KnE Engineering

STARTCON19
STARTCON19 - International Doctorate Students Conference + Lab Workshop in Civil Engineering
Volume 2020

Knowledge E
enriching | engaging | empowering

Conference Paper

Multivariate Data Analysis of the Thermal Performance of Portuguese Residential Building Stock

Rita Andrade Santos[1], Inês Flores-Colen[1], Nuno Vieira Simões[2], and José Dinis Silvestre[1]

[1]CERIS-ICIST, DECivil, Instituto Superior Técnico (IST), Universidade de Lisboa (UL), Portugal
[2]ItеCons / ADAI – LAETA, Depart. of Civil Engineering, FCTUC, Universidade de Coimbra (UC), Portugal.

Abstract

The purpose of this study is to explore the relationship between solar orientation, age (constructive characterization) and energy performance of Portuguese residential building stock and to assess the usefulness of exploring the Portuguese National System for Energy and Indoor Air Quality Certification of Buildings (SCE) database through multivariate analysis techniques. By using principal components technique, it was possible to condense the residential units' features to only four principal components (PC): *solar orientation*; *constructive characterization; geometry* and *energy performance*, making information more workable. Grouping the entities into Clusters with *favourable and unfavourable* solar orientation and *old buildings* allowed to dilute the particularities of each entity, facilitating the interpretation of the data through generalization. A regression model was generated in order to explore/confirm which factors influence summer comfort the most. Using this approach, it was illustrated that the exploration of the SCE database through multivariate data analyses has an enormous potential to convert data into knowledge.

Corresponding Author:
Rita Andrade Santos
andrade.santos@tecnico.ulisboa.pt

Received: 7 January 2020
Accepted: 21 April 2020
Published: 3 May 2020

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the STARTCON19 Conference Committee.

1. Introduction

The majority of the European building stock, which was built before the implementation of energy codes, is a major contributor to CO_2 emissions and climate change. Its thermal retrofitting is essential. Therefore, knowledge concerning the current thermal performance of the building stock becomes of paramount importance.

The purpose of this study is to characterize the Portuguese building stock in order to outline strategies to improve the thermal performance of existing building envelopes.

This work was done in the scope of a PhD Thesis under progress concerning building's thermal rehabilitation, and it allowed to assess the potential of exploring the SCE database through multivariate analysis techniques. In a next step, with an expanded sample, it will be used to assess the usefulness of a specific retrofitting solution in the Portuguese building stock.

Through multivariate data analysis, the relationship between solar orientation, age (constructive characterization) and cooling needs of Portuguese residential units was explored.

2. Multivariate Data Analysis

The Multivariate analysis allows converting data into knowledge, through statistical techniques that simultaneously analyse several dimensions of the individuals under investigation. This was made possible due to improvements in hardware and software permitting to analyse big amounts of data [1]. For this purpose, *STATISTICA* v8 software was used.

2.1. Data collection

In the context of Energy Performance of Buildings Directive (EPBD), the SCE, which aims to ensure and promote the improvement of the energy performance of buildings, has a database of Energy Performance Certificates (EPCs) of mandatory issuance, since 2009, when real estate is transacted [2].

Data collection was made through direct consultation of EPCs. A sample of 35 residential units (entities) was established and 13 quantitative and six qualitative variables were selected.

2.1.1. Variables

Since data was collected directly from EPCs issued prior to the thermal performance regulations currently in force (Decree-Law no. 118/2013, of 20 August), some adjustments in the chosen variables were needed. In particular, it was not possible to collect information about glazing and shading.

The 19 variables collected cover four subjects: location, geometry, construction characteristics and energy performance (Table 1).

2.2. Sample characterization

The residential units were built between 1900 and 2011, with the highest frequency in the 1980s (modal class). The large majority are apartments (77%) with medium thermal inertia. Only 20% of the entities do not have any South orientated facade, none of which

TABLE 1: Variables collected.

	DESIGNATION	ACRONYM	UNITS / VALUES
LOCATION	county	C	
	summer climate zone	S	cool / warm / hot
	winter climate zone	W	warm / mild / severe
	no. of South facade	SFn	Z+
	no. of North facade	NFn	Z+
	no. of East-West facade	EWFn	Z+
GEOMETRY	building type	BT	single-family / apartment
	typology (number of rooms)	Tn	Z+
	floor area	Afloor	m^2
	headroom	H	m
	no. of facades	Fn	Z+
	glazing area < 15% Afloor	Aglazing	yes / no
CONSTRUCTION CHARACTERIZATION	year of construction	Yconst	Z+
	thermal inertia	Ti	light / medium / heavy
	facade thermal transmittance	Ufac	$W/(m^2.K)$
	solar factor	SF	0-1
ENERGY PERFORMANCE	heating energy need	Hen	$kWh/m^2.year$
	cooling energy need	Cen	$kWh/m^2.year$
	energy class	EC	mean value cf the energy class interval in which the residential unit belongs

is a single-family dwelling. Most of the residential units belong to the C energy class and are located in warm Winter and Summer climate zones.

2.3. Principal components analysis

Faced with a large number of variables, *Factor analysis* can be used to observe relationships or underlying patterns between them. Through those techniques it is possible to establish whether the information can be summarized or condensed in a smaller set of factors or components [1].

In order to obtain data reduction, we used the exploratory principal components analysis (PCA), where the total variance present in the set of original quantitative variables is considered.

Since the original variables were found in scales of different measures, it was necessary to standardize them so that their variance became unitary.

The factors were extracted by applying the principal components analysis to the correlation matrix (Table 2). Kaiser and Cattell criteria were used to determine the number of factors to retain in order to represent adequately the initial data. The former, based on excluding factors with eigenvalues smaller than 1.0 and including enough factors to explain at least 70% of the total variance, allowed the retention of four principal components which explained cumulative total variance reaches 75%. The later, known as the scree test [3], implies plotting the latent roots (eigenvalue) against the number of factors in their order of extraction. The cut-off point is given before the inflection

point of the resulting curve [1]. Applying this criterion, five principal components were retained explaining 83% of the cumulative total variance.

Table 2: Correlation Matrix of quantitative variables.

	EC	Yconst	Afloor	H	Hen	Cen	Ufac	SF	NFn	SFn	EWFn	Fn	Tn
EC	1.00												
Yconst	-0.16	1.00											
Afloor	-0.29	0.10	1.00										
H	0.28	-0.52	0.05	1.00									
Hen	0.64	-0.31	-0.28	0.22	1.00								
Cen	0.11	-0.14	0.44	0.30	-0.02	1.00							
Ufac	0.24	-0.78	-0.13	0.41	0.33	0.26	1.00						
SF	0.18	-0.31	0.12	0.14	0.26	0.10	0.31	1.00					
NFn	0.24	-0.02	0.02	0.14	0.22	-0.03	-0.13	-0.20	1.00				
SFn	0.03	0.04	0.14	0.36	0.05	0.14	-0.03	-0.11	0.56	1.00			
EWFn	-0.08	-0.02	0.13	-0.19	0.17	0.08	0.15	0.20	-0.51	-0.71	1.00		
Fn	0.06	0.03	0.25	0.17	0.39	0.16	0.04	-0.01	0.64	0.51	0.06	1.00	
Tn	-0.06	-0.07	0.52	0.33	0.01	0.19	0.15	-0.04	0.23	0.38	0.06	0.57	1.00

The use of four and five principal components was tested using orthogonal rotation methods in order to make them more easily interpretable. The orthogonal *Varimax* method allowed maximizing a variable´s loading on a single factor through an iterative process.

2.3.1. Results and discussion

The four principal components rotated have significant loads (greater than 0.7) for all variables, except for *cooling energy need* (Cen) with the value of 0.49. The variables factor loading in each principal component indicate which component they most correlate with. By identifying and interpreting each principal component, using the load of the variables that compose them, it was easily assigned the following designations: PC1 - *Solar orientation*; PC2 - *Constructive characterization*; PC3 - *Geometry* and PC4 - *Energy performance*. However, this factor solution only accounted for 44% of the variance of the variable *Cen* (communality of 0.44) (Table 3).

With the inclusion of one more component and after rotation, *Cen* factor loading reaches 0.90 in a single component. In this five components solution, *Cen* communality was also higher, resulting in the explanation of 85% of its variance. Nevertheless, the dispersion of variables related to geometry (*Fn, Tn, Afloor*) by two factors is not an improvement, as well as the less significant values (<0.70) in the load of some variables (*NFn, H, Afloor*).

Table 3: Four principal components solution: factor loadings and communality.

QUANTITATIVE VARIABLE	FACTOR LOADING				COMMUNALITY
	PC 1 Solar Orientation	**PC2** Construction Characterization	**PC3** Geometry	**PC4** Energy Performance	
NFn	0.72	-0.16	0.23	0.42	0.77
SFn	0.87	0.06	0.31	0.04	0.86
EWFn	-0.92	-0.05	0.26	0.17	0.94
Yconst	0.04	-0.86	0.05	-0.16	0.77
H	0.32	0.72	0.19	0.10	0.68
Ufac	-0.17	0.84	0.03	0.22	0.78
Afloor	-0.06	-0.03	0.79	-0.38	0.76
Fn	0.29	-0.14	0.71	0.52	0.87
Tn	0.16	0.13	0.81	0.05	0.69
Cen	-0.05	0.42	0.49	-0.15	0.44
EC	0.09	0.25	-0.19	0.71	0.60
Hen	-0.06	0.20	0.00	0.90	0.86

Therefore, the factor solution with only four principal components, easier to analyse and still representative of the initial set, was chosen.

By representing the variables in the space of the respective factors, it is possible to confirm that this data condensation is representative of Portuguese building stock. In Figure 1, regarding *PC2* - Construction characterization it is observed that the more recent the residential units, the lower their headroom and the more thermally isolated.

2.4. Cluster analysis

The main purpose of cluster analysis is to group objects based on their characteristics [1]. This multivariate technique allows classifying objects observing only the similarities or dissimilarities between them, without previously defining inclusion criteria in any group. It tries to organize a set of individuals, for whom detailed information (variables) is known, in relatively homogeneous groups (clusters) [4].

With the purpose of examining hypotheses relating energy performance of residential units, their solar orientation and age, a cluster analysis was performed by *hierarchical* and *non-hierarchical* methods.

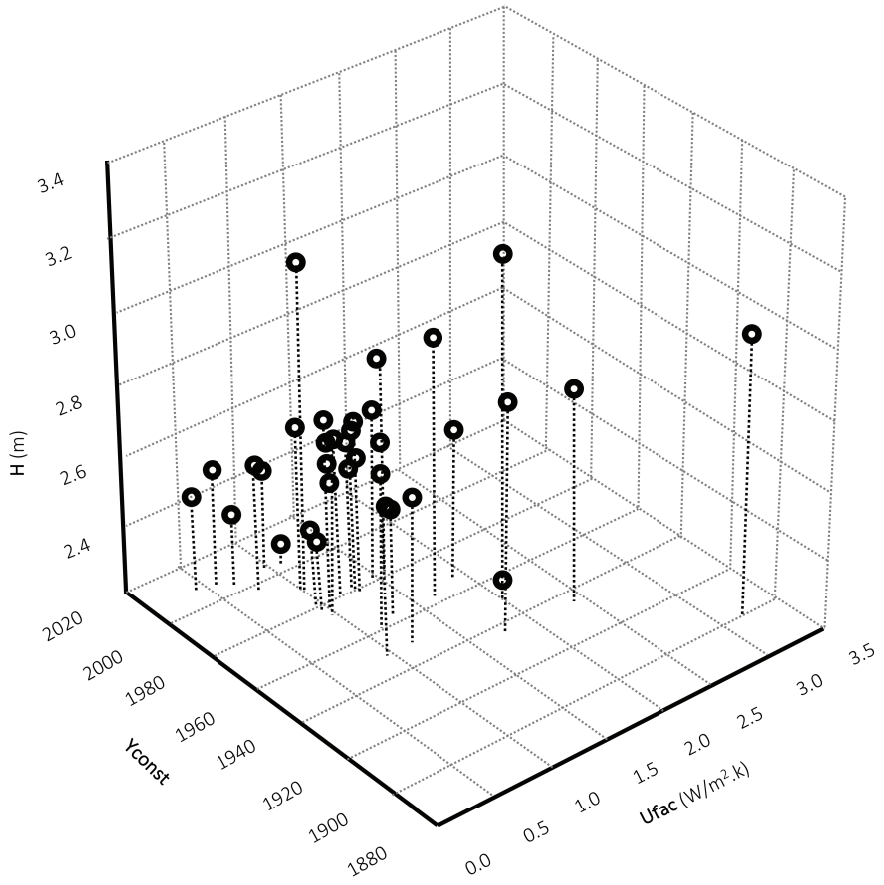

Figure 1: Scatterplot of year of construction (xx axis) against facade thermal transmission coefficient (yy axis) and headroom (zz axis).

Several hierarchical methods with different distances were tested, in order to determine the number of clusters to be applied later to non-hierarchical methods. The latter, being an iterative process, represents an advantage over hierarchical ones, by allowing the reallocation of entities to other groups, seeking the optimization of pre-defined criteria. In the process, different solutions arose with three to six clusters as the ideal number.

Using the non-hierarchical method, *K-means*, several solutions were tested which included using three to six clusters, several options regarding the criterion used to select the seeds and different variable sets.

In a first step of the grouping process, the 13 quantitative variables were considered: *EC; Yconst; Afloor; H; Hen; Cen; Ufac; SF; NFn; SFn; EWFn; Fn*; and *Tn*. Given the difficulty of interpreting the solutions, especially with a high number of clusters, the number of variables was reduced by using the information produced during the PCA.

Avoiding redundancy of information, one variable from each principal component was selected, promoting the selection of variables with low correlation. From *PC1* the choice was the variable with the highest factor loading (-0.92), and whose characteristic is generally complementary to the others, *EWFn*. In most cases the residential units have two opposing facades, orientated to East / West or North / South. The variable elected from *PC2* was *Ufac*. Although *Yconst* have a factor loading slightly higher than *Ufac* (-0.86 / 0.84), it has a greater dispersion (Std. Dev 24.00 / 0.53), so it would have more impact on the value of similitude. Considering that *Ufac* directly characterizes the construction, its choice is intended to preserve the analysis of the cases in which the construction, although old, was retrofitted, already presenting a low *Ufac* value. Although *Cen* have the lowest factor loading in *PC3*, and only is indirectly related with geometry, it was nevertheless elected because it is the most relevant to the purpose of the analysis. From *PC4*, *EC* was rejected against the option to include *Hen*, because it displays several dimensions out of the scope of this analysis. *EC* refers to primary energy, including not only the energy used for heating and cooling, but also hot water production, and incorporating equipment efficiency and energy type used.

Since the original variables were in different scales and units, it was essential to standardize them so that all have equal discriminative capacity. The procedure consisted in making their means zero and their standard deviations one.

2.4.1. Results and discussion

Two solutions of three clusters created with the *K-means* method were compared. Solution A uses the four selected variables from the PCs. The criterion to choose the seed was: *choose observations to maximize initial between-cluster distances*. Solution B uses 13 variables and the criteria to choose the seed was: *sort distance and take observations at a constant interval*.

The solution A was considered suitable for interpretation and intuitive explanation. However, data reduction (only 4 variables) did not allow clusters to be easily labelled (Fig.2).

On the other hand, in solution B (Fig.3), with three clusters extensively characterized, labelling was not difficult.

Cluster 1 - *favourable solar orientation* is mainly constituted by residential units with a high energy class, recent, smaller, with low ceilings, reduced heating and cooling needs, higher thermal isolation, a small number of facades predominantly north / south orientated and fewer interior partitions.

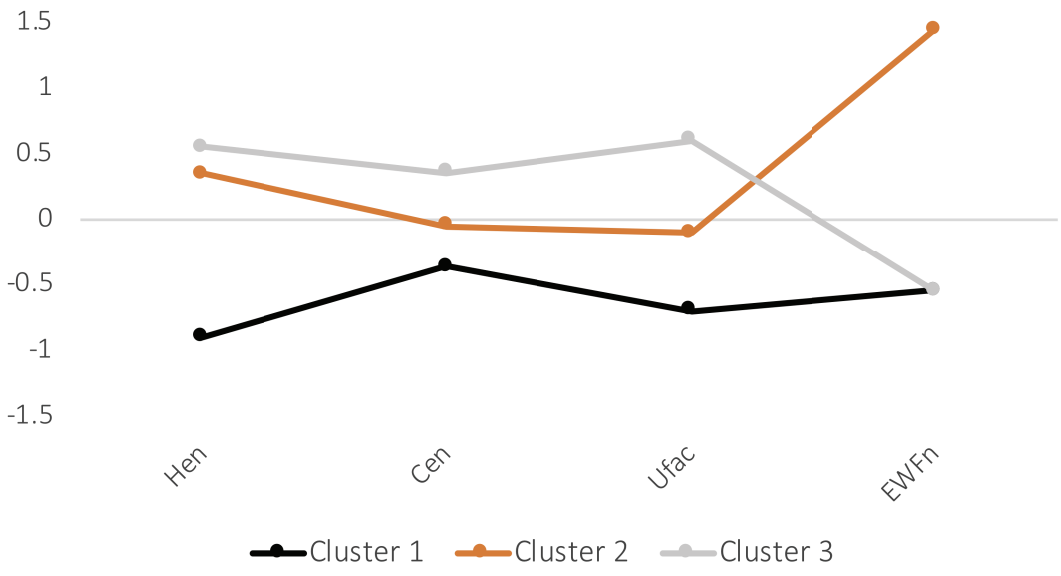

Figure 2: Clusters profile solution A: yy axis - average of the standardized variables; xx axis - four variables selected from each Principal Component.

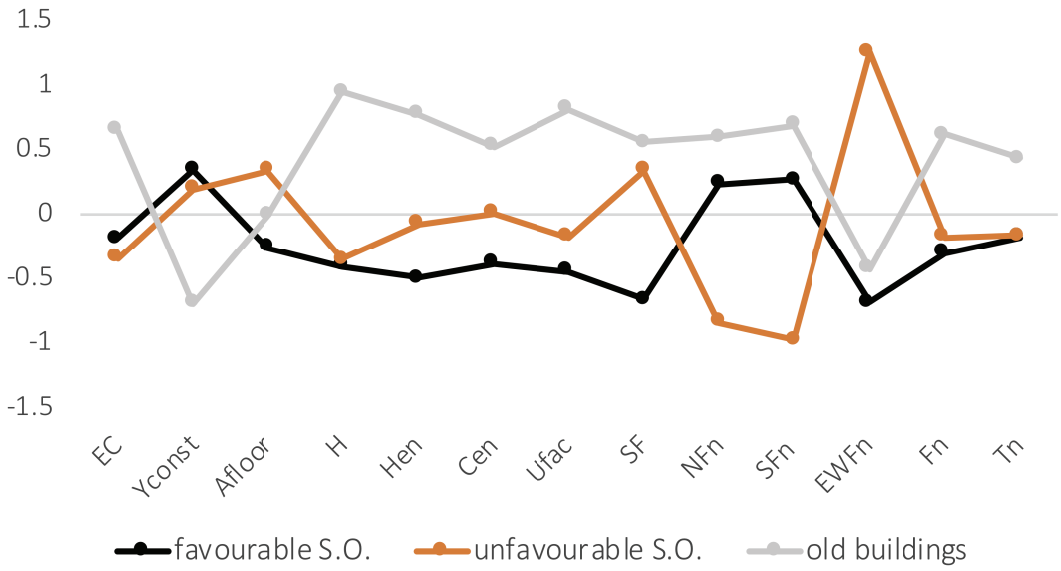

Figure 3: Clusters profile solution B: yy axis - average of the standardized variables; xx axis – 13 quantitative variables.

The residential units from Cluster 2 - *unfavourable solar orientation* have high energy class, are recent, bigger, with low ceilings, medium heating and cooling needs, have few facades, with medium thermal resistance, predominantly East / West oriented, and fewer interior partitions.

Cluster 3 - *old buildings,* groups entities with low energy class, older, medium-sized with high ceilings, high heating and cooling needs, many facades predominantly North / South oriented, with low thermal resistance and many interior compartments.

Through the analysis of the allocation of each entity to a cluster in solutions A and B, it was observed that: two entities, which have facades in contiguous quadrants, are classified both as having favourable and unfavourable solar orientation; four entities that in solution B belong to *favourable solar oriented* buildings, in solution A are assigned to the *Old Buildings* cluster. Considering that the four entities have favourable solar orientation (predominance of North / South facades) and that only one could be considered old (*Yconst* < *Yconst* mean), solution B seems the obvious choice.

The selected solution (B), which groups the entities in three clusters (old buildings, buildings with favourable and unfavourable solar orientation) allowed full confirmation of the expectations initially formulated: i) residential units with north-south solar orientation facades have lower cooling and heating energy needs; ii) residential units with east-west solar orientation facades have higher cooling needs; iii) older residential units have higher heating energy needs.

In order to assess if there was any underlying relationship in the chosen grouping, the characteristics described through qualitative variables were examined. It was observed that cluster 2 of unfavourable oriented buildings has only 18% single-family buildings. The houses, which allow greater freedom of solar orientation, are in greater percentage in clusters 1 and 3 (with 29% and 30%, respectively). In cluster 3 - *Old buildings*, more residential units have heavy thermal inertia (60%) than in clusters 1 and 2 (50% and 36% respectively). This feature is achievable through thick walls, typical of old buildings. These observations give confidence in the choice made of solution B.

2.5. Multiple regression

The relationship between a single dependent variable and several independent variables can be analysed through the generation of a multiple regression model. The regression analysis procedure weights each independent variable, representing their relative contribution to the overall prediction and assists interpretation because it contemplates the influence of each variable in the dependent variable [1].

The present analysis has an explanatory/confirmatory and not predictive purpose. It is intended through the generation of a model the identification/confirmation of the factors, and their relationships, which contribute to the degree of comfort inside the dwellings in the cooling season (Summer).

The variable *cooling energy need* (Cen) was elected as the dependent variable, since it illustrates the dwellings' comfort in the cooling season. Conceptual issues were first

analysed in order to select independent variable candidates, that could influence the dependent variable, to incorporate in the regression model.

From the subject *Location*, two variables were selected: *Summer climatic zone* (S), which indicates the weather severity, and *number of East-West facade* (EWFn), since this solar orientation increases overheating risk. From *Geometry,* the dwelling *headroom* (H) was elected, because the higher this value is, the greater is the stratification of air by temperature, enhancing the convective movement of air, which promotes cooling. From *Energy performance*, the *heating energy need* (Hen) was chosen because the promotion of its reduction may contribute to the increase of the dependent variable (Cen). From the *Constructive characterization*, three variables were selected: *year of construction* (Yconst), since constructive techniques change over time; *thermal inertia* (Ti), since the building heat storage capacity works positively in the summer, generating time-lag effect; and *facade thermal transmission* (Ufac).

Both *Yconst* and *Ufac* characterize the walls' construction. *Yconst* does it indirectly and does not safeguard the situations in which the walls have already been rehabilitated. However, it contains information about all of the envelope, namely facades and roofs. The information overlap of these two variables is showed in the correlation matrix (Table 2), where the value of the correlation coefficient between the two variables (-0.78) is highlighted, revealing that the simultaneous introduction of the two variables in the model may lead to problems of collinearity.

Observing the correlation matrix (Table 2), it can be seen that the variable chosen as dependent, *Cen*, has low correlation coefficients with all the variables, except for *floor area* (Afloor) (0.44), which was not selected. It should be noted that the *Cen* values refer to square meters. The independent variables, *Ufac*, *Yconst*, *H* and *Hen*, previously mentioned as candidates for inclusion in the regression model, have higher correlation coefficients between them, than each of these variables *per se* has with the dependent variable. This may cause problems to the estimation process of the regression model. As multicollinearity results in large portions of shared variance and low levels of single variance, the individual effect of the independent variable becomes less distinguishable. This situation, besides being able to affect the statistical tests of the coefficients or of the whole model, can lead to incorrectly estimated coefficients with the wrong signal, compromising the capacity of the regression procedure, its representation and interpretation [1]. To be precise, the situation initially proposed is the opposite of the ideal, which corresponds to having a number of independent variables highly correlated with the dependent variable, but with low correlation between them [1]. In order to avoid this situation, it was considered the possibility of using as candidates for

independent variables the factors created through the analysis of principal components (not correlated with each other).

As observed in the correlation matrix (Table 4), the relation between the principal components is practically null. It should be noted that the dependent variable, *Cen*, presents only significant correlation with the *PC2* and *PC3*.

The PC1, *Solar Orientation*, is composed by the variables *EWFn*, *SFn* and *NFn* with high complementarity between them, since most buildings present facades in opposite quadrants. The PC2, *Constructive Characterization*, incorporates the variables *Yconst*, *H* and *Ufac*, previously indicated conceptually as candidates for incorporating the regression model, albeit potentially creating problems due to their multicollinearity. The PC3, *Geometry*, contains the dependent variable *Cen*, and *Afloor*, *Fn* and *Tn*, initially conceptually excluded from the regression model. PC4, *Energy Performance*, is composed by *Hen* and *EC*, the latter conceptually excluded because, as previously mentioned, it incorporates information not related to the addressed topic and it contains in itself *Cen* and *Hen*. Although conceptually it makes sense to incorporate *PC1, PC2* and *PC4* into the model; only *PC2* shows a significant correlation value (0.42) with *Cen*. Another alternative to the use of PCs was the selection of a variable from each PC, promoting the selection of variables poorly correlated between them.

TABLE 4: Correlation Matrix of Principal components and dependent variable (*Cen*).

	PC1	PC2	PC3	PC4	Cen
PC1	1.00				
PC2	-0.0000003	1.00			
PC3	-0.0000007	-0.0000003	1.00		
PC4	0.0000008	-0.0000002	0.0000003	1.00	
Cen	-0.05	0.42	0.50	-0.15	1.00

Therefore, three sets of independent variables were considered candidates for inclusion into the regression model. Set 1, which independent variables were selected only by conceptual criteria of the phenomenon under analysis (*EWFn, H, Yconst, Ufac* and *Hen*). Set 2, formed by principal components, in order to avoid the high multicollinearity (*PC1, PC2* and *PC4*). Set 3 composed by a variable of each principal component, promoting the selection of poorly correlated variables (*SFn, Ufac* / *H, Tn and Hen*).

In the three sets, it was considered to include the information from the qualitative variables: *summer climatic zone* (S) and *thermal inertia* (Ti), by transforming them into *dummy* variables. The variable *Ti* has three levels (light, medium and heavy), being that only entities with heavy and medium thermal inertia were observed. Therefore, it is only necessary to represent these two levels. For that purpose, it was only necessary to create one variable: *thermal inertia heavy* (Ti heavy), assigning the value of 1 to the entities that have this characteristic. In the absence of it (value 0), it is assumed that the residential unit has a medium thermal inertia. The variable *S*, related to the climate severity, has three levels (cool, warm and hot), therefore two variables, *S hot* and *S cool*, become necessary. The entities that have zero value for both of these are in the region of *S warm*.

Several multiple regression models were generated by *standard* and *stepwise forward* methods. The *stepwise* method consists of a sequential search process, following the criterion of maximizing the predictive power increment with the least number of independent variables. The variables are individually evaluated for their contribution to predict the dependent variable and added or removed from the regression model based on their relative contribution. *Stepwise* methods have the drawback of exploiting only one variable for selection at a time. It disregards that the simultaneous presence of variables may contribute to the explanatory power, even though individually these variables do not have significant explanatory power. On the other hand, the *standard* method evaluates all possible combinations of independent variables [1]. Any method assumes that the theoretical and conceptual bases of the phenomenon in question are present when selecting candidate variables to be included in the model.

In order to compare the explanatory capacity of the models, the adjusted coefficient of determination, *adjusted* R^2, was used to measure the explanatory capacity considering the number of independent variables included in the model and the size of the sample.

2.5.1. Results and discussion

The chosen model was generated by the *stepwise forward* method with the PCs as candidate variables and the exclusion of two entities suspected of being outliers. For its election, only the criteria of parsimony, adjusted coefficient of determination and global p-value were used, and checked whether the signs of their coefficients made sense in terms of the conceptual meaning of the phenomenon being studied (Tables 5 and 6).

The model interpretation was consistent and logic since the thermal mass positively contributes to the reduction of cooling energy need, as well as the location in a cold

Table 5: Summary statistics of the elected model.

Summary Statistics; DV: Cen	Value
Multiple R	0.77497
Multiple R^2	0.60058
Adjusted R^2	0.55926
F(3,29)	14.53522
p	0.00001
Std.Err. of Estimate	5.82038

Table 6: Regression model summary.

Regression Summary for Dependent Variable: Cen F(3,29)=14,535 p<,00001 Std.Error of estimate: 5,8204						
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(29)	p-level
Intercept			16.76179	1.656965	10.11596	0.000000
PC2	0.479981	0.119140	4.19842	1.042122	4.02872	0.000370
S cool	-0.434222	0.119063	-7.58515	2.079840	-3.64699	0.001034
Ti heavy	-0.303369	0.117495	-5.24059	2.029682	-2.58198	0.015142

summer climate zone. The positive Beta regression coefficient sign of the variable *PC2*, indicates that older houses with higher ceilings and low thermal resistance of the walls have higher cooling needs. This interpretation is consistent with the previously verified relationship between the dependent variable *Cen* and the variables *Yconst*, *H* and *Ufac* constituents of *PC2*, through their correlation coefficients (table 1). It should be noted that it does not confirm the idea that more thermally isolated dwellings, the most recent and with smaller headroom, have higher cooling needs.

According to the generated model, comparing beta coefficients (Table 6), the constructive characteristics, *PC2* (|0.48|; construction year, headroom and facade thermal

transmittance), has higher influence in the cooling need than the climate (|-0.43|). In this model, the thermal mass is the variable which less influences it (|-0.30|).

The selected model, with a coefficient of determination (R^2) of 60%, incorporates non-metric data (dummy variable) related to summer severity (*S Cool*) and thermal inertia (*Ti Heavy*) (equation 1).

$$Cen = 16.76 + 4.20\ PC2 - 7.59\ S\ cool - 5.24\ Ti\ heavy \quad (1)$$

However, it excludes solar orientation information. The low correlation between the *Cen* and the solar orientation variables did not allow to generate a regression model, confirming the formulated hypothesis that solar orientation has a direct influence on the energy cooling needs. Possible reasons for this are: small sample size; limitations in the collection of shading and glazing information and the quality of the data concerning energy needs.

3. Conclusion

Through the application of multivariate analysis techniques to the SCE database, it was possible to:

1. condense to only four principal components the characteristics of the residential units: *solar orientation, constructive characterization, geometry* and *energy performance*; making information more manageable.

2. clustering entities in *favourable* and *unfavourable solar orientation* and *old buildings* allowing to dilute the particularities of each entity, thus simplifying the interpretation of the data through generalization.

3. Generate a regression model in order to explore/confirm which factors influence summer comfort.

Using this approach, it was shown that the exploration of the SCE database through multivariate data analyses has an enormous potential to convert data into knowledge.

Acknowledgment

The authors acknowledge the support of the CERIS, IST, UL, and also FCT for the financial support to the first author through PhD scholarship no. PD/BD/135215/2017.

References

[1] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis: Pearson New International Edition.*, Seventh Ed. (Pearson, Harlow, Essex - UK, 2014).

[2] Concerted Action EPBD, *Implementing the Energy Performance of Buildings Directive - Featuring Country Reports 2012*, no. 1. (ADENE, Porto - PT, 2013).

[3] R. B. Cattell, 'The Scree Test For The Number Of Factors', *Multivariate Behav. Res.*, vol. **1**, no. 2, p. 2010,(1966). https://doi.org/10.1207/s15327906mbr0102_10

[4] E. Reis, *Estatística Multivariada Aplicada*, 2^a Edição. (Edições Sílabo, Lisboa – PT, 2001). (Portuguese)