Original Research Article

Study of Indoor Radon Using Data Mining Models Based on OLAP Cubes

ABSTRACT

This research has focused on a radon measurement campaign that was carried out in two dwellings in a residential building located in Madrid. A new methodology has been used in this field, such as the use of cubes based on On-Line Analytical Processing in SQL Server Analysis Services. The application of this methodology can be of particular interest for analysing thousands of radon measurements and complementary variables, which are easily obtained in any radon measurement campaign.

Keywords: radon; OLAP; indoor air quality; clustering.

1. INTRODUCTION

Radon (or radon-222) is a chemically inert, naturally occurring, radioactive gas. It has no colour, smell, or taste, and is produced from the natural radioactive decay of uranium-238, which is found in rocks and soil. Radioactive gas escapes easily from soils into the air and to concentrate in houses. Soil is the most important source of residential radon. However, there are other sources of radon include building materials and water. Due to its radioactive nature, it represents the second leading cause of lung cancer, after smoking [1]. It is well-known that radon concentration varies throughout the day and also depends on the season of the year [2,3]. In fact, radon levels depend on many factors, like meteorology, radon exhalation rate from building materials and even earthquakes [4-6].

This research has focused on a radon measurement campaign that was carried out in two dwellings in a residential building located in Madrid. In this building, construction techniques have been used to optimise energy efficiency. As a result of this measurement campaign, the author has published several articles on the level of radon in the dwellings and its modelling using computer software [7,8]. This type of households is increasingly sealed off from the outside, which reduces the rate of air renewal compared to older housing. Recently, there has been a lot of research into the quality of indoor air in highly energy-efficient homes and public places [9, 10]. Some studies [11, 12] have included radon concentration as a new parameter to be monitored because of its negative influence on health. This growing interest in radon gas is justified because the Technical Building Codes (TBC) of many European countries, as is the case in Spain, have a specific section on radon gas protection in indoor spaces that provides a guarantee of home habitability [13]. On the other hand, the rise of teleworking due to the coronavirus and, in addition, the confinement of vulnerable people to their homes has led to a considerable increase in the time citizens spend indoors and, therefore, in their exposure to radon [14].

The aim of this article is to gain a better understanding of the behaviour of radon in enclosed spaces and its relationship with external variables such as weather and atmospheric conditions. To carry out this work, a new methodology has been used in this field, such as the use of cubes based on On-Line Analytical Processing (hereinafter, OLAP) in SQL Server Analysis Services. The radon measurements and other complementary information (environmental and atmospheric conditions) were stored in the cubes in order to establish relationships between them. A clustering algorithm (which can be translated as a clustering or segmentation algorithm) was used to find these relationships. The methodology used in this research allows to organize and make quick queries on a large set of data. The application of this methodology can be of particular interest for analysing thousands of radon measurements and complementary variables, which are easily obtained in any radon measurement campaign.

2. METHOD & MATERIAL

2.1 Measurements

The residential building is located in the district of “Barrio de Salamanca” (city of Madrid, Spain) with the coordinates of latitude and longitude 40.4383 and -3.6730, respectively (see Figure 1-a). This building is placed in a high-medium potential radon area, which was identified from the data provided by MARNA (it is an acronym in Spanish of “MApa de Radiación gamma Natural”) [11] and “Radon Project 10 × 10” [12]. The climate of Madrid is characterized by a dry summer and the coldest month averaging above 0ºC.

(a) (b)

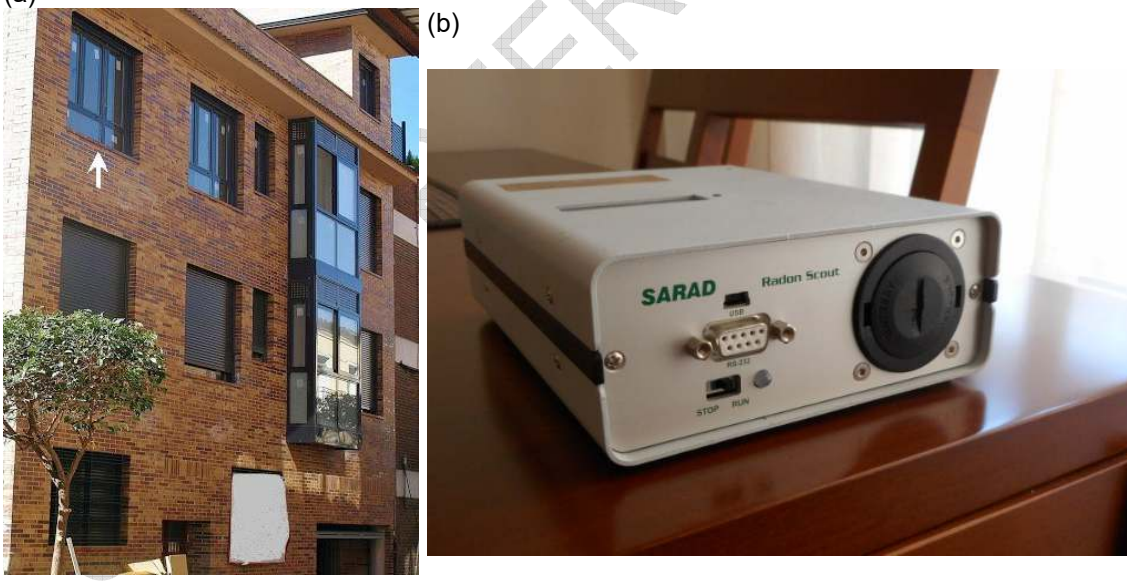

Figure 1. (a) Picture taken in the year 2012 showing the building’s facade. A white arrow points to the location of the dwellings (b) Radon-Scout monitor

Outdoor and indoor data were recorded every three-hourly for a total period of six months between May and November 2014 for two dwellings situated on the second floor in this building. Throughout this paper, the occupied dwelling was labelled with the letter ‘A’ and the unoccupied dwelling with the letter ‘B’. The indoor radon mesurements were carried out with two Radon-Scout devices from SARAD GmbH Company, Germany origin (see Figure 1-b).

The devices consist of a solid state silicon detector that identifies alpha particles emitted by radon gas and its decay products. These monitors are recommended as a highly sensitive detector for indoor radon surveys. More details of this survey can be found at previous papers of the author [7,8].

2.2 OLAP Cube and Data Mining

Data mining is the process of choosing, exploring, and organizing large amounts of data to uncover unknown relationships. These techniques are used commercially for customer grouping, fraud detection, trend analysis, etc. A variety of software tools are available to apply these techniques to our data. In this study we have used Microsoft SQL Server 2005 Analysis Services software that offers online OLAP analytical processing and data mining functions. OLAP is a BI technology (Business Intelligence) that allows multidimensional access to database. The core of this technology is OLAP cubes. An OLAP cube is a data structure specifically designed to quickly respond to MDX (multidimensional expressions) queries based on the dimensions and measures defined about the data source. The high speed of response is due to the fact that queries are not made directly on the data origin but on aggregated and pre-processed data. Therefore, every time there is a change in the data origin it is necessary to process the cube again. The cubes are composed of one or several fact tables, dimensions, measures and cube schemes. The dimensions are the categories in which the data can be classified and within the dimensions, you can establish hierarchies in order to obtain a lower level of data aggregation. Measures are the numerical values that users want to reorganize, aggregate and analyze.

The algorithm used is MicrosoftClustering. With this algorithm, the data are classified in groups so that the variances within each class are as small as possible and the variances between classes are as large as possible. The quotient between the coefficient of intra-class and inter-class variation is an indicator to determine the quality of the clustering process. A number of classes should be chosen for which the value of this index does not decrease as the number of classes increases. It is recommended to consult the documentation available in the official library of Microsoft called MSDN Library for further information on OLAP and the algorithm used [17].

3. RESULTS & DISCUSSION

3.1 OLAP Cube

This section describes the two OLAP cubes obtained from the data acquired by the radon detectors on both floors. The data source for cube 1 is the data obtained on floor 2A, and the data source for cube 2 is the data obtained on floor 2B. The diagram of both cubes resembles a star made up of a table of facts and three dimensions (environmental conditions, atmospheric conditions and time). The fact table has two measurements: radon concentration and the error associated with the measurement. An extract of the table of facts from cube 1 and cube 2 is shown in Table 1.

Table 1. Table of facts (a) Cube 1 y (b) Cube 2

a)	OLAP Cube 1				b)	OLAP Cube 2			
ID	Radon	Error	CondAmb	CondAtm	ID	Radon	Error	CondAmb	CondAtm
1	28.5	30.2	1	4	1	46.5	28.2	2496	4
2	48.3	24	2	5	2	58.3	25.4	2497	5

a)	OLAP Cube 1				b)	OLAP Cube 2			
ID	Radon	Error	CondAmb	CondAtm	ID	Radon	Error	CondAmb	CondAtm
3	65.3	21	3	6	3	96	16.4	1854	6
4	71	20	4	7	4	44.7	28.6	270	7
5	40	26	5	8	5	65	23.8	2498	8
6	57	22	6	9	6	89.3	18	2499	9
7	17.3	33	7	10	7	99.7	15.5	2500	10
8	48.7	25.1	8	11	8	68.7	22.9	2501	11
9	37	28	9	12	9	41.3	29.4	2502	12
10	51	24.5	10	13	10	75.3	21.3	2503	13
11	68.3	20.1	11	14	11	82.3	19.6	2504	14
12	79.7	17.2	12	15	12	72	22.1	2505	15
13	65.3	20.9	13	16	13	75.7	21.2	2506	16
14	34.3	28.7	14	17	14	61.7	24.6	2507	17

The measurements of the environmental conditions dimension are as follows: relative humidity (RelHum in the table), temperature (Temp) and atmospheric pressure (Pres). An extract from the table of this dimension for both cubes is shown in Table 2.

Table 2. Environmental conditions

ID	Temp	RelHum	Pres
1	26	44.5	934.5
2	25	41.3	934.7
3	24.5	42.0	934.0
4	24.2	42.0	934.0
5	24	42.0	934.0
6	24.5	42.3	933.7
7	24.5	42.3	932.0
8	24.5	42.0	930.0
9	25	38.3	930.3
10	24.5	34.7	932.3
11	24.3	37.0	933.7
12	24	37.7	935.0
13	22.3	38.3	937.0
14	22.8	39.0	939.0
15	23.2	39.0	938.7

The measures of the atmospheric conditions dimension are: Zi (m), 10xKz (m^2/s), U*(m/s), Kh (m^2/s), Zo (m), PSQ and the Julian date. The meaning of each of these variables is

shown in the legend in Table 3. This table concerning atmospheric conditions, of which only an extract is shown below, is the same for both cubes.

Tabla 3. Atmospheric conditions

ID	Zi	10xKz	U*	Zo	Kh	Julian Date	PSQ
1	1679	910.4	0.3988	0.75	18240	2456789.00	2
2	1943	1260	0.4684	0.75	25690	2456789.13	2
3	1080	1020	0.452	0.75	12990	2456789.25	2
4	102.3	102.3	0.2323	0.75	2925	2456789.38	4
5	58.45	61.28	0.1521	0.75	1752	2456789.50	5
6	94.43	38.14	0.09933	0.75	1090	2456789.63	5
7	112.5	52.33	0.1159	0.75	1457	2456789.75	5
8	550.3	440.1	0.08528	0.75	5559	2456789.88	3
9	1789	637.5	0.1974	0.75	18920	2456790.00	3
10	2251	1450	0.5424	0.75	28650	2456790.13	1
11	1738	1177	0.5744	0.75	17750	2456790.25	2
12	121.3	183.3	0.3272	0.75	5241	2456790.38	4
13	151.3	160.5	0.2785	0.75	4588	2456790.50	4
14	243.3	197.1	0.3181	0.75	5633	2456790.63	4
15	269.7	194.4	0.3173	0.75	5448	2456790.75	4

JDAY	Julian Date
PSQ	Pasquill stability classess
Zi	Planetary Boundary Layer Height
10xKz	Vertical mixing coefficient x 10 (m^2 s^{-2})
U*	Friction velocity (m/s)
Zo	Lenght of roughness (m)
Zt	Height of terrain (m)
Kh	Horizontal mixing coefficient (m^2 s^{-2})

The dimension tables are related to the fact table by means of a foreign key. In both cases, the date of the data allows to relate the tables. The Julian date of the atmospheric conditions has been approximated to the Julian date of the radon measurements for those values that are not exact, so for a given Julian date of the atmospheric conditions there may be more than one radon measurement. The cubes have been created in a Microsoft Analysis Services server, each one on its own SQL Server 2005 database, as shown in Fig 2.a for cube 1 and Fig.2b for cube 2.

(a) OLAP Cube 1 (b) OLAP Cube 2

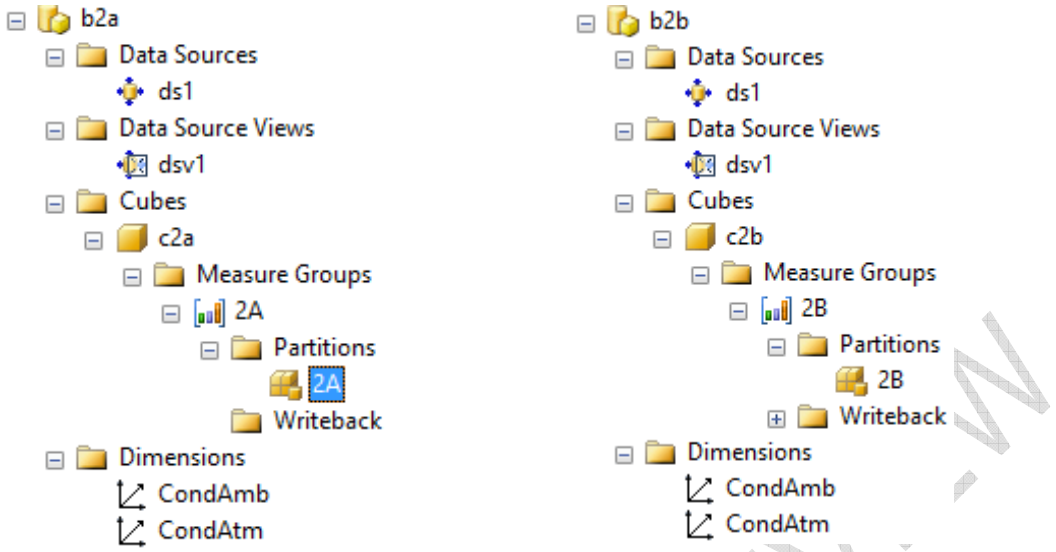

Figure 2. Structure of OLAP Cube 1 and Cube 2

The following hierarchies are defined in the cubes: environmental conditions, with levels of pressure and relative humidity; atmospheric conditions, with levels Zi, Kh and PSQ; and the time hierarchy, with year, month, day, day of the week and hour levels. The following MDX query runs on both cubes:

"select [CondAmb].[LEV_Pres].MEMBERS on ROWS, [CondAtm].[LEV_Zi].MEMBERS on COLUMNS from [c2a] where Measures.[Radon]"

An extract of the answers given to this query based on cube 1 can be seen in Table 4.

Table 4. Result of MDX query on Cube 1

[CondAmb].[HIE_CondAmb].[LEV_ Pres]. [MEMBER_CAPTION]	[CondAtm].[HIE_CondAtm].[LEV_Zi].& [5.004E1]
955	94
956	85

These values are the result of OLAP cube processing, which is explained in detail below. For the Zi value in the column (50.04), we obtain the ID that relates this tuple to the fact table (Table 5a). In the table of facts there are three tuples for that identifier value (Table 5b). The radon measurements corresponding to a pressure of 955 (Table 5c) are added together. In this case there is only one (ID 1387). Table 5. Spreadsheets confirming the result obtained from the MDX query in the table above

Table 5. Spreadsheets about (a) atmospheric conditions, (b) table of facts and (c) table of dimensions.

(a) Atmospheric conditions

ID	Zi	10xKz	U*	Kh	Zo	Zt5rr	Fechajulian	Clic
1839	50,02	0,34	0,01419	9,72	0,75	864,8	2457033,38	
1832	50,04	0,34	0,01419	9,72	0,75	864,8	2457032,50	

(b) Table of facts

ID	Radon	Error	CondAmb	CondAtm
2812	51,00	41,00	1385	18
2813	85,00	32,00	1386	18
2814	94,00	30,00	1387	18

(c) Table of dimensions

ID	RelHum	Temp	Pres
1384	46,00	12,00	957,0
1385	47,00	12,00	957,0
1386	47,00	12,00	956,0
1387	47,00	12,00	955,0

The result of this same query in cube 2 is shown in Table 6. The result can be obtained in the same way as in the previous case.

Table 6. Result of the MDX query on Cube 2

[CondAmb].[HIE_CondAmb].[LEV_Pres].[MEMBER_CAPTION]	[CondAtm].[HIE_CondAtm].[LEV_Zi].&[5.004E1]
955	113
956	123

3.1 Clustering algorithm

As indicated, the algorithm used is MicrosoftClustering. In the following calculation, the dimension corresponding to the environmental conditions CondAmb, the pressure attribute Pres, the key CondAmb.ID and the measure Radon have been used. Figure 3 shows the 10 clusters identified in cube 1.

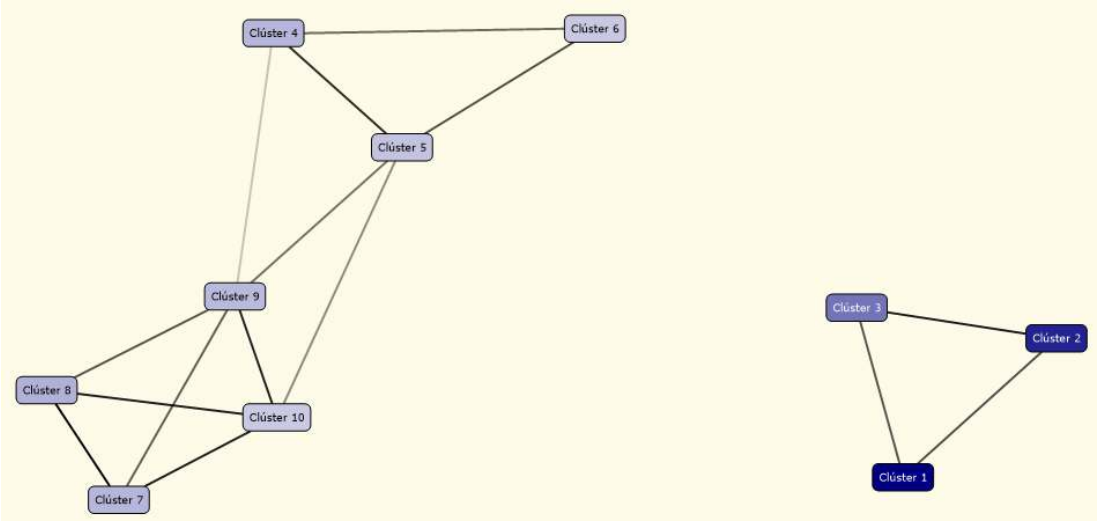

Figure 3. Diagram with the clusters obtained from Cube 1

As a result of this diagram, two well differentiated sets of clusters are detected. The set on the right (see Figure 2) is made up of clusters 1 to 3, highlighted in dark blue, which contain

most of the radon samples and show strong relationships between them (thick lines). This set could be considered as the most representative of the measured radon samples. If we consider the cluster with the highest number of members, cluster 1, the typical radon values are less than 45.8 Bq/m^3 for atmospheric pressure values between 929 and 946 hPa. On the other hand, the group formed by the rest of the clusters - from 4 to 10 - must be considered anomalous cases of radon for certain atmospheric pressures. Specifically, clusters 4 and 6 have a higher proportion of high radon concentration measurements than the rest (over 96 Bq/m^3 for pressures between 930 and 948 hPa).

Figure 4 shows the 10 clusters identified in cube 2. This diagram shows the preponderance of the group formed by clusters 1, 2 and 3, highlighted in the figure with a dark blue color, which are the most representative of the samples, and which correspond to radon concentrations below 54, 3 Bq/m^3 for atmospheric pressure values between 929 and 946 hPa. Radon values greater than 135.1 Bq/m^3 can be considered abnormal.

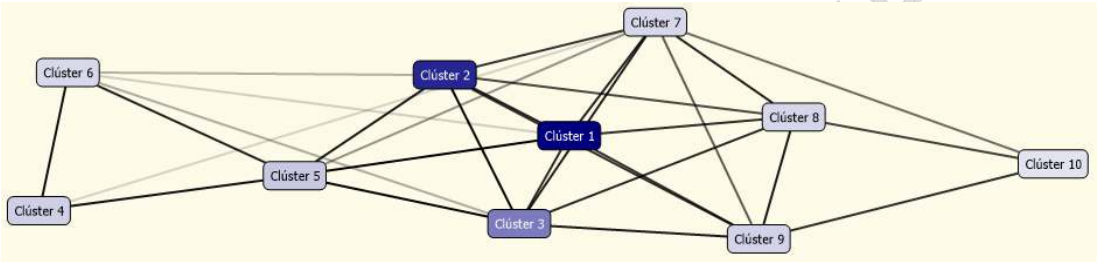

Figure 4. Diagram with the clusters obtained from Cube 2

Table 7 shows the data for the different clusters of both cubes, arranged from highest to lowest number of members.

Table 7. Clusters obtained from Cube 1 and 2

Dwelling 'A'			**Dwelling 'B'**		
Variables	**Values**	**Probability**	**Variables**	**Values**	**Probability**
Mean:	45,7	std. Desv. 74,5	Mean:	54,7	std. Desv. 119,9
Radon	45,8 - 96,0	24,980%	Radon	54,3 - 135,1	24,980%
Radon	96,0 - 269,4	24,885%	Radon	135,1 - 413,9	24,885%
Radon	0,0 - 45,8	23,020%	Radon	0,0 - 54,	17,444%
Pres	932	6,843%	Pres	932	6,843%
Pres	935	6,597%	Pres	935	6,597%
Pres	934	6,427%	Pres	934	6,427%
Pres	933	6,068%	Pres	933	6,068%
Cluster 1			Cluster 1		
Radon	0,0 - 45,8	50,000%	Radon	0,0 - 54,3	50,000%
Pres	934	10,129%	Pres	934	9,597%
Pres	929	9,425%	Pres	929	9,514%
Pres	940	6,706%	Pres	932	7,335%
Cluster 2			Cluster 2		
Radon	0,0 - 45,8	50,000%	Radon	0,0 - 54,3	50,000%
Pres	935	10,160%	Pres	935	9,952%
Pres	946	7,833%	Pres	946	7,789%
Pres	934	7,310%	Pres	947	7,439%
Cluster 3			Cluster 3		
Radon	0,0 - 45,8	50,000%	Radon	0,0 - 54,3	50,000%
Pres	941	8,943%	Pres	941	8,305%

Dwelling 'A'			Dwelling 'B'		
Variables	**Values**	**Probability**	**Variables**	**Values**	**Probability**
Pres	933	7,735%	Pres	933	7,622%
Pres	943	7,666%	Pres	943	7,496%
Cluster 8			Cluster 5		
Radon	45,8 - 96,0	79,352%	Radon	135,1 - 413,9	58,252%
Radon	0,0 - 45,8	19,744%	Radon	54,3 - 135,1	41,273%
Pres	929	9,194%	Pres	934	8,760%
Pres	935	8,646%	Pres	935	8,529%
Pres	933	8,595%	Pres	932	8,089%
Cluster 4			Cluster 4		
Radon	96,0 - 269,4	85,000%	Radon	135,1 - 413,9	86,772%
Radon	45,8 - 96,0	13,765%	Radon	54,3 - 135,1	12,181%
Pres	935	12,182%	Pres	933	11,420%
Pres	933	9,580%	Pres	935	10,315%
Pres	936	9,239%	Pres	932	9,908%
Cluster 7			Cluster 9		
Radon	45,8 - 96,0	73,840%	Radon	54,3 - 135,1	88,862%
Radon	0,0 - 45,8	25,879%	Pres	933	16,045%
Pres	932	15,211%	Pres	932	11,590%
Pres	933	9,751%	Radon	0,0 - 54,3	11,138%
Pres	934	7,889%	Pres	933	16,045%
Cluster 9			Cluster 8		
Radon	45,8 - 96,0	81,907%	Radon	54,3 - 135,1	97,651%
Pres	933	10,159%	Pres	936	14,602%
Radon	96,0 - 269,4	10,066%	Pres	935	10,146%
Pres	930	9,323%	Pres	932	7,777%
Cluster 5			Cluster 7		
Radon	96,0 - 269,4	61,970%	Radon	54,3 - 135,1	97,522%
Radon	45,8 - 96,0	35,142%	Pres	934	10,653%
Pres	932	13,456%	Pres	933	8,680%
Pres	938	8,113%	Pres	940	8,155%
Pres	929	8,091%			
Cluster 6			Cluster 6		
Radon	96,0 - 269,4	49,377%	Radon	135,1 - 413,9	37,479%
Pres	934	9,542%	Pres	943	7,436%
Pres	943	8,661%	Pres	933	6,599%
Pres	947	6,251%	Pres	947	6,346%
Pres	935	5,719%	Radon	54,3 - 135,1	4,984%
Cluster 10			Cluster 10		
Radon	45,8 - 96,0	82,970%	Radon	0,0 - 54,3	51,402%
Radon	0,0 - 45,8	12,092%	Radon	54,3 - 135,1	48,581%
Pres	936	9,321%	Pres	930	8,038%
Pres	935	8,154%	Pres	938	7,981%
Pres	930	8,067%	Pres	935	7,760%

4. CONCLUSIONS

This article describes the methodology for creating OLAP cubes from data obtained by means of a radon gas measurement campaign in homes. The use of an algorithm of clustering on OLAP cubes has allowed the study of anthropogenic and atmospheric pressure effects on radon concentrations for the same soil lithology and atmospheric conditions. As a result of this algorithm, it is possible to group radon measurements according to certain variables. In this case, the most common radon measurements have been identified for a given range of atmospheric pressure. It has also made it possible to group those radon

measurements that are less frequent and, therefore, can be considered as anomalies in the sampling campaign.

As shown in this article, MDX queries on OLAP cubes give satisfactory results. This type of query would also serve to analyse measurements in large-scale radon campaigns. This is the case for sampling campaigns aimed at producing radon risk maps. The great variability of radon, both at spatial and temporal scales, is well known, so it is necessary to have a suitable methodology to manage all this information. This type of consultation, which is easily carried out with such methodology, allows reports to be generated that meet the needs of the project or research at any time.

COMPETING INTERESTS DISCLAIMER:

Authors have declared that no competing interests exist. The products used for this research are commonly and predominantly use products in our area of research and country. There is absolutely no conflict of interest between the authors and producers of the products because we do not intend to use these products as an avenue for any litigation but for the advancement of knowledge. Also, the research was not funded by the producing company rather it was funded by personal efforts of the authors.

REFERENCES

1. WHO handbook on indoor radon: a public health perspective. Available: https://www.who.int/ionizing_radiation/env/9789241547673/en/ Accessed 1 november 2020.

2. Udovičić V., Filipović J., Dragić A., Banjanac R., Joković D., Maletić D., Grabež B, Veselinović N. Daily and seasonal radon variability in the underground low-background laboratory in Belgrade, Serbia. Radiation Protection Dosimetry. 2014:160(1-3), 62–64.

3. De Francesco S., Pascale Tommasone F., Cuoco E., Tedesco D. Indoor radon seasonal variability at different floors of buildings. Radiation Measurements. 2010:45(8), 928-934.

4. Porstendörfer J., Butterweck G., Reineking A. Daily variation of the radon concentration indoors and outdoors and the influence of meteorological parameters. Health physics. 1994: 67(3), 283–287.

5. Baciu A. Radon and thoron progeny concentration variability in relation to meteorological conditions at Bucharest (Romania). Journal of Environmental Radioactivity. 2005: 83(2), 171-189.

6. Woith H. Radon earthquake precursor: A short review. Eur. Phys. J. Spec. Top. 2015:24, 611–627.

7. García-Tobar J. Weather-dependent modelling of the indoor radon concentration in two dwellings using CONTAM. Indoor and Built Environment. 2019: 28 (10), 1341-1349.

8 García-Tobar. J. A Study of Radon Propagation in A Dwelling Using the CFD Modelling Capabilities of CONTAM. To Physics Journal. 2020:5, 72-79.

10. REHVA. Indoor Environment and Energy Efficiency in Schools - Guidebook number 13. Brussels: Federation of European Heating, Ventilation and Air-conditioning Associations (REHVA). 2010.

11 Laverge J., Van Den Bossche N., Heijmans N. et al. Energy saving potential and repercussions on indoor air quality of demand controlled residential ventilation strategies. Building and Environment. 2011:46-7,1497-1503.

12. Vasilyev A.V., Yarmoshenko I.V., Zhukovsky M.V.. Low air exchange rate causes high indoor radon concentration in energy-efficient buildings. Radiation Protection Dosimetry, 2015: 164(4), 601-605.

13. Derbez M., Berthineau B., Cochet V. et al. Indoor air quality and comfort in seven newly built, energy-efficient houses in France. Building and Environment. 2014:72, 173-187.

14. Paniagua J. Nueva sección HS 6 del Código Técnico de la Edificación: protección frente a la exposición al radon. Cercha revista de los aparejadores y arquitectos técnicos. 2020:143, 10-15. Spanish.

15. Maya J, Mohamadou L, Mbembe S, Likéné A, Mbembe B, Boubakari M. Radon Risks Assessment with the Covid-19 Lockdown Effects. Journal of Applied Mathematics and Physics. 2020:8,1402-1412.

16. Sainz-Fernandez C., Fernandez-Villar A., Fuente-Merino I., et al. The Spanish indoor Radon mapping strategy. Radiation Protection Dosimetry, 2014:162, 58-62.

17. Rolph G., Stein A. and Stunder B.. Real-time Environmental Applications and Display sYstem: READY. Environmental Modelling & Software. 2017: 95, 210-228.

18. Available on Web “MSDN, Microsoft Developer Network”: https://msdn.microsoft.com Accessed 1 november 2020.