validation shows in how many of the ten times the algorithm was run the combination selected as the most consistent and best describing the dataset. Thus, it is performed with all combinations of variables until the one that best represents the analysis dataset is obtained. Subsequently, a permutation can be applied to find a p-value to test whether the model is statistically significant (Dai et al., 2012).

III. Results

The use of this methodology has made it possible to determine which combination of variables is the most representative of the entire data set that was initially available, and which groups are at high or low risk of exceeding energy consumption within these variables. This considerably reduces the dimensions of the problem, going from a problem with many variables and no determined interactions between them and the control variable, to a problem with defined results by limiting the number of variables and only one dimension, whether or not energy consumption is exceeded.

Results are obtained for combinations of different number of variables and, according to their cross-validation and balanced accuracy, it is selected which of them is the most consistent for the analysis of the problem (Coffey et al., 2004).

In this way, it is possible to perform a more effective analysis using the variables obtained as a result, it is also possible to tailor the projection of new health centres within the groups with the lowest risk of excess energy.

In the phase of defining the input variables, a careful choice must be made, as inconsistent results may be obtained. It is also necessary to have a large number of observations in order to obtain a consistent and statistically significant result, so that this result can be used to analyse the problem and propose improvements.

When analysing many variables and a large sample of data, MDR can require high computational power from computer hardware (Hahn et al., 2003). This is due to the large number of combinations it performs.

Using combinations of many variables, more than three, it is very likely to have groups without data as in the case of Figure 3. Due to the large number of possible combinations, data can be included in certain groups and not in others, leaving empty

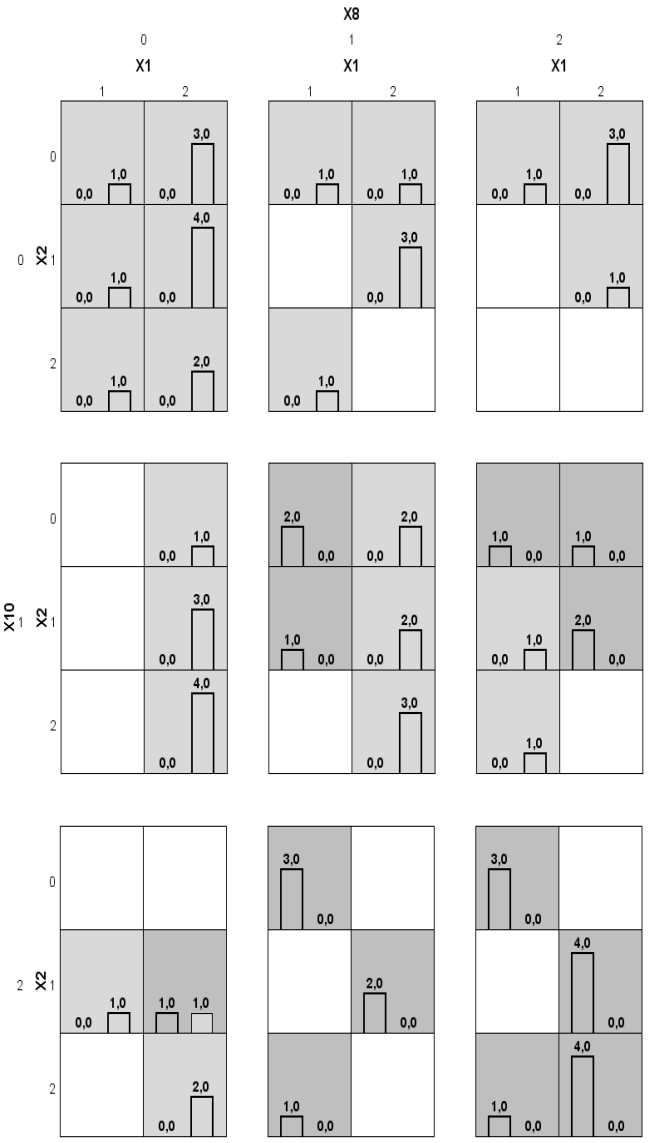

Fig. 3. Example of combination of four variables with missing data

combinations of data. This means that the model does not have sufficient significance and that future data cannot be analysed to fit these data combinations, making it necessary to study all of its variables.

The result of MDR is not always easy to interpret; when combinations of four or more variables are used to model the data, it is not easy to find patterns of low or high risk.

A. Results discussion

The main advantage is that it allows a better analysis of all multivariate data by reducing the number of variables describing the problem. This is achieved with different combinations of variables and their classification as high or low risk. For the simulated data, the consistency of the cross-validation is maximised and the average prediction error is minimised (Winham et al., 2010).

Another advantage of using MDR is that it has been developed as a non-parametric method and without a predefined model, i.e., the results obtained will depend only on the data analysed. For example, no regression of the dataset is performed, and no expert judgement is needed for data analysis. As there is no predefined model, some variables are not valued more highly than others.

Finally, the optimal combination of variables is selected objectively due to the use of cross-validation. For this selection, only the model's ability to predict high or low risk, the balanced accuracy and the consistency of the cross-validation are considered.

In the future, this methodology can be used for decision making in other engineering fields, such as equipment maintenance analysis, or waste generation analysis (García-Sanz-Calcedo et al., 2017).

IV. Conclusions

Thanks to the use of this methodology, it has been possible to reduce the number of variables that affect the study so that the data can be analysed to propose improvements. In this way, much more effective analyses can be carried out, thus avoiding working with very complex systems that make it difficult to obtain a result due to the large number of variables involved. Furthermore, it is possible to develop improvements that reduce energy consumption, improve equipment maintenance or reduce waste generation.

Unlike other methodologies, MDR is objective thanks to the use of Machine Learning with cross-validation, which allows defining which cases are high or low risk by using a smaller number of variables. By classifying them into high or low risk groups, it is possible to propose improvements aimed at those that are high risk. In short, the application of MDR is appropriate for data analysis in engineering.

Acknowledgements

This research has been supported by the European Regional Development Fund through the Research Project GR21098 linked to the VI Regional Plan for Research, Technical Development and Innovation of the Regional Government of Extremadura (2022).

References

Y.M. Cho, M.D. Ritchie, J.H. Moore, J.Y. Park, K.-U. Lee, H.D. Shin, H.K. Lee, K.S. Park, Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus, Diabetologia. 47 (2004) 549–554. https://doi.org/10.1007/s00125-003-1321-3.

C.S. Coffey, P.R. Hebert, M.D. Ritchie, H.M. Krumholz, J.M. Gaziano, P.M. Ridker, N.J. Brown, D.E. Vaughan, J.H. Moore, An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation., BMC Bioinformatics. 5 (2004) 49. https://doi.org/10.1186/1471-2105-5-49.

R.L. Collins, T. Hu, C. Wejse, G. Sirugo, S.M. Williams, J.H. Moore, Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis, BioData Min. 6 (2013) 4. https://doi.org/10.1186/1756-0381-6-4.

H. Dai, M. Bhandary, M. Becker, J.S. Leeder, R. Gaedigk, A.A. Motsinger-Reif, Global tests of P-values for multifactor dimensionality reduction models in selection of optimal number of target genes, BioData Min. 5 (2012) 3. https://doi.org/10.1186/1756-0381-5-3.

J. García-Sanz-Calcedo, M. Gómez-Chaparro, Quantitative analysis of the impact of maintenance management on the energy consumption of a hospital in Extremadura (Spain), Sustain. Cities Soc. 30 (2017) 217–222. https://doi.org/10.1016/j.scs.2017.01.019.

D. Gola, J.M. Mahachie John, K. van Steen, I.R. König, A roadmap to multifactor dimensionality reduction methods, Brief. Bioinform. 17 (2016) 293–308. https://doi.org/10.1093/bib/bbv038.

J. González-Domínguez, G. Sánchez-Barroso, J. García-Sanz-Calcedo, N. de Sousa Neves, Cox proportional hazards model used for predictive analysis of the energy consumption of healthcare buildings, Energy Build. 257 (2022) 111784. https://doi.org/10.1016/j.enbuild.2021.111784.

J. González-Domínguez, G. Sánchez-Barroso, J. García-Sanz-Calcedo, M. Sokol, Condition-based maintenance of ceramic curved tiles roof in Primary Healthcare buildings using Markov chains, J. Build. Eng. 43 (2021) 102517. https://doi.org/10.1016/j.jobe.2021.102517.

A. González González, J. García-Sanz-Calcedo, D.R. Salgado, A quantitative analysis of final energy consumption in hospitals in Spain, Sustain. Cities Soc. 36 (2018) 169–175. https://doi.org/10.1016/j.scs.2017.10.029.

L.W. Hahn, M.D. Ritchie, J.H. Moore, Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions, Bioinformatics. 19 (2003) 376–382. https://doi.org/10.1093/bioinformatics/btf869.

S. Lee, M.-S. Kwon, J.M. Oh, T. Park, Gene-gene interaction analysis for the survival phenotype based on the Cox model, Bioinformatics. 28 (2012) i582–i588. https://doi.org/10.1093/bioinformatics/bts415.

X.-Y. Lou, G.-B. Chen, L. Yan, J.Z. Ma, J. Zhu, R.C. Elston, M.D. Li, A Generalized Combinatorial Approach for Detecting Gene-by-Gene and Gene-by-Environment Interactions with Application to Nicotine Dependence, Am. J. Hum. Genet. 80 (2007) 1125–1137. https://doi.org/10.1086/518312.

D.Q. Ma, P.L. Whitehead, M.M. Menold, E.R. Martin, A.E. Ashley-Koch, H. Mei, M.D. Ritchie, G.R. DeLong, R.K. Abramson, H.H. Wright, M.L. Cuccaro, J.P. Hussman, J.R. Gilbert, M.A. Pericak-Vance, Identification of Significant Association and Gene-Gene Interaction of GABA Receptor Subunit Genes in Autism, Am. J. Hum. Genet. 77 (2005) 377–388. https://doi.org/10.1086/433195.

E. Martinez de Salazar, J. García Sanz-Calcedo, Study on the influence of maintenance operations on energy consumption and emissions in healthcare centres by fuzzy cognitive maps, J. Build. Perform. Simul. 12 (2019) 420–432. https://doi.org/10.1080/19401493.2018.1543351.

A.A. Motsinger, M.D. Ritchie, Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene - gene interactions in human genetics and pharmacogenomics studies, Hum. Genomics. 2 (2006) 318. https://doi.org/10.1186/1479-7364-2-5-318.

A.A. Motsinger, M.D. Ritchie, The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction, Genet. Epidemiol. 30 (2006) 546–555. https://doi.org/10.1002/gepi.20166.

M.D. Ritchie, L.W. Hahn, J.H. Moore, Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity, Genet. Epidemiol. 24 (2003) 150–157. https://doi.org/10.1002/gepi.10218.

M.D. Ritchie, L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, J.H. Moore, Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer, Am. J. Hum. Genet. 69 (2001) 138–147. https://doi.org/10.1086/321276.

G. Sánchez-Barroso, J. González-Domínguez, J. García-Sanz-Calcedo, Impact of urban mobility on carbon footprint in healthcare centers in Extremadura (Spain), Int. J. Sustain. Transp. (2021) 1–18. https://doi.org/10.1080/15568318.2021.1914794.

S.J. Winham, A.J. Slater, A.A. Motsinger-Reif, A comparison of internal validation techniques for multifactor dimensionality reduction, BMC Bioinformatics. 11 (2010) 394. https://doi.org/10.1186/1471-2105-11-394.