CHAPTER 38

DIMENSIONAL ANALYSIS - SPURIOUS CORRELATION

by

M S Yalin
Professor, Department of Civil Engineering,

and

J W Kamphuis
Associate Professor, Department of Civil Engineering,
Queen's University at Kingston, Canada

ABSTRACT

Dimensional analysis and errors leading to spurious correlation are presented using a number of practical examples

INTRODUCTION

The theory of dimensions has been subject of considerable controversy and only in the last few decades has it been agreed that dimensional methods can serve as a powerful tool in experimental investigations of physical phenomena Its effectiveness becomes especially noticeable when the number of factors involved is large and the theoretical knowledge is insufficient

However, like any other mathematical tool, the theory of dimensions can supply correct and useful results only if it is properly applied One example of improper application is the introduction of an exaggerated and sometimes simply of a nonexistent correlation between the experimental results The purpose of the present paper is to analyse these kinds of spurious correlation

OUTLINE OF GENERAL PRINCIPLES

For detailed information on the principles of the theory of dimensions the reader is referred to Refs (1), (2), (3), here only those points will be mentioned which have a direct bearing on the present analysis

In general, a physical phenomenon corresponding to a specified geometry can be described by a number of independent quantities a_1, a_2 a_n referred to as the characteristic parameters Accordingly any quantitative property b of the phenomenon under consideration must be given by a functional relation such as

$$b = f(a_1, a_2, \qquad a_n) \qquad (1)$$

where the quantities a_i and b are usually dimensional Since all physical relations must be dimensionally homogeneous, using the Π theorem of the theory of dimensions, one can express the above relation in a dimensionless form as

$$Y = \Phi\ (X_1, X_2, \quad X_{n-k}) \tag{2}$$

Here, k is the number of fundamental units, while Y and X_j ($j = 1, 2, \quad n-k$) are the following dimensionless power products

$$Y = b \prod_{i=1}^{k} a_i^{\beta_i} \tag{3}$$

$$X = a_j \prod_{i=1}^{k} a_i^{\alpha_i} \tag{4}$$

The dimensionless quantities Y and X_j can be interpreted as the dimensionless versions of b and a_j respectively The k parameters $a_i (i = 1 \quad k)$ which are common in the expressions of X_j and Y are often called the basic quantities or the "repeaters"

Although either system may be used in experimental work, Eq 2 has some definite advantages over Eq 1 and these are described in standard works on the theory of dimensions, e g Refs (1, 2, 3) The most important advantages are -

a) Eq 2 is independent of the system of units used,
b) the number of variables on the right hand side of the equation has been reduced by k,
c) the dimensionless variables (X_j) are criteria of similarity

VARIATION OF THE DIMENSIONLESS CHARACTERISTICS, SPURIOUS CORRELATION

In a correctly designed experimental investigation, the variation of Y is achieved by varying only one X_j at a time In such cases, one deals with a set of functions of only one variable, as below

$$Y_j = \Phi\ (\text{Const}_1, \text{const}_2, \quad X_j, \quad , \text{const}_n) = \phi_j (X_j) \tag{5}$$

Accordingly in the following, the consideration of Eq 2 will be replaced by its special case, Eq 5, while the subscript j will be omitted

Differentiating Eqs 3 and 4, one arrives at the following relations which represent the most general versions of variation of Y and X_j

$$dY = b \prod_{i=1}^{k} a_i^{\beta_i} \left[\frac{db}{b} + \sum_{i=1}^{k} \beta_i \frac{da_i}{a_i}\right] \tag{6}$$

$$\frac{dY}{Y} = d(\ln Y) = d(\ln b) + \sum_{i=1}^{k} \beta_i\ d(\ln a_i) \tag{7}$$

$$dX = a \prod_{i=1}^{k} a_i^{\alpha_i} \left[\frac{da}{a} + \sum_{i=1}^{k} \alpha_i \frac{da_i}{a_i}\right] \tag{8}$$

$$\frac{dX}{X} = d(\ln X) = d(\ln a) + \sum_{i=1}^{k} \alpha_i \, d(\ln a_i) \qquad (9)$$

Three typical methods of variation of Y and X_j are discussed below

Case I Variation of Y and X is achieved by varying b and a only and by keeping the basic quantities a_i constant Substituting $da_i = 0$ into Eqs 6 and 8, one arrives at

$$\frac{dY}{dX} = C\,\frac{db}{da} = C'\,\frac{db}{dX} \qquad (10)$$

where

$$C = \frac{\prod_{i=1}^{k} a_i^{\beta_i}}{\prod_{i=1}^{k} a_i^{\alpha_i}} \quad \text{and } C' = \prod_{i=1}^{k} a_i^{\beta_i} \qquad (11)$$

Hence, the variation of Y with X is proportional to the variation of b with a (and of b with X) and therefore the dimensionless relationship between Y and X is merely a scaled down version of the dimensional relationship between b and a (The scale of the ordinate is C times the scale of the abscissa)

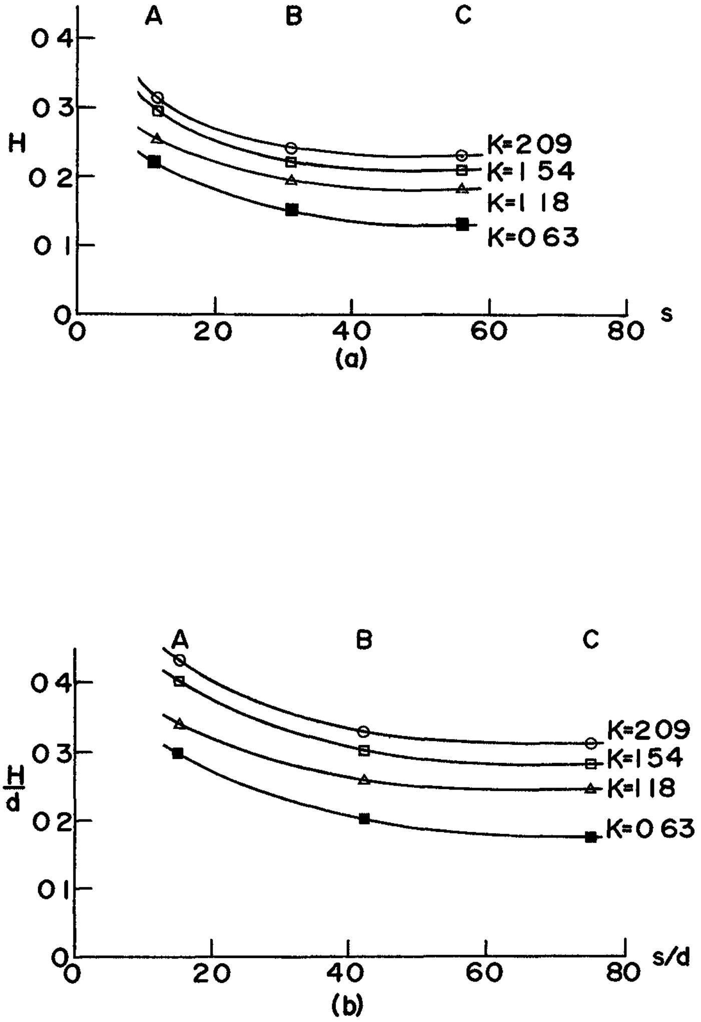
For example, consider the decay of an impulsively generated wave with distance (4) The wave height H (which should be identified with b) may be plotted against the distance, s (to be identified with a) as in Fig 1a On the other hand, one can plot the dimensionless quantities $Y = \frac{H}{h}$ and $X = \frac{S}{h}$ against each other - Fig 1b Since the common parameter h, the water depth (which is to be identified with one of a_i) has not been varied, Figs 1a and 1b are merely scaled down versions of each other For the present example C = 1 and thus the curves in Figs 1a and 1b are geometrically similar From Eqs 7 and 9 it follows that

$$\frac{d(\ln Y)}{d(\ln X)} = \frac{d(\ln b)}{d(\ln a)} = \frac{d(\ln b)}{d(\ln X)} \qquad (12)$$

is also valid, which implies that, in the case under consideration, the variation of Y and X in a logarithmic system is identical to that of b and a (or of b and X)

Case II Variation of Y and X is achieved by keeping the parameter a constant and by varying b and one or more of basic quantities a_i Substituting da = 0 into Eqs 7 and 9 one obtains

$$\frac{d(\ln Y)}{d(\ln X)} = \frac{d(\ln b)}{d(\ln X)} + m \qquad (13)$$



(a)

(b)

Fig 1 Dimensional and Dimensionless Plots

where

$$m = \frac{\sum\limits_{i=1}^{k} \beta_i \frac{da_i}{a_i}}{\sum\limits_{i=1}^{k} \alpha_i \frac{da_i}{a_i}} = \frac{\sum\limits_{i=1}^{k} \beta_i \, d(\ln a_i)}{\sum\limits_{i=1}^{k} \alpha_i \, d(\ln a_i)} \tag{14}$$

If only one of the basic quantities, e g a_1 is varied, then the expression for m reduces to

$$m = \frac{\beta_1}{\alpha_1} \tag{15}$$

The present case is very common in practice

Consider, for instance, the familiar plot of the drag force acting on a sphere in an infinite fluid flow (Fig 2) The dimensional relationship - analogous to Eq 1 - is

$$F = f(U, D, \rho, \mu) \tag{16}$$

where F is the drag force, U the velocity of the undisturbed flow, D, the sphere diameter, ρ the fluid density and μ the dynamic viscosity The quantities F, μ, U, D and ρ must be identified with b,a, a_1, a_2 and a_3 of Eq 1 respectively

The dimensionless form - analogous to Eq 2 - is

$$Y = \frac{F}{U^2 D^2 \rho} = \phi\ (X) = \phi\ (\frac{UD\rho}{\mu}) \tag{17}$$

Note that Eq 16 contains four variables whereas Eq 17 contains only one, i e the problem has been simplified considerably by introducing the dimensionless form! Let us now assume that the variation in flow is achieved by varying only one basic quantity (common parameter) U The variation of U alone will certainly induce the variation of the drag force Γ and the experimental procedure carried out in this manner will form an example for the present case The experimental relation between Y and X shown in Fig 2 indicates that

$$\frac{d(\ln Y)}{d(\ln X)} = -1$$

is valid On the other hand from the expression for Y and X (Eq 17) it follows that m = -2 Accordingly, Eq 13 gives

$$\frac{d(\ln b)}{d(\ln X)} = \frac{d(\ln Y)}{d(\ln X)} - m = 1$$

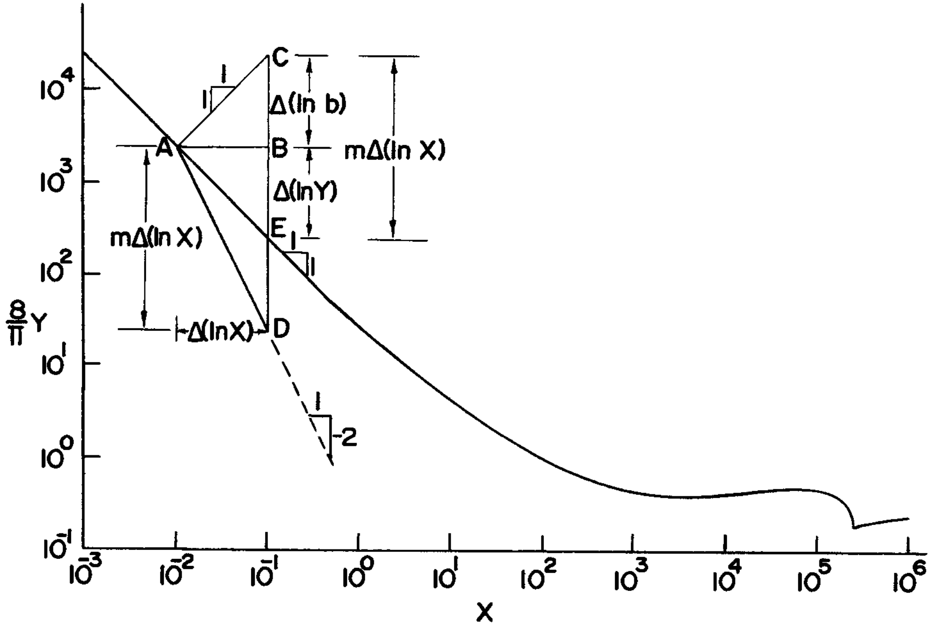


Fig 2 Drag Force on a Sphere

and in terms of finite differences

$$\Delta\,(\ln Y) = \Delta\,(\ln b) + m\,\Delta\,(\ln X)$$

In Fig 2 the increment $\Delta(\ln X)$ along the abscissa is shown by the horizontal distance $\overline{AB}$, the decrement $\Delta(\ln Y)$ along the ordinate being $\overline{BE}$ In Fig 2 the decrement $\overline{BE}$ is shown as the algebraic sum of the increment $\Delta(\ln b) = \overline{CB}$ and the decrement $m\Delta\,(\ln X) = \overline{BD} = \overline{CE}$, i e as implied by the equation above It follows, that the correlation between Y and X, achieved by varying a_i (and consequently b) is as legitimate as that achieved by varying a (and b) as in the previous case The difference is that in the case I the variation $\overline{AE}$ is achieved directly, whereas in the present case it yields itself as the "resultant" of the "component variations" $\overline{AC}$ and $\overline{AD}$

Note, however, that the variation of Y with X can characterise the variation of b with a only, as has been shown in the case I It cannot characterise the variation of b with any of the common parameters a_i Observe from Fig 2 that the value of Y decreases (along $\overline{AE}$ with X, whereas the value of b increases (along $\overline{AC}$) with a_i = U

Case III Variation of Y is achieved without varying b Substituting $d(\ln b) = 0$ into Eqs 7 and 9, one obtains

$$\frac{d(\ln Y)}{d(\ln X)} = \frac{m}{1 + \dfrac{d(\ln a)}{\sum_{i=1}^{k} \alpha_i \ d(\ln a_i)}} \qquad (18)$$

and if a is kept constant,

$$\frac{d(\ln Y)}{d(\ln X)} = m \qquad (19)$$

Eq 19 is simply Eq 13 without the first term which reflects the influence of the variation of b, the dimensional term responsible for the existence of Y Hence a correlation between Y and X, obtained as described above is completely spurious Indeed the correlation implied by the present case can only be due to the variation of the common quantities a_i If the parameters a_i vary only, then the quantities Y and X also vary, and Y can be plotted against X in the form of a curve, even if b in actual fact does not depend on a at all With reference to Fig 2, the curve of spurious correlation is the line $\overline{AD}$ which implies the following case of Eq 19

$$\frac{d(\ln Y)}{d(\ln X)} = -2$$

and which is due to the variation of the common quality a_i = U alone Hence, even if F had not been a function of μ at all, by applying the procedure described in the present case, one could still obtain a correlation between Y and X, in the form of the line $\overline{AD}$, which would, of course, be completely spurious

Integrating Eq 19, one obtains

$$Y = CX^m \tag{20}$$

which is a straight line in the logarithmic system of co-ordinates and where C is given by

$$C = \frac{b}{a^m} \prod_{i=2}^{k} a_i^{(\beta - m\alpha_i)} \tag{21}$$

It follows that

a) any correlation between Y and X obtained without varying b is spurious (unless a special subset of Eq 16 is chosen, where U, D, ρ and μ are not independent, thus violating the basic assumptions),

b) in a log-log system of co-ordinates, this spurious correlation forms a straight line,

c) the position and slope of the straight line of spurious correlation is, a priori, predictable, it is determined by Eqs 19 and 20

When analysing experimental plots, in case of doubt it is advisable to determine and plot the family of straight lines S_1 of spurious correlation (Fig 3) and to check the trend of the experimental points accordingly If the dimensional analysis has been performed improperly and the original quantities are not truly independent, then the variation of a_1 brings about unexpected variation in a or b and spurious correlation will take place along a curve Γ (Fig 3) This is also a point to be kept in mind

A special case occurs when experimentally

$$Y = \phi\,(X) = C\,X^m \tag{22}$$

and the lines of spurious correlation and genuine correlation coincide

Consider, for example, the length $\Lambda = b$ of dunes forming on the bed of a two dimensional rough turbulent flow with a mobile bed According to (5), this length may be expressed by the following functional relation

$$\Lambda = f\,(D, \rho, v_*, h) \tag{23}$$

where $D = a_1$ is the grain size, $\rho = a_2$ is the fluid density, $v_* = a_3$ is the shear velocity and h = a is the flow depth

The relation above can be expressed in dimensionless form

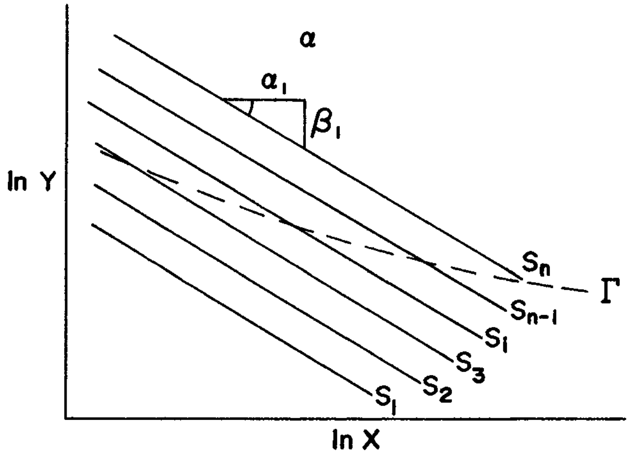


Fig 3 Lines of Spurious Correlation

as follows

$$Y = \frac{\Lambda}{D} = \phi(X) = \phi\left[\frac{h}{D}\right] \tag{24}$$

If the grain size $D = a_1$ is varied only, then Eqs 20 and 21 give

$$Y = CX \tag{25}$$

which implies that the lines of spurious correlation in a log-log system are straight lines with slope $m = 1$ On the other hand, the experimental results (obtained from experiments where $\Lambda = b$ is varied over a large range) indicate that the genuine correlation between Y and X is given by the linear relation

$$Y = 2\pi \quad X \tag{26}$$

which also a straight line with the slope $m = 1$

The lines of spurious correlation and the real relationship can coincide as in example above only if the parameter a_1 is a "spurious parameter" This can be deduced from Eqs 3, 4 and 20 and indeed, the relation above which can be expressed as

$$\Lambda \simeq 2\pi h \tag{27}$$

indicates (contrary to the theoretical expectation) that the dune length Λ actually does not depend on the grain size $a_1 = D$

EXAGGERATION OF THE ACTUAL CORRELATION

In practical applications, an existing correlation is often exaggerated by plotting common quantities along both co-ordinate axes (6) The explanation is given here in terms of a dimensionless system, however, the same argument is true for dimensional quantities

Consider the functional relation

$$Z = \psi(X) \tag{28}$$

where Z is the product

$$Z = Y\,X^N \tag{29}$$

and where Y implies

$$Y = \phi(X) \tag{30}$$

From Eq 28 one obtains

$$d(\ln Z) = d\,(\ln Y) + N\,d(\ln X)$$

and thus

$$\frac{d\ (\ln Z)}{d\ (\ln X)} = \frac{d(\ln Y)}{d(\ln X)} + N \qquad (31)$$

This equation indicates how in a log-log system of coordinates the rate of change with X can be increased by plotting the product $Z = YX^N$ rather than simply Y versus X

For example if

$$\frac{d(\ln Y)}{d(\ln X)} = \tan \theta = 1$$

then the plot Y vs X is a 45° straight line, as shown in Fig 4 a It is assumed that the experimental points forming this straight line are scattered in a "ribbon" of the thickness w_a

If the product $Z = Y X^N$ is used as ordinate, and if for example the value of N is 2, then from Eq 30

$$\frac{d(\ln Z)}{d(\ln X)} = 1 + 2 = 3$$

is valid, which implies that the straight line becomes steeper by a factor 3, while the thickness of the scatter-ribbon decreases by a factor

$$\frac{w_b}{w_a} = \sqrt{\frac{1 + N^2 \tan^2 O}{1 + \tan^2 O}} = \sqrt{\frac{5}{2}} = 1\ 58$$

(Fig 4b) Thus the relative thickness of the scatter ribbon, i e

$$\frac{\text{thickness of scatter-ribbon}}{\text{length of the line}}$$

is reduced by factor

$$\frac{1 + N^2 \tan^2 O}{1 + \tan^2 O} = \frac{5}{2} = 2\ 5$$

One can say that the correlation can be improved 2 5 times by plotting $Z = YX^2$ rather than simply Y against X Such an "improvement" is nothing else but an optical illusion and is not legitimate, for the dimensionless version of the quantity b under investigation, as supplied by the Π - theorem, is Y, not YX^N These kinds of illegitimate plots are recognisable by the presence of the parameter a (which should be present only in the expression of X) in the power product implying the ordinate

SUMMARY

The experimenter is free to choose between a dimensional representation, Eq 1, or a dimensionless representation, Eq 2 of the

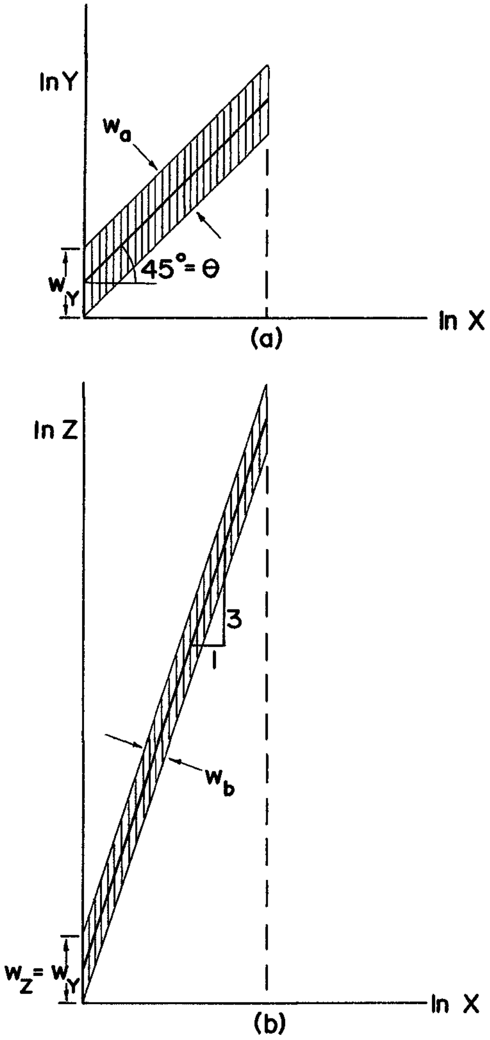


Fig 4 Exaggerated Correlation

phenomenon under investigation, The latter is found to have certain obvious advantages Once a decision is made, all the subsequent analysis is expressed in terms of the system chosen For instance if the dimensionless system is chosen, then the results should not be interpreted in terms of the individual dimensional parameters that form the dimensionless variables since the real variables of the experiment are the dimensionless variables

The usual cause of spurious correlation is the appearance of common quantities along both co-ordinate axes This is true for both a dimensional and a dimensionless system This must not be confused with the appearance of common dimensional parameters along both co-ordinate axes when dimensionless variables are plotted against each other

In addition, when considering dimensionless variables, care must be taken that the relationship between Y and X is not brought about only by variation of one or more of the repeating dimensional parameters common to both

ACKNOWLEDGEMENTS

The authors wish to thank Dr A Brebner who urged the writing of this paper

APPENDIX 1 - REFERENCES

1 Sedov, L I , Similarity and Dimensional Methods in Mechanics, Academic Press, New York, 1959

2 Langhaar, H L , Dimensional Analysis and Theory of Models, John Wiley and Sons, New York, London, 1962

3 Birkhoff, G , Hydrodynamics, a Study in Logic, Fact and Similitude, Princeton, Harvard University Press, 1950, Dover, 1955

4 Kamphuis, J William and Bowering, Richard "Impulse Waves", 12th Conference on Coastal Engineering, Washington, Sept 1970

5 Yalin, M S , "Geometric Properties of Sand Waves" Journal of the Hydraulics Division, ASCE, Vol 90, No HY5, Sept 1964, pp 105-119

6 Benson, Manuel A , "Spurious Correlation in Hydraulics and Hydrology", Journal of the Hydraulics Division, ASCE, Vol 91, No HY4, July 1965, pp 35-42

APPENDIX II - NOTATION

a_i	=	dimensional basic quantities (repeaters),
a_j	=	non repeating dimensional independent quantities,
b	=	dimensional dependent parameter,
C	=	constant,
D	=	diameter or grain size,
F	=	force,
f	=	dimensional function,
H	=	wave height,
h	=	depth of flow,
K	=	constant,
k	=	number of fundamental units,
m	=	constant (usually slope),
N	=	exponent,
s	=	distance,
U	=	approach velocity,
v_*	=	shear velocity,
w	=	scatter band width,
X_j	=	dimensionless independent variable,
Y	=	dimensionless dependent variable,
Z	=	dimensionless dependent variable,

α_i	=	exponent,
β_i	=	exponent,
Λ	=	dune length,
μ	=	dynamic viscosity,
Π	=	product,
ρ	=	density,
Σ	=	sum,
Φ	=	dimensionless function,
ϕ	=	dimensionless function of a single dimensionless variable