Hausdorff dimension method reduces the error of box counting

Zhenyu Jin

Nanjing Foreign Language School, No.30, Beijing Road, Nanjing, Jiangsu, China

rkendall51653@student.napavalley.edu

Abstract. The Hausdorff dimension of an object is a topological measure of the size of its covering properties. To compute the Hausdorff dimension of an object, this report reviews the method of box counting, a method of gathering data for analyzing complex patterns by breaking a dataset, object, image, etc. into smaller and smaller pieces, typically "box"-shaped, and analyzing the pieces at each smaller scale. However, in this process, errors are always generated because of the extreme cases in which boxes are just covered with a small corner of the pattern but still should be counted. As a result, the number of boxes counted becomes larger than usual resulting in inaccurate results. In order to reduce the error in the final result, this report applies an improvement to this method to reduce the errors in the result and provides the example of Koch snowflake, and also examines the effect of the improvement.

Keywords: Hausdorff Dimension, Koch Snowflake, Box Counting.

1. Introduction

Both fractal geometry and dynamical systems have a long history of development [1]. Nowadays, commercial applications of deterministic fractal geometry have emerged in the areas of image compression, video compression, computer graphics, and education [2]. In the system, the Hausdorff dimension has its basic position. The theory of the Hausdorff dimension provides a general notion of the size of a set in a metric space [3]. With the Hausdorff dimension, people can define objects to be n-dimensional where n is not a natural number, but instead a nonnegative real number, which gives people further tools to analyze sets that are not interesting in integer dimensions [4]. To calculate the dimension of a set it is often important to understand its infinitesimal structure [5]. Box counting is a way to calculate the dimension of a pattern.

This paper introduces a way to reduce the error generated in this process by moving the pattern in one direction and counting several times in order to reduce the influence of extreme cases.

2. Introduction to box-counting dimensions

2.1. Derivation of formula

Define ε as the side length of each box, and N(ε) as the total number of boxes.

By observing the ε and $N(\varepsilon)$ of the image such formula can be obtained:

$$N(\varepsilon)\text{=c} \times (\frac{1}{c})^{D} \tag{1}$$

Therefor applying log to both sides:

$$\log N(\varepsilon) = \log[\text{c} \times (\frac{1}{c})^{D}] \tag{2}$$

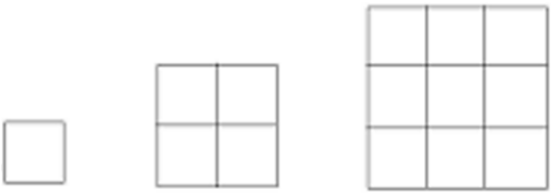

Figure 1 The boxes that will be put onto the pattern.

Hence, by covering an object with multiple boxes as shown in Figure 1 with side length ε, and counting the number of boxes, its limiting behaviour can be found using the equation [6]:

$$D = \lim_{\varepsilon \to 0} \frac{\log c - \log N(\varepsilon)}{\log(\varepsilon)} \tag{3}$$

Therefore, when scaling "ε" infinitely close to 0, there is:

$$D = lim_{\varepsilon \to 0} \frac{- \, log \, N(\varepsilon)}{log(\varepsilon)} \tag{4}$$

2.2. Computing the dimension with linear regression

The dimension can be counted by computing a graph with $log("\varepsilon"\,)$ on the x-axis and $log(N("\varepsilon"\,))$ on the y-axis. A line of best fit is expected there, and the gradient is the value of D [7].

$$logN(\varepsilon) = \log c + d \times \log\,(\varepsilon) \tag{5}$$

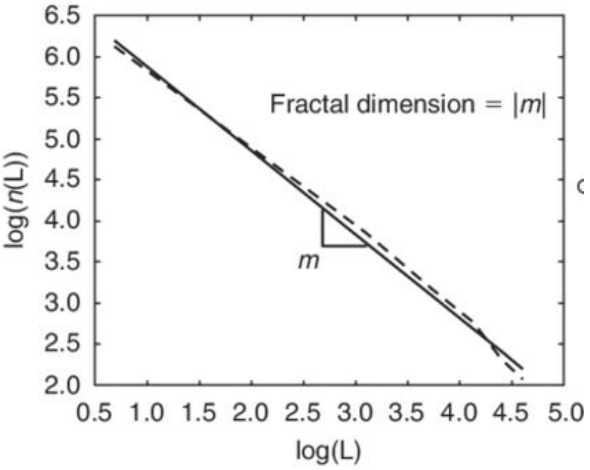

Figure 2. The dimension equals to the absolute value of the gradient of the line that best fit the two lists of $\log N(\varepsilon)$ and $\log(\varepsilon)$ [7].

3. The method of computing the box dimension D of a pattern with the example of the Koch snowflake (D=1.26) [8]

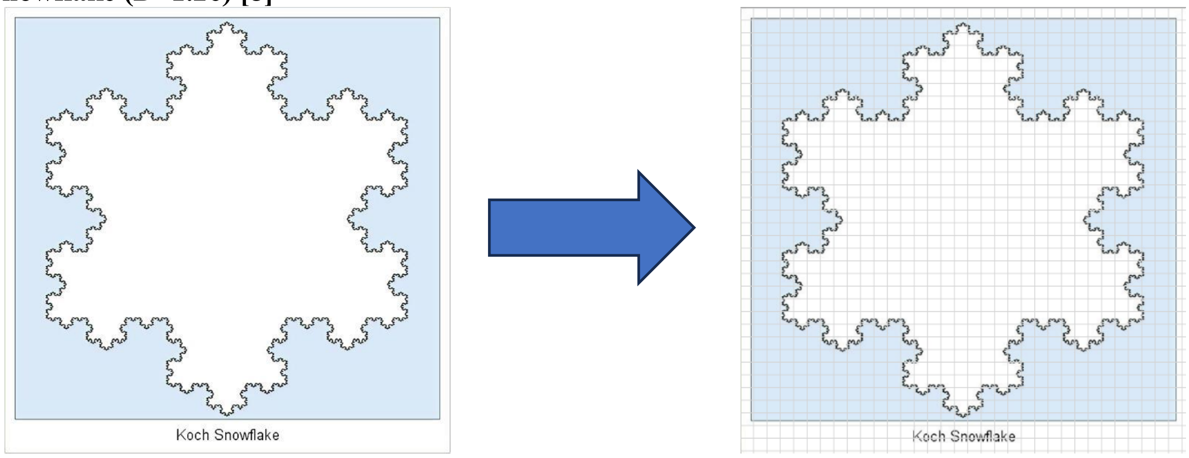

Figure 3. The process of covering the pattern with boxes.

3.1. Cover the image with boxes

As shown in figure 3, Koch Snowflake is made up of four ‘quarters’ each similar to the whole, but scaled by a factor 1/3 [9]. It can also be thought as a replacement construction, in which a line segment is replaced by an appropriately scaled copy of a polygonal curve [10]. After repeating this action infinite times on a straight line, Koch snowflake can be reached. The reason to choose the Koch snowflake is that many studies have been done to this pattern, so its dimension is easily accessible.

3.2. Repeat the previous step 9 times with different ε

This passage picked 8 ε values between the range of 1.715073 to 8.33. (Easier for the software to draw the picture)

Then cover the Koch snowflake with grids of boxes of corresponding ε values as shown in Figure 4.

This passage previously mentioned that when scaling "ε" infinitely close to 0, there is $\mathbf{D} = \lim\limits_{\boldsymbol{\varepsilon}\to\mathbf{0}} \frac{-\log \mathrm{N}(\varepsilon)}{\log(\varepsilon)}$.

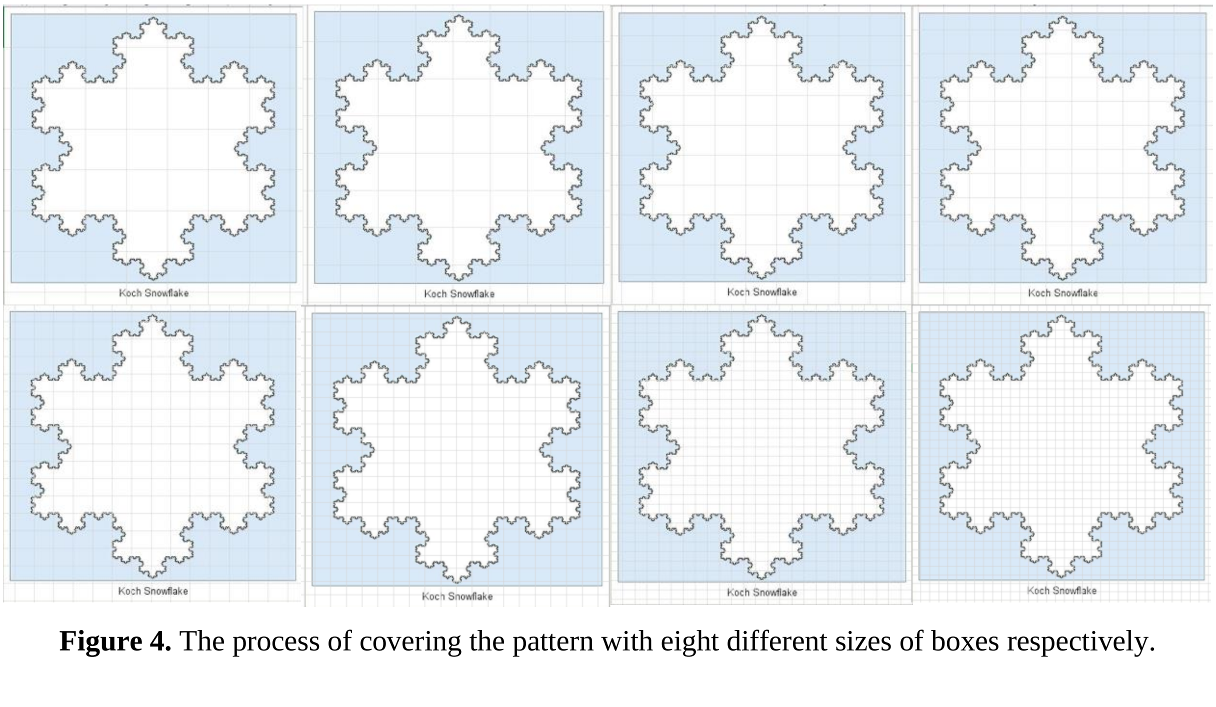

Figure 4. The process of covering the pattern with eight different sizes of boxes respectively.

Then apply log to each pair of ε and $N(\varepsilon)$ values, which are our data for later linear regression.

Table 1. The sizes of the boxes used and the numbers of the boxes covered by the Koch snowflake with the log of these two lists of values in order to calculate the dimension.

ε	$N(\varepsilon)$	$log(\varepsilon)$	$log(N(\varepsilon))$
8.33	28	0.920645	1.447158
7.497	41	0.874888	1.61278
6.07257	43	0.783373	1.633468
4.918782	56	0.691858	1.748188
3.585792	75	0.554585	1.875061
2.614042	102	0.417313	2.0086
1.905637	140	0.28004	2.146128
1.715073	217	0.234283	2.33646

3.3. Linear regression

With the log data shown in table 1, use software to do the linear regression. Based on the error function, the software presents the best fit line. The eventual gradient for the line is about -1.1005 as figure 5 shows, so the dimension of Hainan coastline is 1.1005.

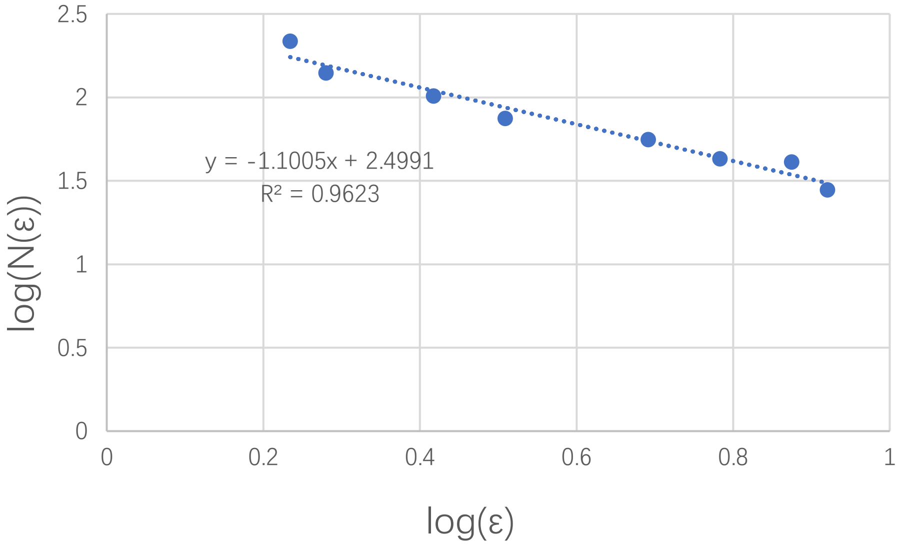

Figure 5. The best fit line for the two lists of log(N(ε)) and log (ε).

4. A way to improve the accuracy of computing the box dimension D of a pattern

4.1. Introduction to an improvement

$$Error: 1.26 - 1.1005 = 0.1595$$

$$Error\ percentage: \frac{0.1595}{1.26} = 12.659\% \tag{6}$$

As shown above, the error is quite significant. That is because when using box counting, some boxes containing only a tiny edge of the pattern are still counted. As a result, the error of box counting will be high. In order to reduce the error, this passage moves the pattern up and right together n times to move a whole box with boxes unmoved and then counts the number of boxes the pattern is in. Afterward, this passage uses the same method above to calculate the dimension.

This paper then apply this method to Koch snowflake ($D = 1.26$) the with $n = 10$ below, which means moving the pattern ten times to move the pattern upwards and rightwards respectively distance of ε. After every movement, this passage counts the number of boxes that are covered by the perimeter of the pattern and calculate the average number of boxes covered by adding all the ten numbers together and divide it by 10.

4.2. Application of an improvement

As shown in Table 2, the size of the boxes used which is the same as those in Table 1 above. Similarly, this table also shows the log of these to values in order to draw the graph of the linear regression.

Table 2. The average numbers of boxes covered by the Koch snowflake after ten times of counting.

ε	$N(\varepsilon)$	$log(\varepsilon)$	$log(N(\varepsilon))$
8.33	27.1	0.920645	1.432969
7.497	37.4	0.874888	1.572872
6.07257	44.6	0.783373	1.649335
4.918782	54.5	0.691858	1.736397
3.585792	72.9	0.554585	1.862728
2.614042	105.2	0.417313	2.022016
1.905637	144.6	0.28004	2.160168
1.715073	214.3	0.234283	2.331022

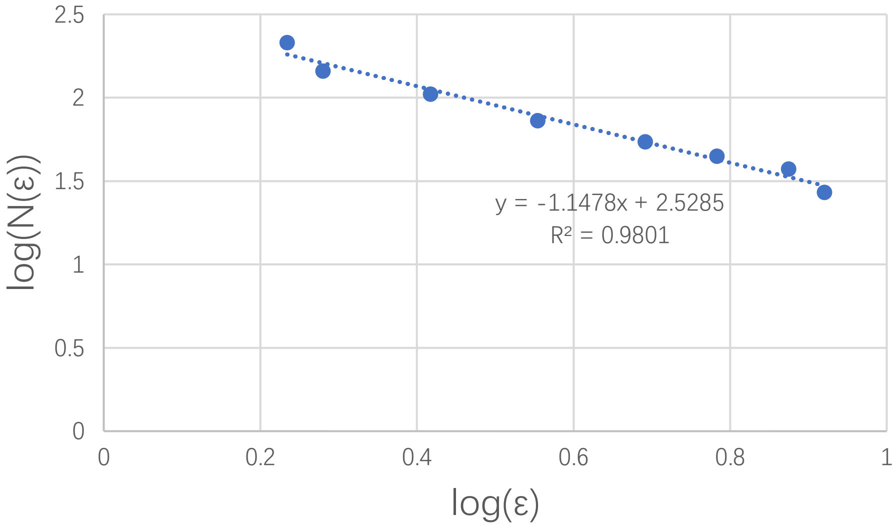

Figure 6. The best fit line for the two lists of $log(N(\varepsilon))$ and $log\ (\varepsilon)$ after the improvement.

The best fit line for the two lists of $log(N(\varepsilon))$ and $log\ (\varepsilon)$ after the improvement.

As shown in Figure 6, again use software to fine the best fit line. Based on the error function, the software presents the best fit line with the eventual gradient of about -1.1478.

So the dimension of Hainan coastline is 1.1478, the absolute value of -1.1478.

4.3. Improvement in the final result

Table 3. The error and percentage of error before and after the improvement respectively in order to compare the results to check the effect of the improvement.

	error	error percentage
Previous	0.1595	12.659%
After improvement	0.1122	8.90%

As Table 3 shows, the error reduces at about 0.05 and the percentage of error reduces at about 3.7% to the real dimension of the Koch snowflake which equals 1.26.

5. Conclusion

This research finds a way to reduce the error generated when using box counting to calculate the dimension of a pattern. This passage moves the pattern and count several times in order to reduce the influence of extreme cases. As the result shows, the error reduces at about 3.7% to the correct answer. Considering the size of boxes may be too large in this passage, the result could be closes to the correct answer if boxes are smaller. In a word, the improvement does show its effectiveness according to the result above and improves the result of box counting.

References

[1] Pesin Y B, Climenhaga V 2019 American Math. Soc.
[2] Barnsley M F 2014 Fractals everywhere Aca. Pre.
[3] Shah J 2009 Hausdorff dimension and its applications Manuscript.
[4] Rogers C A 1998 Hausdorff measures Camb. Uni. Press
[5] Fraser J, Howroyd D, and Käenmäki A 2019 P. Am. Math. Soc. 147 11 4921-4936
[6] Morgan A 2018 Uni. of Water.
[7] Zembrowska K, Kuźma M 2002 The Phy. Tea. 40 8 470-473
[8] Taylor S J 1986 Math. Proc. of the Camb.e Philosophical Soc. 100 3 383-406
[9] Falconer K 2014 Fractal geometry: mathematical foundations and applications John Wiley Sons,
[10] Baccelli F, Haji-Mirsadeghi M O, Khezeli A 2021 Elec. Jou. of Prob. 26 1-64