International Journal of Advanced Smart Convergence Vol.4 No.1 93-98 (2015)
http://dx.doi.org/10.7236/IJASC.2015.4.1.93

IJASC 15-1-12

Correlation Analysis of Radon Levels using Cluster Algorithm

Myeong Hwan Oh*, Yong Gyu Jung*†, Min Soo Kang*, John Lee**

**Dept. of Medical IT Marketing, Eulji University, Korea*
**†Dept. of Medical IT Marketing, Eulji University, Korea*
ygjung@eulji.ac.kr
***CEO, Trulogic Pty Ltd, NSW, Australia*

Abstract

Recently, Radon has been gotten attention for problems of Nuclear Generating Station and a variety of nuclear. It is naturally arises that is accumulated in the interior through the soil with radioactive materials. People exposed to indoor a Radon increase the high risks of lung cancer. The data are consisted of regional Country, The Location, Average Radon pCi/L, Geo Mean and Geo S.D etc. The research is experimented using E-M algorithm. The research result appears to make a division of soil distance, regional and cluster. It requires in effort to minimize exposure to people who live in areas with high radon levels. A country must apprise to people about Radon risk and needs to work out measures plan.

Keywords: *Radon, Lung-cancer, E-M algorithm, Cluster*

.

1. INTRODUCTION

Recently, due to the government's "National Housing Radon survey", the interest of people about Radon is increased. National Institute of Environmental Research reports that the result which investigated national house radon level to target 6048 national house during winter indicates approximately 1100 house which exceed the multiuse facility recommendation 148 Bq/m^3 in the three months of February from December 2014.

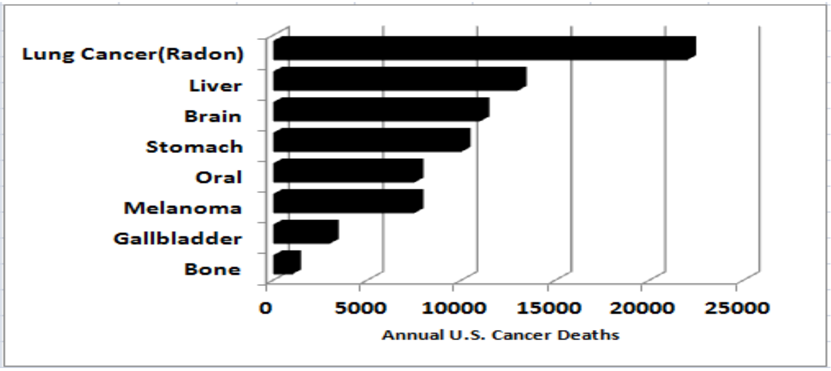

Figure 1. Incidence of radon-related lung cancer versus other types of cancer

Manuscript Received: Jan. 23, 2015 / Revised: Feb. 17, 2015 / Accepted: Mar. 22, 2015
Corresponding Author: ygjung@eulji.ac.kr
Tel: +82-31-740-7178
Dept. of Medical IT marketing, Eulji University, Korea

Radon that generated a process of several uranium decay in rock, a soil and a construction materials is a radioactive substance that exists a gas of colorless, scentless and tasteless anywhere in the earth. Although not well known in our country, all over the world is operating the radon management standard in interior. in the United State, it is configured 4pCi/L in repaired building. Environmental Protection Agency(EPA) concluded that radon in indoor air is the second leading cause of lung cancer in the U.S after cigarette smoking. The National Academy of Science(NAS) estimated that 15,000~22,000 American die every year from radon-related lung cancer. When people who smokes are exposed to radon as well.

Bulk of Indoor radon is entered though a crack of a wall or a floor. Other trace it inflows in building materials or in the ground water. the accumulated radon in an enclosed space causes problems which continually expose people to radioactive substance and they increase lung cancer incidence.

This paper experiments data though U.S. government's open data. The data which has the regional information investigates relations between region and radon levels. And the data which has the soil data investigates relations between soil distances and radon levels. it finds out the character of clusters through clustering of regional radon levels and effect of radon levels in the regional also suggested ways to reduce radon levels.

2. Related research

2.1 clustering

There is the clustering that appears to divide randomly several groups to receive mass data which analyzed the distribution character consisting of no labeling information. In other words, because information is not the labeling, the data with similar inputted the data are configured in the same cluster. In order to divide the cluster, the clustering is required a measure for indicating the degree of similarity between objects. The most commonly used way is the distance. The Euclidean distance of object i between object j defined as follows:

$$d(x_i, x_j) = \sqrt{\sum^{p}(X_{ai} - X_{aj})^2} = \sqrt{(x_i - x_j)^T (x_i - x_j)}$$

The data object of grouping differs the classification analyze that not include a state variable to indicate category. In other word, clustering is an unsupervised learning. It can be useful class labels are not given for data, as well as the number of samples costly to manually labeled with a class for each data so many.

2.2 E-M algorithm

E-M algorithm is a common learning method using to estimate parameter in variety of the probability models, not restricted to a Gaussian Mixture Model. it is especially used to estimate the parameters and hidden variable together in probability model with concealed variable and it has a discrimination feature compared to optimized algorithms. Also for a Gaussian mixture model, even if hidden variables in the model not exist originally, this can be expressed by applying the E-M algorithm by modifying a model with hidden variables.

The steps of the general E-M algorithm

① given data set {$x_1, x_2, \dots, x_n$}, the Probability model with

Hidden parameter that describes the model p(x, z| Θ) define.

② To set the initial values $\Theta^{(0)}$ of the parameters as desired.

③ [E-Step] t-th iteration step, the Q function is obtained by using a given parameter $\Theta^{(t)}$.

$$Q(\theta,\theta^{(r)}) = E_z[Inp(X,z|\theta] = \sum_z Inp(X,z|\theta)p(z|X,^{(r)})$$

④ [M-step] calculates a parameter $\Theta^{(r+1)}$ for maximizing the Q function obtained in the E-step.

$$\theta^{(r+1)} = \mathrm{arg}max_\theta Q(\theta,\theta^{(r)})$$

⑤ the parameters to be up or until the desired Q value is obtained when the convergence repeat E-step(③) and M-step(④).

3. Experiments

3.1 Experiments tool

For this experiment, WEKA is used as a tool developed at the University Wikato. It was developed as Java language and disseminate under general public license of GNU. Almost all platforms can be performed and it performs in Linux, window and Macintosh. It was designed to be the design of existing algorithms to target a new set of data from the user a quick and flexible way. It provides a wide range of support about the entire process of data mining experimentation such as input data preparation, statistical learning scheme evaluation, way to visualize input data and result data.

3.2 data set

It was used to Radon_Test_Result_By_Country_Beginning
_1987 provided by the home of U.S. government's open data.
This data has 128 instances and is consisted of Country, The location in home, Average radon pCi/L, GeoMean, Geo S.D and Highest(pCi/L). Attributes of Country having a country name and the location in home having a Basement or 1st Floor are varchar type. The rest of attribute is Numeric type.

Table 1. Data set for Experiment

Variable Name	Type	Variable Value
Country	Varchar	Country Name
The Location in home	Varchar	Basement or1st Floor
Average Radon Pci/L	Numeric	Continuous from 0 to 100
Geo Mean	Numeric	Continuous from 0 to 100
Geo S.D	Numeric	Continuous from 0 to 100
Highest (pCi/L)	Numeric	Continuous from 0 to 100
Number of Radon Tests <4 pCi/L	Numeric	Continuous from 0 to 100
Number of Radon Tests <20 pCi/L	Numeric	Continuous from 0 to 100
Location 1	Numeric	Continuous from -78 to 43

In order to measure more than accurate, the data having a null point and incorrect data that attribute have not equal variable values except the experiment.

3.3 Selection of clustering number

In order to select suitably, this experiment uses maximum likelihood function. It is how to obtain the parameters of the random variable that base on the values sampled from a random variable. When a parameter is given, it selects way to the parameters making up to maximum likelihood.

Table 2. Clustering log likelihood

The number	2	3	4	5	6
Log likelihood	-6.27	-6.18	-6.08	-6.01	-5.9

The result to analyzed data present Cluster2 to appear -6.27, Cluster3 to appear -6.18, Cluster4 to appear -6.08, Cluster5 to appear -6.01 and Cluster6 to appear -5.9. According to maximum likelihood function, clustering number is selected cluster6.

4. Experimental Result

Indoor radon enters the building through crack in the floor or wall from the soil. it through the soil layer is piled up the building interior of low pressure than the exterior. The radon is a United States based on the reference value that has been recommended 4pCi/L. The clustering result is as follows:

Table 3. Experimetal result of Clustering

Cluster number	Cluster instance
0	26(20%)
1	10(8%)
2	13(10%)
3	29(23%)
4	26(20%)
5	24(19%)

Cluster0 accounted for 20% having 26 instances. Clutser1 accounted for 8% having 10 instances. Cluster2 accounted for 10% having 13 instances. Cluster3 accounted for 23% having 29 instances. Cluster4 accounted for 20% having 26 instances. Cluster5 accounted for 19% having 24 instances.

As the result of the soil distance, the average of 1stFloor appears to 2.39 and the average of basement appears to 5.31. the lowest Radon levels area is 1stFloor of HAMILTON to appear 0.3 and the highest Radon level area is the basement of CORTLAND to appear 14.66.

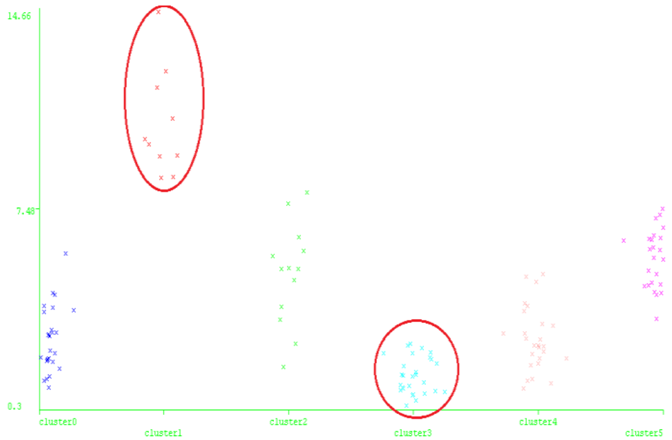

Figure 2. The graphic using Average Radon Pci/L

In the Figure 2, it indicates clusters using radon levels. Each clustering has the average and standard deviation. Cluster0 represents the average 2.8 and standard deviation 1.1. Cluster1 represents the average 10.1 and standard deviation 1.8. Cluster2 represents the average 5 and standard deviation 1.8. Cluster3 represents the average 1.3 and standard deviation 0.6. Cluster4 represents the average 2.8 and standard deviation 0.9. Cluster5 represents the average 5.7 and standard deviation 1.

Feature of Cluster1 and Cluster3 was the most prominent in clustering result. Cluster1 based on instances of basement consisting of average 10.1 has least 10 instances. Cluster3 based on instances of 1stFloor consisting of average 1.3 has the most instances.

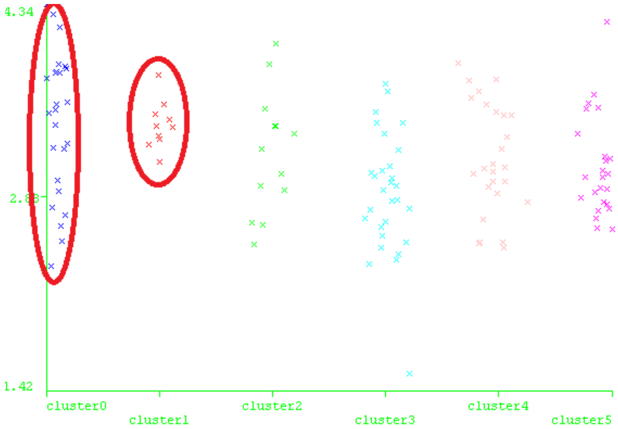

Figure 3. The graphic using Geo S.D

In the Figure 3, it indicates clusters using regional standard deviation. Feature of Cluster0 and Cluster1 was the most prominent in clustering result. Cluster0 has values of a standard deviation of the radon and most wide range. Cluster1 has values of a standard deviation of the radon and least extent.

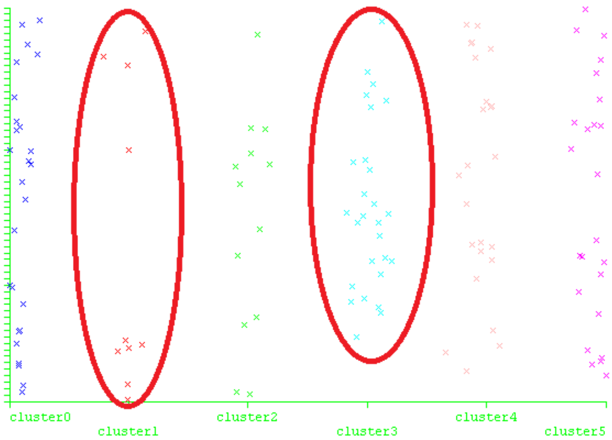

Figure 4. The graphic using Location

In the Figure 4, it indicates clusters using Location. Each cluster has the data relative the location. Compared to location, the relationship among data in Cluster1 showed the lowest. It indicates that the distance among regional data far. The relationship among data in cluster3 showed the highest. It indicates that the distance among regional data close.

5. Conclusions

Radon is inhaled to the body through breathing emit alpha rays causing the decay. Emitted alpha rays destroy lung tissues. A people continually exposed radon cause the lung cancer. The area of Cluster1 includes COTLAND, CHEMUNG, STEUBEN and etc. the area of Cluster3 includes LEWIS, FULTON, YATES and etc. Cluster1 and Cluster3 is the difference between average values of approximately 9.8. Cluster1 makes up the most area consisting of NEWYORK approximately 0.5% also Cluster3 includes the most area consisting of NEWYORK approximately 0.38. It proofs to different radon level according to soil or other circumstances in same area.

Result of the radon levels of soil, 1stfloor appear to the average 2.39 and basement appear to the average 5.31. it is a space of structure that is made structure in the soil. The basement surrounded by the soil exposures radon than 1stfloor. The basement make a window limited size or not make the window. The limited size window provides limited ventilation. The basement shorted ventilation so more increases the risk of lung cancer to exposure people. The insufficient ventilation of basement that exposures people radon increases the risk of lung cancer. Radon levels of basement appear approximately double more than radon levels and approximately 1pCi/L more than U.S recommendation.

To reduce indoor radon levels coming from the soil, it must frequently air out window for people of basement, and paint on crack using the reinforcement. Also the government created well-ventilated basement when made house and create a regulation that can minimize your exposure to radon.

References

[1] Jafari-Khouzani, Kourosh, and Hamid Soltanian-Zadeh, “Rotation-invariant multiresolution texture analysis using Radon and wavelet transforms,” *Image Processing, IEEE Transactions on*, Vol. 14, No. 6, pp. 783-795, 2005.

[2] Harris, Joyce M., et al., “Variations in atmospheric methane at Mauna Loa Observatory related to long-range transport,” *Journal of Geophysical Research: Atmospheres (1984–2012) 97.D5*, pp. 6003-6010, 1992.

[3] Sung A Kang, Donghyun Han, Chong-Yeal Kim, “A Study on the Correlation between the Volume of Indoor Space and the Measured Concentration of Indoor Radon,” *The Journal of Radiation Protection*, Vol. 32, No. 3, 2007.

[4] Hakl, J., et al., “Radon transport phenomena studied in karst caves-international experiences on radon levels and exposures,” *Radiation Measurements*, Vol. 28, No. 1, pp. 675-684, 1997.

[5] Ramola, R. C., et al., “A study of seasonal variations of radon levels in different types of houses,” *Journal of environmental radioactivity*, Vol. 39, No. 1, pp. 1-7, 1998.