Towards a multidimensional segmentation of the real estate market of apartments in Santiago de Chile: a probabilistic approach from the energy building simulation

Felipe Encinas Pino[1*], Francisco José Sanchez De La Flor[2], Carlos Aguirre Nunez[3], Alejandro Rincon Casado[4]

[1]Architecture et Climat, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[2] Máquinas y Motores Térmicos, Escuela Superior de Ingeniería, Universidad de Cádiz, Cádiz, Spain
[3] Centro de Política de Suelo y Valoraciones, Universidad Politécnica de Cataluña, Barcelona, Spain
[4] Ingeniería Mecánica y Diseño Industrial, Escuela Superior de Ingeniería, Universidad de Cádiz,Spain

Abstract

It is clear that the thermophysical properties of materials, occupancy patterns and internal gains represent some of the most important sources of uncertainty in the field of building simulation. Uncertainty and sensitivity analysis deals with this situation, since it can generate a great range of forecast values based on the distribution of the input variables. However, most of the building energy simulation programs are deterministic, rather than probabilistic and consequently their results frequently are not expressed in terms of probabilities. On the contrary, the probabilistic approach requires a more complex process, since parameters quantification requires not only an assessment of the point estimate, but also an assessment of the uncertainty. Probably this situation appears as particularly critical when the focus is place on a group of cases instead of a single building case. This research aims to define the uncertainty of the predicted energy performance of apartments from the real estate market of Santiago de Chile by means of the Monte Carlo Analysis (MCA) method. These units were previously classified using the k-means clustering method with the aim of generating representative market segments. A total of 7 input parameters constitute the basis for these analyses using the standard EN ISO 13790 as calculation algorithm for estimating the annual heating demand.

Keywords: energy performance, real estate market, cluster analysis

1. Introduction

It is clear that the introduction of building performance simulation models represents an important contribution to decision-makers' understanding and overview of the decision-problem in the field of energy performance. Indeed, many of the answers that can be formulated at the light of these models can improve the certainty level with respect to aspects such as the material definition of the envelope or the glazing ratio of the building. Several studies have shown that uncertainties in design evaluations may be quite substantial that makes it imperative to convey them to the decision makers [1]. Consequently, if the design process is carried out based on informed decisions, the possibilities to reach the desirable level of energy performance and/or thermal comfort are considerable.

At the same time, most energy building simulation programs are deterministic rather than probabilistic, and as such, offer their results in the form of deterministic values. Also, none of these programs offer methods to propagate uncertainty of the input parameters on the model output. On the contrary, the probabilistic approach requires a more complex process, since parameters quantification requires not only an assessment of the point estimate, but also an assessment of the uncertainty. Probably this situation appears as particularly critical when the focus is place on a group of cases instead of a single building case. The probability theory allows conducting an uncertainty assessment of the simulation output (dependent variable Y) based on the uncertainty of one or more input parameters (independent variables, $X = \{X_1,...,X_n\}$) and where $Y = f(X)$. In this respect, literature identifies two closely related methods to assessing uncertainties, which can be defined as uncertainty analysis (UA) and sensitivity analysis (SA). UA assesses the uncertainty in model outputs that derives from uncertainty in inputs parameters. SA assesses the contribution of the inputs parameters to the total uncertainty in analysis outcomes.

This research aims to define the uncertainty of the predicted energy performance as well as the input parameters that causes this uncertainty from a representative database with 21,902 apartments for sale in Santiago de Chile between 2004 and 2009. A total of 7 input parameters constitute the basis for these analyses using the calculation algorithm of EN ISO 13790 standard for estimating the annual heating demand.

* **E-mail:** f.encinaspino@uclouvain.be

2. Methodology

2.1. Introduction to the real estate market of apartments in Santiago de Chile

Santiago, capital city of Chile, with 5,638,820 inhabitants, concentrates the 37% population of the country and at the same time, the 65% of the private real estate supply [2]. Indeed, the real estate market of Santiago presents an important evolution in the field of real estate marketing during the last decades, due to it is possible to observe a very competitive market, strong marketing campaigns and complex real estate products. In that sense, variables such as orientation and micro-localization are very relevant for customers since the real estate supply develops their products with the aim of integrating more attributes when the external attributes decrease [3]. For example, in an apartment submarket focused on young couples with children, if a real estate project presents disadvantageous conditions of location with respect to the metropolitan centres and subcentres, it will increase its attributes, adding complementary public areas such as swimming pool and gym, and/or improving the technical conditions of its apartments.

A study carried out from a representative database with 21,902 apartments for sale in Santiago between 2004 and 2009, thanks to Portal Inmobiliario.com, shows as some market segments provide higher freedom for architects to positively discriminate a specific orientation or view. Indeed, this freedom can be applied, for example, assigning a higher living area to the north or northeast oriented apartments during the design process of the building layout [4].

Therefore, the identification of housing market segments is a matter of great importance. In the field of the supply analysis, typologies are usually defined according to price ranges, establishing supply segments. However, this kind of approach forgets that the real estate market can compensate the lack of specific attributes in their products with others. In consequence, it is required a multidimensional approach that integrate a series of attributes – additionally to the variable of price - allowing that these will be developed at the level of their chain value.

The already mentioned database contains a total of 5 and 17 continuous and categorical (discrete) variables, respectively. Between the first ones, it appears the supply price and area of apartments and between the second ones, the number of floors, bedrooms and bathrooms, orientation and location (boroughs from the Metropolitan Region) real estate attributes such as level of quality finishing or presence of central heating. In the following point, cluster analysis is presented as the selected tool to carry out the multidimensional approach from the information contained in this database.

2.2. Definition of market segments by means of a cluster analysis

There are different to identify market segments and the selection of real estate products using clustering methods is one of these. This numerical taxonomy technique attends to identify a group of entities (e.g. people, markets, organizations) that share certain common characteristics (attitudes, purchase propensities, habits, etc). Indeed, the seminal study carried out by Punj and Stewart in 1983 clearly established that cluster analysis is a statistical method of classification [5]. At the same time, according to Vivanco (1999), the use par excellence of cluster analysis is the generation of typologies [6]. A typology is a group of cases that present a strong similarity, which can be metrically defined and quantified.

To define the market segments by means of this method, the architectural layout (in terms of number of bedrooms and bathrooms) has been kept constant with the aim of considering the family life cycle and establishing supply niches. Table 1 presents the number of cases from the different architectural layouts in the Portal Inmobiliario.com database.

Table 1. Architectural layouts in the Portal Inmobiliario.com database, according to the combination between the number of bedrooms and the number of bathrooms

		Number of bedrooms								Total	
		1		2		3		4			
		N	%	N	%	N	%	N	%	N	%
Number of bathrooms	1	5,971	27.3%	1,814	8.3%	730	3.3%			8,515	38.9%
	2			6,304	28.8%	4,991	22.8%			11,295	51.6%
	3					671	3.1%	931	4.3%	1,602	7.3%
	4							490	2.2%	490	2.2%
Total		5,971	27.3%	8,118	37.1%	6,392	29.2%	1,421	6.5%	21,902	100.0%

Between the different possibilities of clustering method, k-means procedure from factor scores was selected for its application, since it was the method that generated clusters with higher differences between them. Initially, the two-stage method was also studied – because it is appropriate when there are both continuous and categorical variables – the obtained results did not allow a clear separation between the continuous variables. On the contrary, when the separation of clusters is carry out from a Principal Component Analysis (generating factor scores), the groups appear significantly different in their continuous variables per each supply niche. Table 2 presents the mean and standard deviation for the variables of price (expressed in terms of UF[1]) and living area by cluster. As it can be observed, a total of 9 different clusters were generated.

Table 2. Mean and standard deviation for the variables of price and living area by cluster

Architectural layout	Cluster			Price [UF[1]]		Living area [m²]	
	Id.	N	% of total	Mean	St. Dev.	St. Dev.	St. Dev.
1 bedroom 1 bathroom	1.1	4,036	18.4%	1,199	244	31.9	3.9
	1.2	1,935	8.8%	2,033	473	38.7	6.2
2 bedrooms x bathrooms	2.1	5,648	25.8%	1,939	550	51.4	9.4
	2.2	2,470	11.3%	3,058	770	66.2	11.3
3 bedrooms x bathrooms	3.1	5,099	23.3%	2,532	1,116	70.3	16.6
	3.2	192	0.9%	3,315	897	89.0	11.8
	3.3	1,101	5.0%	5,101	1,920	105.0	26,7
4 bedrooms x bathrooms	4.1	1,179	5.4%	5,662	1,333	119.6	14.5
	4.2	242	1.1%	6,973	2,446	138.2	31.2
Total		21,902	100.0%				

3. Results

3.1. Approach to the energy performance of the apartments from the obtained clusters

With the aim of establishing an approach to the thermal performance of apartments from the real estate market of Santiago, a thermal simulation model based on the bottom-up structure was proposed. This logic suppose the simulation of numerous individual cases, which in their ensemble provide a general approach by virtue of the common criterion that group them.

EN ISO 13790:2004 was chosen as the calculation algorithm for implementation of routines for energy building simulation. This standard gives a simplified calculation method for assessment of the annual energy use for space heating of a residential or a non-residential building [7].

Table 3 present the 7 considered input parameters, which are related to building envelope definition and architectural design criteria. Weather data is defined for Santiago de Chile from the NCh 1079-2008 standard and considered as fixed scenario for all simulations [8]. As can be observed, an amount between 2 and 4 variables were defined by each parameter, which were established using a triangular probability density function. This kind of distribution is generally used as a subjective description of a population for which there is limited sample data or where the relationship of variables is roughly known but data is scarce. These parameters and their distributions were applied with the aim to obtain the annual heating demand for a sample of cases that can be representative of the real estate market of apartments in Santiago de Chile.

Table 4 presents the 3 assigned values for the input parameter of living area by cluster. It can be observed that the proposed values for upper bound, intermediate value and lower bounds were defined by cluster based on their distribution function and statistical indicators (where kurtosis – understood as a measure of peakedness of the probability distribution – represents a key element for the decision).

[1] UF (*Unidad de Fomento*) is an indexed unit of account used in Chile. The exchange rate between the UF and the Chilean peso is constantly adjusted to inflation so that the value of the UF remains constant. It has become in the predominant measure for determining construction costs, values of housing and secured loans.
1 UF = 32.7 € according to the Banco Central de Chile, 17th May 2011

Table 3. Description of the input parameters used in the energy building simulation models

Input parameters	Unit	Variables by parameter	
		N	Description
U-value of walls	W/m²K	3	3 levels of thermal insulation in concrete walls (without insulation, 10 and 20 mm EPS 15 kg/m3)
U-value of windows	W/m²K	2	2 levels of thermal insulation corresponding to single and double glazing
Internal gains	W/m²	3	Low (5 W/m²), medium (10 W/m²) and high (15 W/m²) internal gains
Shading correction	(0-1)	2	Average shaded fraction of the area of opening for windows: with and without balcony
Living area	m²	3	3 levels of living area (upper bound, intermediate value and lower bound) according to statistical distribution criteria (Table 4)
Glazing ratio	%	3	Ratio of the glazed area to the total area of the exposed facade (30%, 40% and 50%)
Orientation	°	4	0°= North; 90°= East; 180°= South; 270°= West

Table 4. Assigned values for the input parameter of living area by cluster

	Living area [m²] by cluster								
	1.1	1.2	2.1	2.2	3.1	3.2	3.3	4.1	4.2
Upper bound	29,4	34,8	42,4	59,1	50,0	79,8	84,4	105,1	123,6
Intermediate value	31,8	36,9	50,5	66,3	69,4	88,3	97,6	119,3	130,8
Lower bound	34,9	41,6	60,8	73,5	93,9	96,9	126,6	134,8	144,0
Kurtosis	1,4	4,0	-0,8	1,5	-0,4	0,3	8,5	-1,0	7,6
Standard deviation	3,9	6,2	9,4	8,8	16,6	8,6	26,7	14,5	31,2
Criteria*	a	a	b	a	b	c	a	b	a

* Criteria for the definition of the upper bound, intermediate value and lower bound by cluster:
a. For kurtosis >0: percentile 20, percentile 50 (median) and percentile 80, respectively
b. For kurtosis <0: percentile 10, Huber's M-Estimator and percentile 90, respectively
c. For kurtosis=0; median – standard deviation, median and median + standard deviation, respectively

3.2. Comparison between Factorial Design and Monte Carlo models

From the given number of variables by parameter (presented in the Table 3), all the possible combinations of them produce a total of 1296 cases (where $\prod_n$ = 3*2*3*2*3*3*4), which correspond to the complete sample of the factorial design. However, with the aim of obtaining a more economical model - in terms of the number of simulations to obtain a sufficiently accurate result – a model based on the Monte Carlo Analysis (MCA) method was applied. According to literature, this method represents the most widely applicable and easy to implement solution for this purpose [9].

MCA performs multiple evaluations with randomly selected input parameters, generating input and output distributions useful in assessing model and parameter uncertainties in a global sense. The range of each variable is divided into N non-overlapping intervals of equal probability 1/N. In a first approach, MCA was considered for just 36 cases, which were obtained from the square root of the complete sample (Factorial Design). This estimation is based on the concept that MCA relies on the central limit theorem – which establishes the mean of a sufficiently large number of independent random variables will be approximately normally distributed - and consequently, the number of trails could be defined by means of $\sqrt{N}$ [10]. However, this amount would be lower than the recommended number of cases, since it has been estimated than only marginal improvements in accuracy are obtained after 60-80 simulations [11]. To determine the level of statistic similarity between the complete sample of Factorial Design and different MCA samples according to their number of trials, the t-test for unequal sample sizes was applied.

Table 5 presents the results of the parametric t-test for both Factorial Design and MCA for cluster 2.1 (the group that concentrates the greatest number of cases, accounting for the 25.8% of the total) and considering four alternatives of number of trials. This test assesses the statistically significant difference between the means of two conditions (in this case, unpaired samples). The procedure basically consists in the estimation of the t-value (a ratio between a measure of the between-groups variance and the within-groups variance) and the probability of obtaining this statistic by sampling error ('t-critical' in the table). This probability determines the null hypothesis (H_0) of the test.

For the four MCA models levels, the obtained t-value is higher than t critical and consequently, the null hypothesis (H_0) is rejected. At the light of these results, it is possible to conclude that the differences between the means of Factorial Design and the four MCA samples are no significant and therefore their distributions can be considered as statistically similar. However, according to their obtained results, the MCA model of 100 cases was selected since it presents an adequate level of accuracy with a reasonable number of trials.

Table 5. T-test for Factorial and the three Monte Carlo models (according to the number of cases) for cluster 2.1

	Factorial	Monte Carlo			
		N=36	N=100	N=500	N=1000
Mean	2,018	3,150	3,175	2992	3,023
Variance	907,869	1,027,027	1,047,789	891033	968,837
Observations	1,296	36	100	500	1,000
Degrees of freedom		37	113	914	2114
t-value		6.6	10.9	19.6	24.6
P(T<=t)		9.06E-08	1.88E-19	1.98E-71	1.24E-117
t-critical		2.0			

Figure 1 presents the results of heating demands (in kWh/year) in terms of cumulative frequency for all clusters using a MCA model with 100 cases by each one. The results range for the total of 900 cases varies from 778 kWh for an apartment of cluster 1.1 (of 31.8 m², double glazing, 20 mm EPS 15 kg/m3 in walls, high internal gains, 30% glazing ratio, without balcony and west oriented) to the 9595 kWh of an apartment of cluster 4.2 (of 144 m², single glazing, without thermal insulation in walls, low internal gains, 50% glazing ratio, with balcony and south oriented). These two extreme cases shows the most favourable and unfavourable situations from the point of view of heating demand, respectively. In general terms, it is possible to observe that while higher the living area of clusters (as it was presented in Table 4), greater the heating demand. However, cluster 3.1 – while it presents greater living area than clusters 2.1 and 2.2 – shows in many cases a more favourable thermal performance than these others. This situation can be attribute to the fact that cluster 3.1 presents the greatest percent of walls with thermal insulation and double glazing in comparison to all the other clusters. This suggests that U-value of walls and windows are two very relevant input parameters in terms of their influence on the output model (annual heating demand).

This hypothesis can be tested by means of the uncertainty and sensitivity analysis by each cluster. Figure 2 presents the Spearman's rank correlation coefficient (rho) for all the input parameters with respect to the annual heating demand. This index represents a non-parametric quantitative estimation of the correlation (using a monotonic function) between the input and output values. The obtained coefficients show that in general, parameters that depend on the building quality are most sensitive than the others that do not depend on that. Between them, U-value of walls and windows are first and second in the rank respectively, according to their contribution to uncertainty prediction. On the contrary, the contribution of shading correction ratio (by means of balcony) is negligible, since its correlation is not statistically significant.

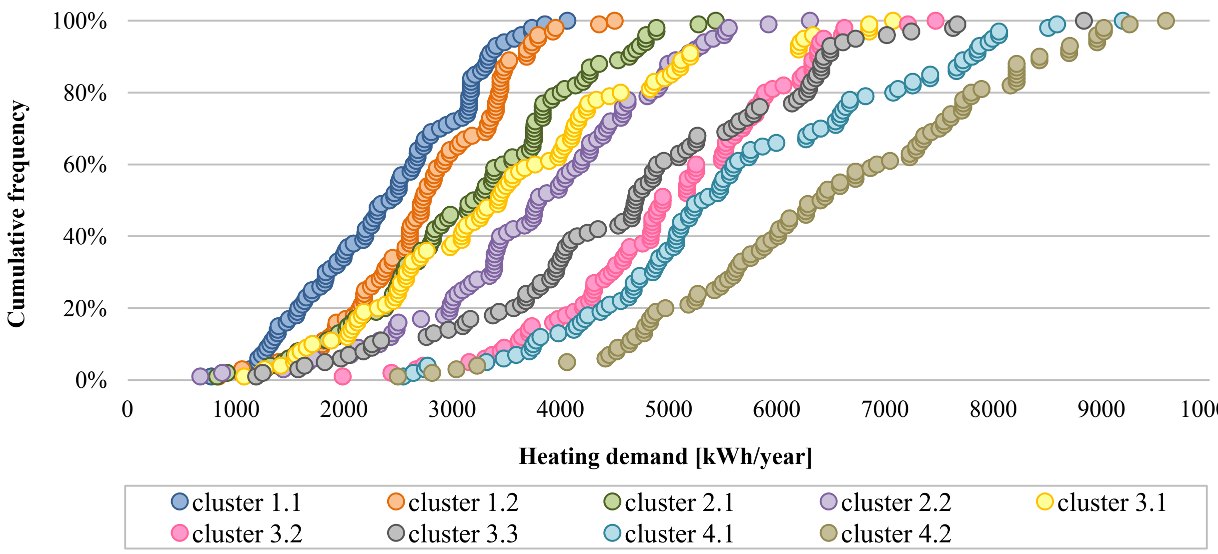

Figure 1. Cumulative frequency for heating demand by cluster

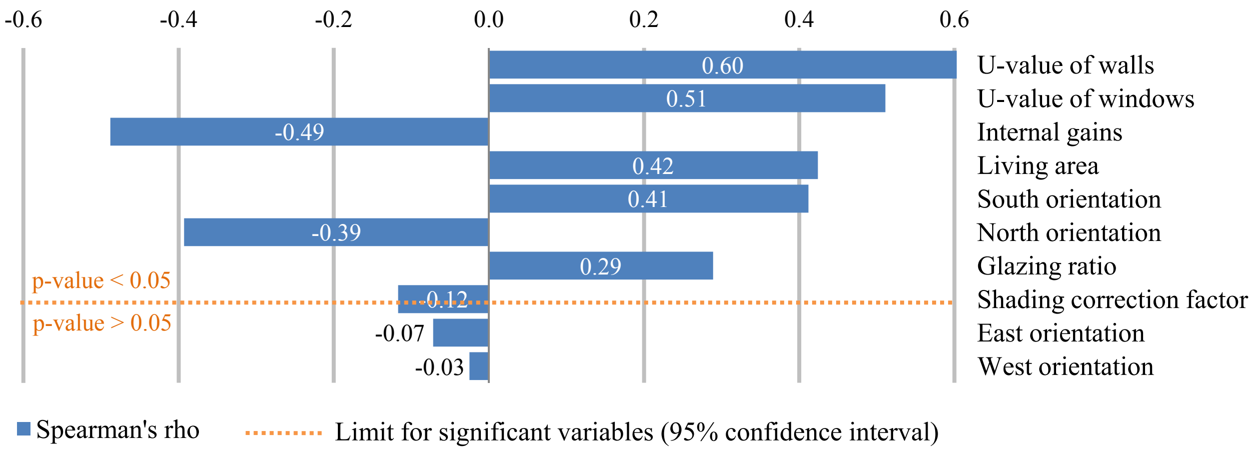

Figure 2. Spearman's rho for the different input parameters with respect to the heating demand for cluster 2.1

4. Conclusions

Monte Carlo Analysis (MCA) appears as an appropriate tool to propagate uncertainty of the input parameters on the model output in the context of the energy performance assessment (using the ISO 7730 standard as calculation algorithm). This was observed by means of the comparison between a Factorial design with 1296 cases and a MCA with 100 cases, where both sample distributions are statistically similar.

At the same time, uncertainty analysis (UA) and sensitivity analysis (SA) were used to investigate the uncertainty of the predicted thermal performance in a group of cases, which are representative of the real estate market of apartments in Santiago de Chile for the period 2004-2009. The U-value of both walls and windows and the internal gains are the most important and sensitive input parameters since they have the largest impact on the heating demand. Other parameters with a moderate influence on the assessment results are the living area, north and south orientation and glazing ratio, while shading correction factor can be considered as negligible.

Acknowledgements

This study was carried out as a part of a PHD Thesis at the "Architecture et Climat" research centre from the Université catholique de Louvain in Belgium. The study is funded by the "Bourse de coopération au développement" scholarship from the same university.

The authors would like to thank to Cristian Maturana, Managing Director of Portal Inmobiliario.com for providing the information about the real estate market that makes possible this study.

References

[1]De Wit, S. & Augenbroe, G., "Analysis of uncertainty in building design evaluations and its implications', *Energy and Buildings 34*, 951-958 (2002)

[2] CChC, "Mercados inmobiliarios, Análisis y Proyecciones, Primer Semestre 2008" (M. Vergara, Ed.), *Cámara Chilena de la Construcción*, Santiago (2008).

[3] McDonald, J.F., McMillen, D., "Urban Economics and Real Estate: Theory and Policy", *Blackwell Publishing* (2007)

[4] Encinas, F., "El impacto de la orientación en el precio de un departamento", *Diario Inmobiliario, Portal Inmobiliario.com*, Santiago de Chile (2010). http://www.portalinmobiliario.com/diario/noticia.asp?NoticiaID=14990

[5] Punj, G., & Stewart, D. W., "Cluster Analysis in Marketing Research: Review and Suggestions for Application", *Journal of Marketing Research 20* (2), 134-148 (1983).

[6] Vivanco, M., "Análisis Estadístico Multivariante. Teoría y práctica", Editorial Universitaria, Santiago (1999)

[7] CEN, "EN ISO 13790:2004 Thermal performance of buildings – Calculation of energy use for space heating", *European Committee for Standardization (CEN)*, Brussels (2004).

[8] INN, "NCh 1079-2008. Arquitectura y Construcción - Zonificación climático habitacional para Chile y recomendaciones para el diseño arquitectónico", *Instituto Nacional de Normalización (INN),* Santiago (2008)

[9] Hopfe, C.; Hensen, J & Plokker, W, "Introducing uncertainty and sensitivity analysis in non-modifiable building performance software" *Proceedings of the 1st IBPSA Germany/Austria Conference BauSIM*, International Building Performance Simulation Association, 9-11 October 2006, Munich