Journal of Probability and Statistical Science
24(2), 1-15Aug. 2026

Graph Fitting by Compartmentalization Method, Calculation of In-Set Differentiation, Propensity Score and Permutation Statistical Test

M. Yasin Pala
Istanbul Technical University, Turkey

ABSTRACT

By using points of sufficient density on the coordinate plane, any desired degree of graph can be drawn by dividing the coordinate plane into equal parts. Furthermore, whether a numerical set is divided within itself and whether the quantities formed by two numerical sets are separate from each other can be determined using averages without making any assumptions. Besides, robust p values in permutation statistical test can be obtained within large datasets through correction factor.

Keywords: Graph fitting, numerical unity, propensity score, permutation p value.

1. Introduction

Although many methods have been proposed for drawing graphs that fit given points, a comprehensive method for creating graphs of any desired degree using a sufficient number and density of points in the coordinate plane had not yet been developed. However, this is possible using the compartmentalization method, and the graph that fits best can be determined using the fitness coefficient.

In addition, a method for measuring how consistent a numerical set is within itself without using assumptions, and a formula for determining whether two numerical sets differ in terms of their quantities, had not yet been explained. These can be clarified using the in-set differentiation coefficient and propensity score, respectively.

Also, in permutation statistical test, calculation of robust and significant p values which are not equaling zero within large datasets has been problematic since many zero-p-values are useless and large datasets extremely decrease p value threshold. However, this problem can be overcome by a simple correction factor [1-8].

2. Theory

2.1. Compartmentalization Method on Graph Fitting

2.1.1. Multicompartmentalization Method

If a first-degree graph is to be created using the x and y points on the coordinate plane, the range of x values starting from the maximum x value on this plane and ending at the minimum x value is divided into only two equal parts. The arithmetic mean of the x and y values in each interval is calculated, and these mean values are evaluated as if they were the single point of each compartment. Then, these two points are connected to draw the first-degree graph. The equation of the graph is written using the general equation of first-degree graphs and the coordinates of the mean points.

Similarly, if a second-degree graph is to be constructed using the available points, the range of x values formed by these points is divided into only three equal parts. The arithmetic mean of the x and y values in each compartment is calculated. Thus, using the three points

- Received August 24, 2026, in final form August 2026.
- M. Yasin Pala (corresponding author) is affiliated with Istanbul Technical University, Turkey (graduate university).
  Email: palamu@itu.edu.tr

formed, the equation of the graph is written based on the general equation of quadratic graphs, and the graph is drawn.

If a third-degree graph is to be drawn, the x-value range is divided into only four equal parts. The graph and equation are obtained in the same way as described above, based on the average x and y values in each compartment. With this method, if there are sufficient points and point density, graphs can be drawn and equations can be established for all degrees. In this regard, when the coordinate plane is divided into compartments, if any point falls exactly on a boundary, the points should be distributed as evenly as possible among the compartments. If it does not matter which compartment the point falls in, it can be accepted on either side of the boundary.

2.1.2. Single Compartmentalization Method

As an alternative to the multicompartmentalization method, the first-degree graphs of the available points can be drawn using arithmetic and geometric means, considering the coordinate plane as a single part. Two new points (($x_{\text{arithmetic mean}}$, $y_{\text{arithmetic mean}}$), ($x_{\text{geometric mean}}$, $y_{\text{geometric mean}}$)) are created based solely on the arithmetic and geometric means of the x and y values, respectively. Then, as described above, a first-degree equation and graph are obtained from the new points. If the new points coincide, the slope is set to one and the graph is drawn. However, the alternative method differs from the previously described multicompartmentalization method because it uses the geometric mean. Therefore, it should be evaluated separately.

The degree to which the generated graph fits the given points will be calculated using the fitness coefficient. This coefficient is indicator of the fit between the points and the graph fitted to these points. It is calculated based on the initial x values, using the distance between each initial y value and each y value on the graph. These distances are processed based on the initial y values. The overlap of all initial and final y values indicates greater fit as the number of these y values increases. Hence, a term dependent on the number of y values is added to the formula to reflect this distinction. As the coefficients approach zero, the graph's fit to the initial points increases.

Compartmentalization Fitness Coefficient of Dot (CFCD) = ($n^{-0.25}$ + 10 $\sum_{i=1}^{n}(|(graph\ y_i - initial\ y_i)\ /\ initial\ y_i\ |)$)

Compartmentalization Fitness Coefficient of Whole (CFCW) = ($n^{-0.25}$ + 10 $\sum_{i=1}^{n}|(graph\ y_i - initial\ y_i)|\ /\ \sum_{i=1}^{n}|(initial\ y_i)|$)

where i is order of initial x, therefore order of y based on that x, n is the last order of initial x for any graph fitting approach since those are easily practicable to other methods. Fitness coefficients ought to be always considered simultaneously in order to rule out bias of infinites coming from null initial y_i.

2.2. Calculation of In-Set Differentiation

The in-set differentiation coefficient is an indicator of how consistent a numerical set is within itself, and consequently, whether it is numerically divided or not.

The differentiation coefficient, when considering absolute values, is directly proportional to the deviation of the numbers in the set from the mean and inversely proportional to the mean of the set [1, 2]. Therefore, as the coefficient approaches zero, the numerical integrity of the set will increase. As the coefficient increases, the harmony within the set will decrease, and after a certain value, the set will be numerically divided. In addition, since the overlap between the average and the numbers in the set will show greater numerical integrity as these numbers increase, a term dependent on the number of elements in the set is added to the formula.

In-Set Differentiation Geometric Coefficient (IDGC) = ($n^{-0.25}$ + 10 $\sum_{i=1}^{n}|(x_i - gm)\ /\ gm|$)

In-Set Differentiation Arithmetic Coefficient (IDAC) = ($n^{-0.25}$ + 10 $\sum_{i=1}^{n}|(x_i - am)\ /\ am|$)

Balanced Mean (bm) = $(am^a\ gm^b)^{1/(a+b)}$

In-Set Differentiation Balanced Coefficient (IDBC) = $(n^{-0.25} + 10 \sum_{i=1}^{n} |(x_i - bm) / bm|)$

Where i is order of x, a number in given set, and n is the last order of x. am and gm are arithmetic and geometric mean of x values, respectively. If gm equals zero, formulas containing it should not be utilized. All differentiation coefficients ought to be always considered simultaneously in order to rule out bias of infinites coming from null mean.

2.3. Calculation of Propensity Score Between Two Sets

Unlike conventional statistical tests, the propensity score should examine not the overall difference between numerical sets, but rather the differences in the quantities of the numerical sets. Hence, the propensity score is an indicator of the distance between the numerical sizes of the sets and should be calculated based on the means of the sets [1, 2].

The numerical difference between two sets depends on the absolute value of the difference between their means when their means are considered jointly. As the propensity score increases, the separation between the sets will increase.

Geometric Propensity Score (GPS) = $100 \left| (gm_1 - gm_2) / ((gm_1 + gm_2) / 2) \right|$, where gm is geometric mean for set 1 and 2.

Arithmetic Propensity Score (APS) = $100 \left| (am_1 - am_2) / ((am_1 + am_2) / 2) \right|$, where am is arithmetic mean for set 1 and 2.

Balanced Propensity Score (BPS) = $100 \left| (bm_1 - bm_2) / ((bm_1 + bm_2) / 2) \right|$, where bm is balanced mean for set 1 and 2.

The methods in question were applied to sample data sets, and the results are presented in the results section.

2.4. P Value Calculation in Permutation Statistical Test

Although p-values are undependable, in order to improve them, by avoiding easy and useless p values equaling zero in permutation test, a correction factor can be added to conventional p value formula. Within a dataset, as multiple testing count increases, significancy probability of some items increases as well, in that, proportion of insignificant items will increase.

On the other hand, as permutation count increases, a probability to unduly skip significant items increases, in that, excessive permutation count will eliminate significant items due to overpermuting.

Hence, in p value formula, the correction factor ought to be derived from multiple testing and permutation counts through reverse proportion:

$$p\ value = (e / n) + f_p$$

$f_p = (m\ n)^{-1}$, where e is extreme count, n is permutation count, f_p is the correction factor and m is multiple testing count.

3. Results

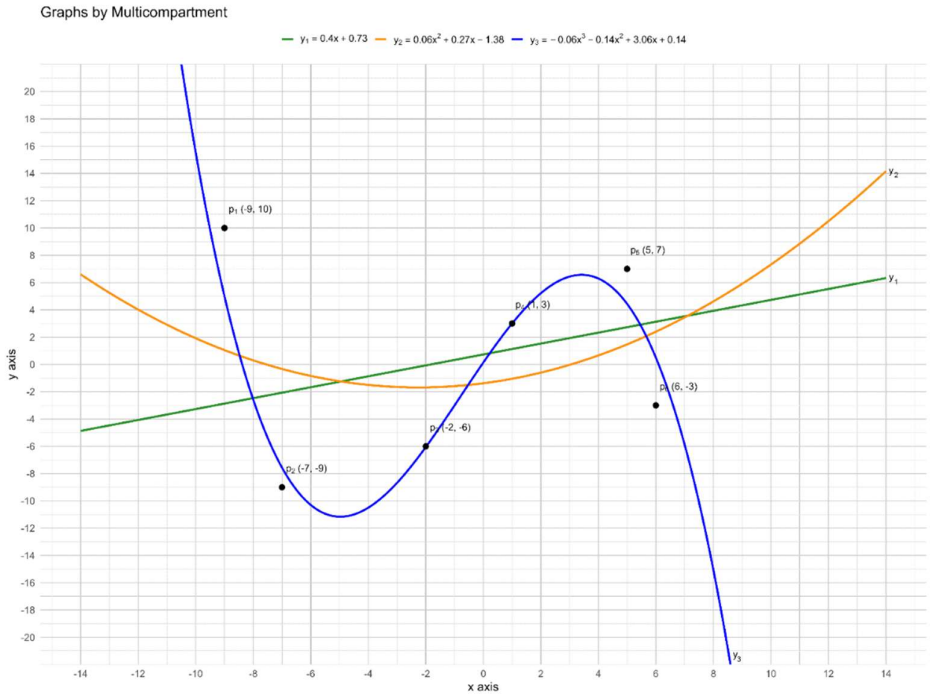


Figure 1: Fitting graphs of multicompartment. Equations at top

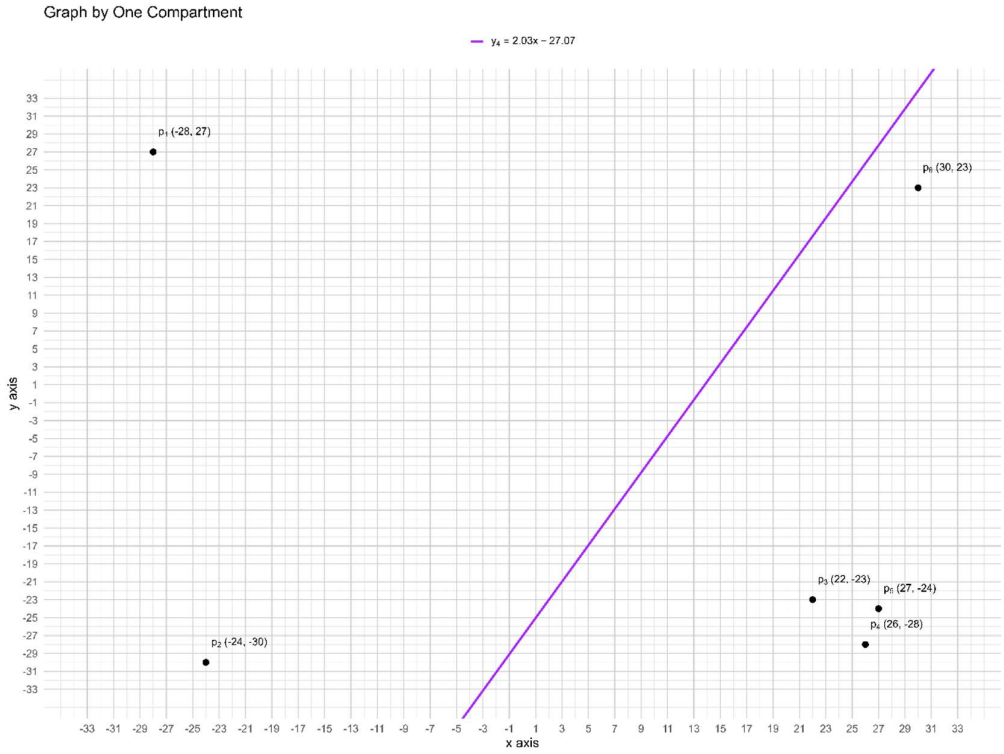


Figure 2: Fitting graph of single compartment. Equation at top

Table 1: Characteristics of graph fittings, proposed sets of disunity and unity by inches. A) Fitness coefficients by graph fittings. B) Set elements, related in-set differentiation coefficients and propensity scores.

A)	CFCD	CFCW	Equation	
Multicompartmentalization:				
1. Degree	63.86	10.64	y_1	
2. Degree	65.99	10.34	y_2	
3. Degree	18.44	3.28	y_3	
Single Compartmentalization	119.89	20.84	y_4	
B)				
Set of Disunity by Inches:	IDGC	IDAC	bm (a=b=1)	IDBC
1, 2, 3, 4, 5, 10, 11, 12, 13, 14	81.35	60.56	6.46	70.22
Set of Unity by Inches:	75.61	57.70	6.11	66.03
1, 2, 3, 4, 5, 9, 10, 11, 12, 13				
GPS= 4.40, APS= 6.90, BPS= 5.57				

Upon clearly seen in the tables and graphs, as the fit graphs approach the points given on the coordinate plane, the fitting coefficients approach zero. The in-set differentiation coefficients increase as the separation within the sets increase. Propensity scores, on the other hand, are low because the characteristics of the sets are very similar to each other.

As for permutation statistical test, in a case that m equals five, and, p value threshold 0.05/5=0.01; when n equals ten, the lowest possible p value equals $0 + 50^{-1} = 0.02$. Hence, there is no significant p value possibility.

Yet, when n equals a hundred, the lowest possible p value equals $0 + 500^{-1} = 0.002$. Thus, significant p value possibility emerges. Said implementation can easily expand to datasets with much higher m.

4. Discussion

The multicompartmentalization method is a graph fitting method that can be easily applied if the distance between consecutive points along the x-axis is not high and sufficient points are present. The single compartmentalization method can be applied to all points within certain mathematical limits, regardless of distance. With these methods, the most suitable exponential graph can be fitted to the given points without resorting to exponential or logarithmic equations.

Whether a numerical set is divided or not can be determined using the coefficients of a reference set. The set of disunity by inches is a set that is well suited for reference. The thresholds IDGC≥80 and IDAC≥60 should be used. When a=b and the degree of the root expression is 0.5, the threshold IDBC≥70 should be used.

Propensity scores, on the other hand, have proven to yield highly consistent results by yielding very low values between two sets with very similar characteristics. However, due to its nature, the propensity score measures the separation of one entire set's quantity from another entire set's quantity, thus limiting its statistical applicability. Therefore, other methods should be used rather than propensity scores.

Permutation statistical test including subject correction factor does facilitate significant p value calculations, particularly in large datasets.

5. Conclusions

Instead of the assumption-based and problematic p-value calculations which are inconsistent with actuality and within themselves, even when same datasets are considered [4-7, 9, 10], the easier and comprehensive method of the in-set differentiation coefficient should be used. All the methods should be employed in graph fitting and statistical tests to enhance the consistency and validity of scientific studies.

Declarations

All content is available to be used citing exactly the paper and the author has no competing interest.

References

[1]. Roxy Peck, Chris Olsen and Jay Devore. (2008). *Introduction to Statistics and Data Analysis*. 3rd Edition. Thomson Brooks/Cole. Belmont, California.

[2]. Amanda Ross and Victor L. Willson. (2017). *Basic and Advanced Statistical Tests*. Sense Publishers. Rotterdam, the Netherlands.

[3]. M. Yasin Pala. (2025). Assessment of Genetic Distance Measurements to Shed Light on Population Histories. *International Journal for Research Trends in Social Science & Humanities.* 3(4), 1-9. DOI: http://doi.org/10.5281/zenodo.17077331

[4]. Rudolf Gaško. (2003). Statistical hypothesis testing--how exact are exact p-values? *Bratislavske Lekarske Listy*. 104(1), 36-9.

[5]. Sander Greenland, Stephen J. Senn, Kenneth J. Rothman et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol.* 31, 337–350. DOI 10.1007/s10654-016-0149-3

[6]. Joel J. Gagnier and Hal Morgenstern. (2017). Misconceptions, Misuses, and Misinterpretations of P Values and Significance Testing. *J Bone Joint Surg Am.* 99,1598-1603.

[7]. Sang-June Park and Youjae Yi. June (2023). The Overall t-test Investigating a Conflict between the Individual t-test and the Overall F-test in Regression Analysis. *Seoul Journal of Business*, 29(1), 1-24.

[8]. Belinda Phipson and Gordon K. Smyth. (2010). Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology*. 9(1). Article 39. DOI:10.2202/1544-6115.1585

[9]. Yudi Pawitan. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Clarendon Press. Oxford, England.

[10]. Ronald N. Forthofer, Eun Sul Lee and Mike Hernandez. (2007). *Biostatistics: A Guide to Design, Analysis, and Discovery*. pp: 103-133. Elsevier Academic Press. MA, USA. ISBN 13: 978-0-12-369492-8. ISBN 10: 0-12-369492-2