SPURIOUS OR VIRTUAL CORRELATION ERRORS COMMONLY ENCOUNTERED IN REDUCTION OF SCIENTIFIC DATA

Robert N. Meroney, Professor
Fluid Mechanics and Wind Engineering
Civil Engineering Department
Colorado State University
Fort Collins, CO

ABSTRACT:

Flawed analysis either intentional or though misunderstanding of commonly accepted data analysis methods can lead to erroneous results and presumption of correlation of cause and effect, when, in fact, there is little or none. Classical dimensional analysis combined with statistical regression of such scaled data may produce apparent correlation of information resulting in "virtual" or "spurious" correlation. Such inadvertent correlation errors can result in inappropriate conclusions and self-deception concerning the actual relationships between scaling variables and the associated reduction in variance found in tables and graphs. A number of dimensionless expressions used in meteorology and wind engineering induce large magnitudes of spurious correlation when plotted against other commonly accepted parameters. For example when drag coefficients, C_D , pressure coefficients, Cp, dimensionless shear, S*, or dimensionless concentrations, K, are regressed against Reynolds numbers, Re, dimensionless height, z/L_{MO} , or stratified Jensen numbers, z_o/L_{MO}, then inherent virtual correlations can exist with values from 50 to 95% even when random numbers are used to generate the component parts of the dimensionless groups!

KEY WORDS:

Data correlation, error analysis, similitude

CORRESPONDING AUTHOR:

Robert N. Meroney
Civil Engineering
Colorado State University
Fort Collins, CO 80523

Phone: (970) 491-8574
FAX: (970) 491-8671
E-mail: meroney@engr.colostate.edu

SPURIOUS OR VIRTUAL CORRELATION ERRORS COMMONLY ENCOUNTERED IN REDUCTION OF SCIENTIFIC DATA

Robert N. Meroney, Professor
Civil Engineering Department
Colorado State University

1.0 Introduction

The art of dimensional analysis was introduced in the year 1765 by the great scientist Euler, and since then, dimensional analysis has been accepted as a valuable interpretation technique in all scientific fields. It is a method for combining significant variables associated with a given physical phenomena into a set of dominant dimensional groups (usually fewer than the original variable set), which universally characterize the phenomena. These dimensional groups often contain common variables which can be the source of spurious correlation among the dimensionless terms.

In 1897 Karl Pearson [1] introduced the concept of virtual or spurious correlation. Pearson showed that if one has three sets of variables x_1, x_2, and x_3, which are entirely uncorrelated, and these variables are combined into two ratio sets x_1/x_3 and x_2/x_3, spurious correlation would exist between these ratios even though all variables are uncorrelated. Such inadvertent correlation errors can result in inappropriate conclusions and self-deception concerning the actual relationships between scaling variables and the associated reduction in variance found in tables and graphs.

In 1921 Reed [2] extended Pearson results to the general case of any two functions of two sets of variables, and defined spurious correlations as:

> *"Though no correlation exists between any two of a set of variables there will still exist correlation between any two functions of these variables whenever these two functions have any of the variables in common. The correlation existing under these conditions will be called spurious correlation"*

Chayes [3] using Pearson's original formula derived formulae for specific cases and cited in petrographic literature different cases where spurious correlations were found. Benson [4] derived additional examples for ratio and product relations. He also reviewed several cases in the hydraulics and hydrology fields where spurious correlations exist. He mentioned that:

> *"It is apparent that dimensional analysis is a potential field for the dangers of spurious correlation to sprout in. Many of the dimensionless ratios that are plotted against one another contain random elements in common. This practice is not wrong per se, nor are the correlation coefficients computed between such ratios (or sums) wrong, provided that the interpretation of correlation is made only in terms of ratios (or sums) and not in terms of the individual factors."*

The tendency for some regressions to imply erudition was identified by Kohn [5] with tongue in cheek as one of the principle methods of "obscurantism." He noted that "*one particularly elegant method for obtaining the desired correlation when correlation coefficient comes out at 0.00 with a distribution shaped like a cake with rains is to use Rudin's correlation sheet, which by proper stretching will produce the correlation needed*" (See Figure 1). Rudin's rubber sheet produces mechanically what one obtains analytically by inducing spurious correlation through proper combination of variables.

Hicks [6] investigated the existence of the artificial correlation in the atmospheric sciences, and he pointed out that a number of published power law relationships are artifacts of the analysis. Hicks also stressed the fact that virtual correlation arguments can be applied as objective tests of analysis procedures.

Several examples, where researchers have inadvertently introduced virtual correlation, are given by Benson [4], Hicks [6, 7, 8, 9], Meroney [10], and Elbahdry [11]. These examples are in the fields of fluid mechanics, hydraulics, hydrology and meteorology. An analytic framework from which to assess the presence of such spurious correlation errors is provided in this paper.

2. Theoretical Considerations:

Let x_1 , x_2, x_3,..., x_n, and y_1, y_2, y_3,....,y_k be two sets of n and k variables respectively. m_{x1}, m_{x2}, m_{x3},...., m_{xn}; S_{x1},S_{x2}, S_{x3},...., S_{xn}; and $r_{x1\,x2}$, $r_{x1\,x3}$, $r_{x1\,x4}$,....,$r_{xn-1\,xn}$ are means, standard deviations and coefficients of correlation of paired variables of the first set of variables. The same notations will be used for the means, standard deviations and coefficients of correlation of the second set of variables.

Now let $T = f(x_1, x_2, x_3, \ldots, x_n)$, and $Z = F(y_1, y_2, y_3, \ldots, y_k)$ represent any analytical functions. The correlation between these two functions is measured by

$$corr(T,Z) = \frac{cov(T,Z)}{\sqrt{var\ T\ var\ Z}} \tag{1}$$

or by taking the Taylor series approximation for T and Z in terms of m_i, s_i and r_i, one can develop the formula from Reed [2],

$$corr(T,Z) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} f_i\, F_j\, r_{x_i y_j}\, S_{x_i}\, S_{y_j}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} f_i\, f_j\, r_{x_i x_j}\, S_{x_i}\, S_{x_j} \sum_{i=1}^{k}\sum_{j=1}^{k} F_i\, F_j\, r_{y_i y_j}\, S_{y_i}\, S_{y_j}}} \tag{2}$$

where

$$f_i = \frac{\partial f(m_{x_i}, m_{x_2}, \ldots, m_{x_n})}{\partial m_{x_i}} \tag{3}$$

$$F_j = \frac{\partial F(m_{y_1}, m_{y_2}, \ldots, m_{y_k})}{\partial m_{y}}$$

If the two functions $T(x_1, x_2, ..., x_n)$ and $Z(y_1, y_2, , y_k)$ are linear functions, but variables $x_1 ... x_p$ are identical to the variables $y_1, ... y_p$, where $p \le k$ and $p \le n$, and variables $x_{p+1} ... x_k$ and $y_{p+1} ... y_n$ are uncorrelated, then $r_{xiyi} = 1$ if $i \le p$ and $r_{xiyi} = 0$ if $i > p$. Hence, using Equation (2) the spurious correlation $S_{corr(T, Z)}$ is evaluated as

$$S_{corr(T,Z)} = \frac{\sum_{i=1}^{p} f_i F_i S_{x_i} S_{y_i}}{\sqrt{\sum_{i=1}^{n} f_i^2 S_{x_i} \sum_{j=1}^{k} F_j^2 S_{y_j}^2}} \qquad (4)$$

Even if T and Z are nonlinear functions of x_i and y_j the same Equation (4) applies. Also if the variables x are unique functions of y rather than identically equal, spurious correlations will exist.

3.0 Examples from Fluid Mechanics, Meteorology and Wind Engineering

3.1 Limitations of Entrainment Rate Relations in Fluid Mechanics

Elbahdry [11] evaluated the presence of spurious correlation in expressions proposed for interfacial mixing driven by turbulent buoyant jets. Researchers have suggested a relationship exists between entrainment rate, $E = w_e / u$ and the bulk Richardson number, $Ri = \Delta b\, l / u^2$, where Δb is the buoyancy step across the interface ($g\, \Delta\rho / \rho$), l is a layer length scale, and u is a turbulent velocity scale. One can take u, w_e, and $\Delta b\, l$ as variables x_1, x_2 and x_3, respectively. Typical values of standard errors encountered are 0.03, 0.20, and 0.05, respectively. Using these values, we find the spurious coefficient as predicted by Equation (4) will be about 0.12. These spurious correlation values are small compared to the values of the overall correlations generally calculated for these cases which ranges between 0.75-0.85. The accuracy of the measurements appear to have a major effect on the values of spurious correlation. If the uncertainty values of the shared variables increase the spurious correlation values will increase.

3.2 Limitations of Popular Meteorological Boundary Layer Expressions

Hicks [6] pointed out that in plots of the dimensionless wind shear $S^* = (kz/u^*)(du/dz)$ against stability scale, $-z/L_{MO} = (kgzH_T)/(\rho c_p u^{*3} \Theta)$, one can take (du/dz), u^* and H_T as variables x_1, x_2 and x_3, respectively. In very unstable conditions one might encounter standard errors of 0.30, 0.50 and 0.30, respectively. Using these values, we find the spurious coefficient as predicted by Equation (4) will be about 0.84.

Another example discussed by Hicks [6] concerns an attempt to relate the drag coefficient of open water surfaces to wave effects. A regression of the drag coefficient $C_d = u^{*2}/u^2$ on the normalized wave velocity c/u^* may be characterized by u^{-2}, (x_1, $\sigma_{x1} = 0.2$), u^*, (x_2, $\sigma_{x2} = 0.5$), and c, (x_3, $\sigma_{x3} = 0.2$) where again u^* is the shared variable. The corresponding value of the induced correlation coefficient is now 0.91!

Hicks [7, 12] considered the proposal by scientists to correlate turbulence statistics in the lower atmosphere, σ_u / u^*, σ_v / u^*, σ_w / u^* and σ_T / T^* in and above the surface boundary layer with the depth of the mixing layer, z_i / L_{MO} or z/L_{MO}. [Panofsky et al., 13]. Regression of the original data sets and data produced by randomizing the friction velocities produce almost identical results. Secondary correlations between σ_i and u^* and between H and u^* are insignificant compared to the interdependence between u^* and L_{MO}, and the often noted 1/3 power law dependence between σ_i / u^* and z_i / L_{MO} is merely an artifact of the grouping of variables. The spurious correlation for the surrogate data sets produced from random numbers yielded values of 0.67.

Hicks [8] examined apparent dependencies of turbulent exchange coefficients for sensible heat and water vapor, K_H / K_W, on Bowen ratio, $\Delta T / \Delta e$, as presented by Verma et al. [14]. Hicks created another data set from random numbers to provide values simulating variables over ranges measured but completely uncorrelated. Then he calculated mixing ratios and regressed the results in the same manner as Verma et al. This produced correlation coefficients with higher values than Verma et al. obtained from the real data ($r_{artificial}$ = 0.76-0.91 versus r_{real} = 0.56-0.78).

3.3 Limitations of Popular Wind Engineering Expressions

Even a random number constrained to vary between 0.5 to 1.0 will give the impression of strong correlation to a totally independent parameter when both the ordinate and abscissa are scaled by functions of the independent parameter. Given R a random variable, then a highly correlated plot with some independent variable say A may be produced by plotting: $R\, f_1(A)$ versus $f_2(A)$. For example, it is dangerous to infer high correlation between fluid mechanic type parameters if a variable of interest is scaled by the friction velocity, u*, generally a small parameter when data is plotted against characteristic height, H, divided by some small scaling length such as roughness length, z_o. Spurious correlation or a biased plot occurs because $U_H / u^* = f(H/Z_o)$.

There indeed may be variance of the dependent parameter explained by a functional relationship between such a scaled grouping and H/Z_o. However, if 90% or more of the variance is explained by the velocity relationship above, then it is difficult to separate such correlation from random scatter of experimental data.

Consider examples selected from Ranga Raju et al.[15] and Bachlin et al. [16]. First, in Ranga Raju et al. the drag of a two-dimensional sharp-edged fence is plotted dimensionally as $C_{Do} = F/(\rho\, U_\infty^2\, H/2)$ versus δ/H. Let us re-scale the dependent parameter presuming a power-law relationship for the velocity profile as follows:

$$C_{Do} = \frac{F}{(\rho U_H^2 H/2)} \left(\frac{U_H}{U_\infty}\right)^2 , \textit{ but since}$$
$$\frac{U_H}{U_\infty} = \left(\frac{H}{\delta}\right)^\alpha , \textit{ then} \qquad (5)$$
$$C_{Do} = C_{DH} \left(\frac{H}{\delta}\right)^{2\alpha}$$

Since $\alpha = (10)^{1/2}\, (u^*/U_\infty)$, then for the range of conditions considered by Ranga Raju et al., $0.220 < \alpha < 0.238$. Figure 2 compares their data to the function above for a value of $0.8 < C_{DH} < 1.0$. Equation 2 yields $R^2 = 0.97$.

Second, consider the corrected data for Good and Joubert [17] provided by Ranga Raju et al. In this case $C_o^* = F/(\rho\, u^{*2}\, H/2)$ is plotted verus $H^+ = u^* h/\nu$. Let us again re-scale the dependent variable as follows:

$$C_o^* = \frac{F}{(\rho U_H^2 H/2)} \left(\frac{U_H}{u^*}\right)^2 , \textit{ but since}$$
$$\frac{U_H}{u^*} = \frac{1}{k}\ln(Hu^*/\nu) + B = \frac{1}{k}\ln(10\, H^+) , \textit{ then} \qquad (6)$$
$$C_o^* = C_{DH} \left(\frac{1}{k}\ln(10 H^+)\right)^2$$

Figure 3 compares the Good and Joubert data presented by Ranga Raju et al. against the relation above. Notice that the velocity expression explains almost all the variance shown by the data. Within experimental scatter one might argue that C_{DH} is nearly a constant. Equation (2) yields $R^2 = 0.98$.

Finally, for the same paper consider C_o^* plotted versus H/Z_o. Following the same procedure, a recasting of the definition of the dependent variable produces:

$$C_o^* = \frac{F_{Do}}{(\rho U_H^2 H/2)} \left(\frac{U_H}{u^*}\right)^2 , \text{ but now}$$
$$\frac{U_H}{u^*} = \frac{1}{k}\ln(H/Z_o) , \text{ then} \qquad (7)$$
$$C_o^* = C_{DH} \left(\frac{1}{k}\ln(H/Z_o)\right)^2$$

Figure 4 extracted from Ranga Raju et al. compares their data as well as the Good and Joubert data to the relation above. Again the correlation is strikingly good when $C_{DH} = 0.9$. Equation 2 yields $R^2 = 0.98$. If one is committed to the power law approach an alternative equation can be derived as follows:

$$C_o^* = C_{DH} \left(\frac{Z_o}{\delta}\right)^{2\alpha}\left(\frac{H}{Z_o}\right)^{2\alpha}\left(\frac{1}{C_f}\right) , \text{ where}$$
$$C_f = 0.1\ \alpha^2, \text{ also} \qquad (8)$$
$$\frac{Z_o}{\delta} = 0.15\ \exp\left(\frac{-1}{\alpha}\right), \text{ so that}$$
$$C_o^* = C_{DH} \frac{1}{0.1\ \alpha^2 e^2}(0.15)^{2\alpha}\left(\frac{H}{\alpha}\right)^{2\alpha}$$

Equation 2 yields $R^2 = 0.99$. If one limits the plot to the u^*/U_∞ range and H/Z_o provided by Ranga Raju et al., then the appearance of correlation against only H/Z_o exists.

Consider the plot of $(Cp^*)_{roof\ max}$ versus (H/z_o) prepared by Bachlin et al. [16]. (Henceforth, let this parameter be designated by C* for compactness.) Reformulate the dependent parameter as follows:

$$C^* = -\frac{P_{max}}{(\rho U_H^2/2)}\left(\frac{U_H^2}{u^*}\right)^2\frac{1}{C_f}, \text{ where}$$
$$C_f , \frac{Z_o}{\delta} \text{ are defined as before, then} \qquad (9)$$
$$C^* = Cp_H \frac{1}{0.1\alpha^2 e^2}(0.15)^{2\alpha}\left(\frac{H}{Z}\right)^{2\alpha}$$

Figure 5 displays this new correlation. If one limits the plot to the power law coefficient, α, and H/z_o ratio ranges provided by Bachlin et al., then the appearance of a correlation C* proportional to $(H/z_o)^{0.3}$ exists which agrees with the empirical expression proposed in their paper. In this case one need only limit the Cp_H variation to the region 0.7-

0.93 to obtain almost perfect correlation. Equation 2 yields $R^2 = 0.99$. The individual data sets actually seem to agree better with the slopes produced by the biased correlation than the proposed 0.29 slope.

The coefficient Cp_H does vary systematically with H/δ or H/z_o, but these plots shrink the ordinate and stretch the abscissa so much, that it is not possible to differentiate the variation from data scatter on such a chart. In conclusion, it would appear better to avoid dimensionless groups for pressure coefficient or drag force which are not of the O(1) in magnitude.

Acknowledgments: The author gratefully acknowledges references and discussions about virtual errors provided by Dr. Hesham Elbahdry, Cairo-Egypt, and Dr. Bruce Hicks, Air Resources Lab., NOAA.

References:

1. Pearson, K. (1896), "On a Form of Spurious Correlation Which May Arise When Indices are Used in the Measurements of Organs," Proceedings, Royal Soc. Of London, Vol. 60, pp. 489-502.
2. Reed, J. L. (1921), "On the Correlation Between Any Two Functions and Its Application to the General Case of Spurious Correlation," J. of the Washington Academy of Science, Vol. 11, pp. 449-455.
3. Chayes, F. (1949), "On Ratio Correlation in Petrography," J. of Geology, Vol. 57, No. 3, pp. 239-254.
4. Benson, M.A. (1965), "Spurious Correlation in Hydraulics and Hydrology," J. of the Hydraulics Division, Proceedings of the ASCE, Vol. 91, NO. HY4, pp. 35-42.
5. Kohn, A. (1970), "Principles and methods of obscurantism," New Scientist, Vol. 29, pp. 213-214.
6. Hicks, B. B. (1978a), "Some Limitations of Dimensional analysis and Power Laws," Boundary Layer Meteorology, Vol. 14, pp. 567-569.
7. Hicks, B.B. (1978b), "Comments on 'The Characteristics of Turbulent Velocity Components in the Surface Layer under Convective Conditions' by H.A. Panofsky, H. Tennekes, D.H. Lenschow, and J.C. Wyngaard," Boundary-Layer Meteorology, Vol. 15. Pp. 255-258.
8. Hicks, B.B. (1979), "Comments on 'Turbulent Exchange Coefficients for Sensible Heat and Water Vapor Under Advective Conditions' by S.B. Verma, N.J. Rosenberg, and B.L. Blad" Journal of Applied Meteorology, Vol. 18, pp. 381-382.
9. Hicks, B.B. (1985), "Behavior of Turbulence Statistics in the Convective Boundary Layer," Journal of Climate and Applied Meteorology, Vol. 24, No. 6, pp. 607-614.
10. Meroney, R.N. (1986), "Apparent Correlation of Data Produced by Using Correlated Variables in Ordinate and Abscissa Parameters," Appendix A, Fluid Mechanics Report No. CEM86-87RNM48, Colorado State University, Fort Collins, 9 pp.
11. Elbahdry, H.M. (1993), "Spurious Correlation as a Limitation of Dimensional Analysis," Chapter 5.0 of Interfacial Mixing in Diffusive Systems, Ph.D. Dissertation, Civil Engineering, Colorado State University, Fort Collins, 126 pp.
12. Hicks, B.B. (1981), "An Examination of Turbulence Statistics in the Surface Boundary Layer," Boundary-Layer Meteorology, Vol. 20, pp. 389-402.
13. Panofsky, H.A., Tennekes, H., Lenschow, D.H., and Wyngaard, J.C. (1977), "The Characteristics of Turbulent Velocity Components in the Surface Layer under Convective Conditions," Boundary-Layer Meteorol., Vol. 11, pp. 355-361.
14. Verma, S.B., Rosenberg, N.J. and Blad, B.L. (1978), "Turbulent exchange coefficients for sensible heat and water vapor under advective conditions," J. Appl. Meteor., Vol. 17, pp. 330-338
15. .Ranga Raju, K.G., Loeser, J. and Plate, E.J. (1976), "Velocity profiles and fence drag for a turbulent boundary layer along smooth and rough flat plates, " J. Fluid Mech., Vol. 76, part 2, pp. 383-399.
16. Bachlin, W., Plate, E.J., and Kamarga, A. (1982), "Influence of the ratio of building height to boundary layer

thickness and of the approach flow velocity profile on the roof pressure distribution of cubical buildings, J. Wind Engr. Ind. Aero., Vol. 11, pp. 63-74.

17. Good, M.C. and Joubert, P.N. (1968) "The form drag of two-dimensional bluff-plates immersed in turbulent boundary layers," J. Fluid Mech., Vol. 31, ppr. 547ff..

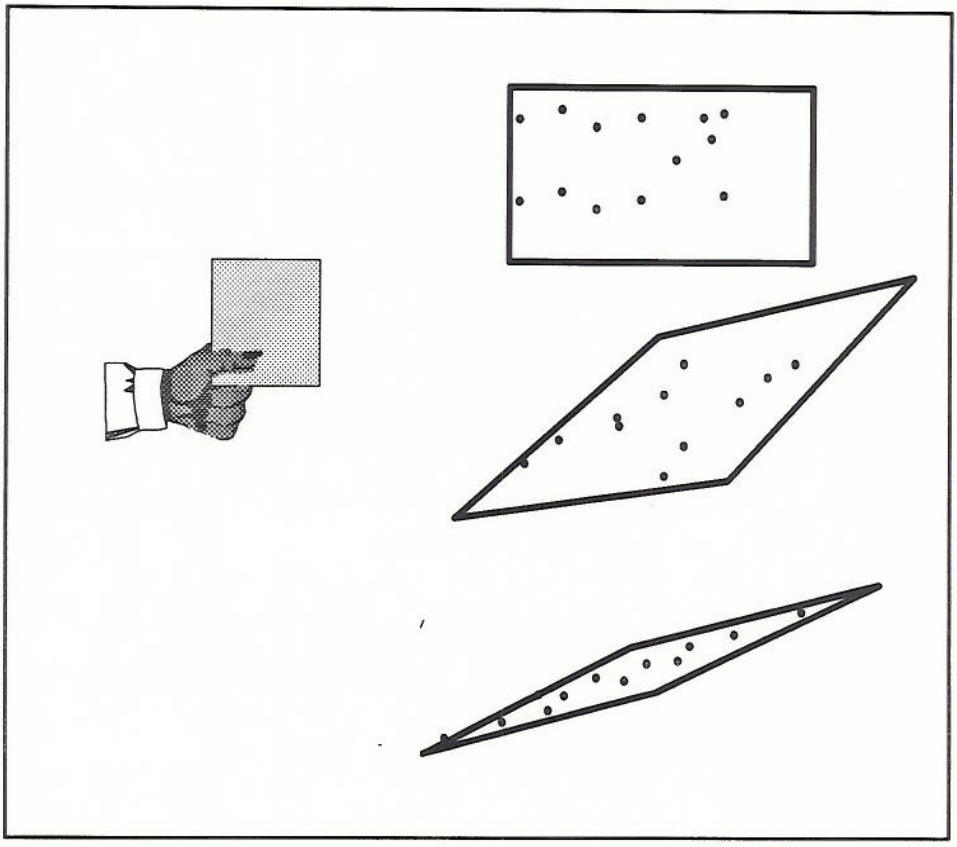

Figure 1: Rudin's rubber correlation sheet, Kohn (1970)

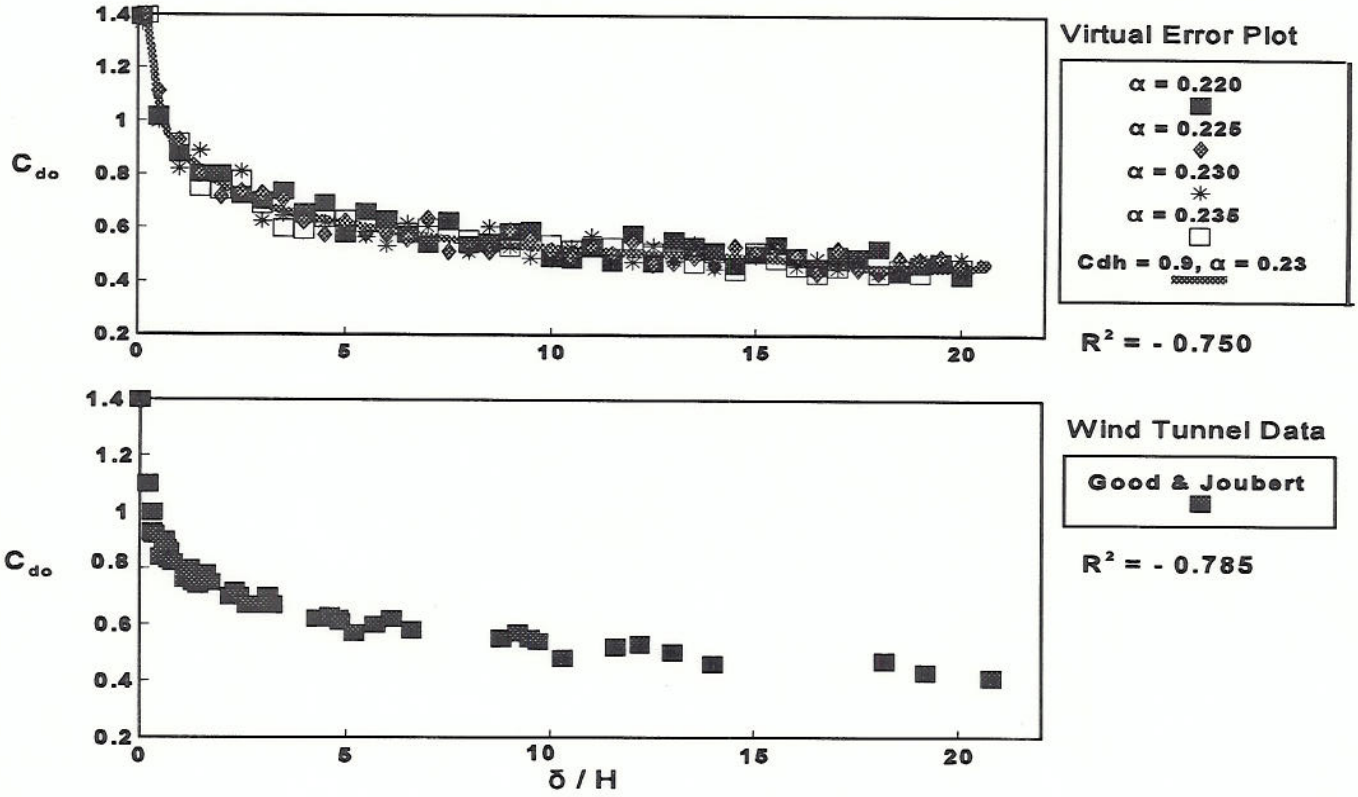

Figure 2: Power law correlation line for drag coefficient data for a 2-d fence on a smooth plate from Ranga Raju et al., 1976, Figure 1.

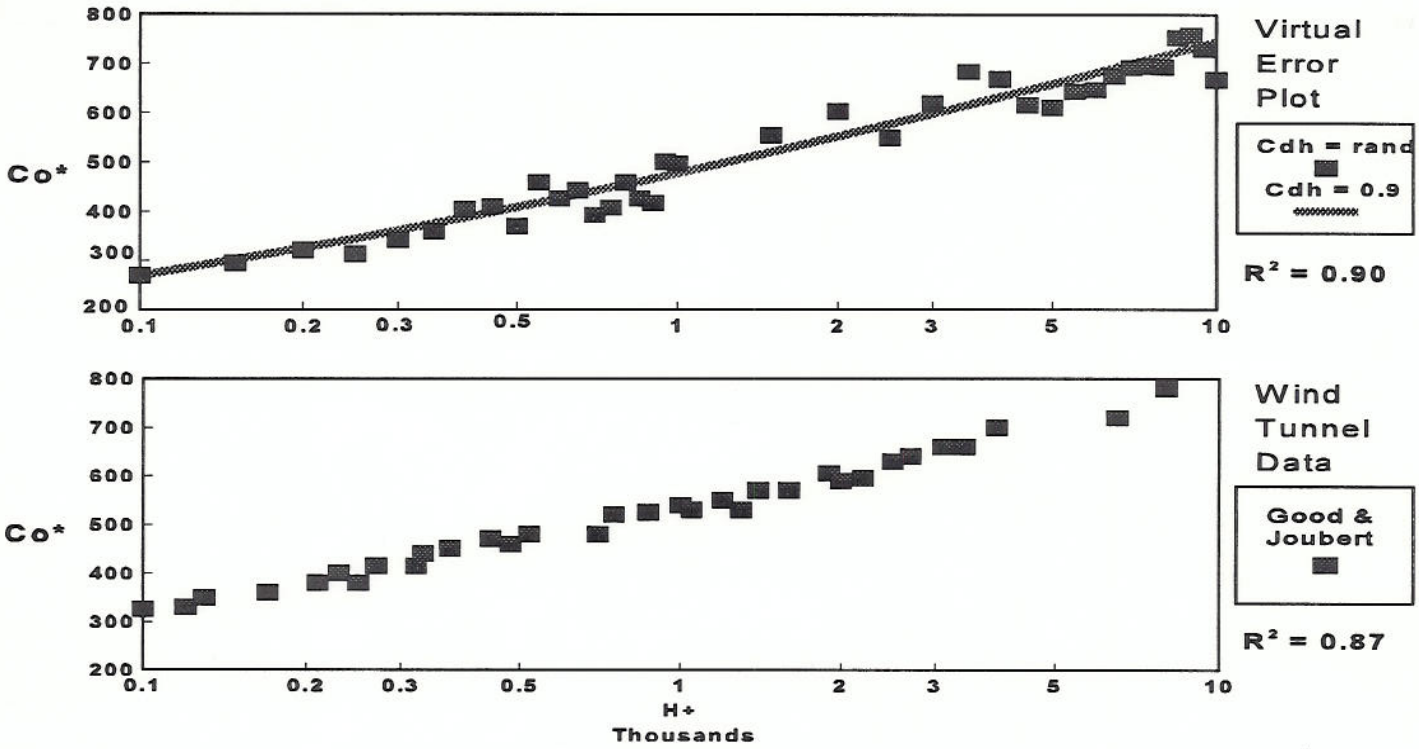

Figure 3: Log law correlation line for Good & Joubert fence data mounted on a smooth floor from Ranga Raju et al., 1976, Figure 3.

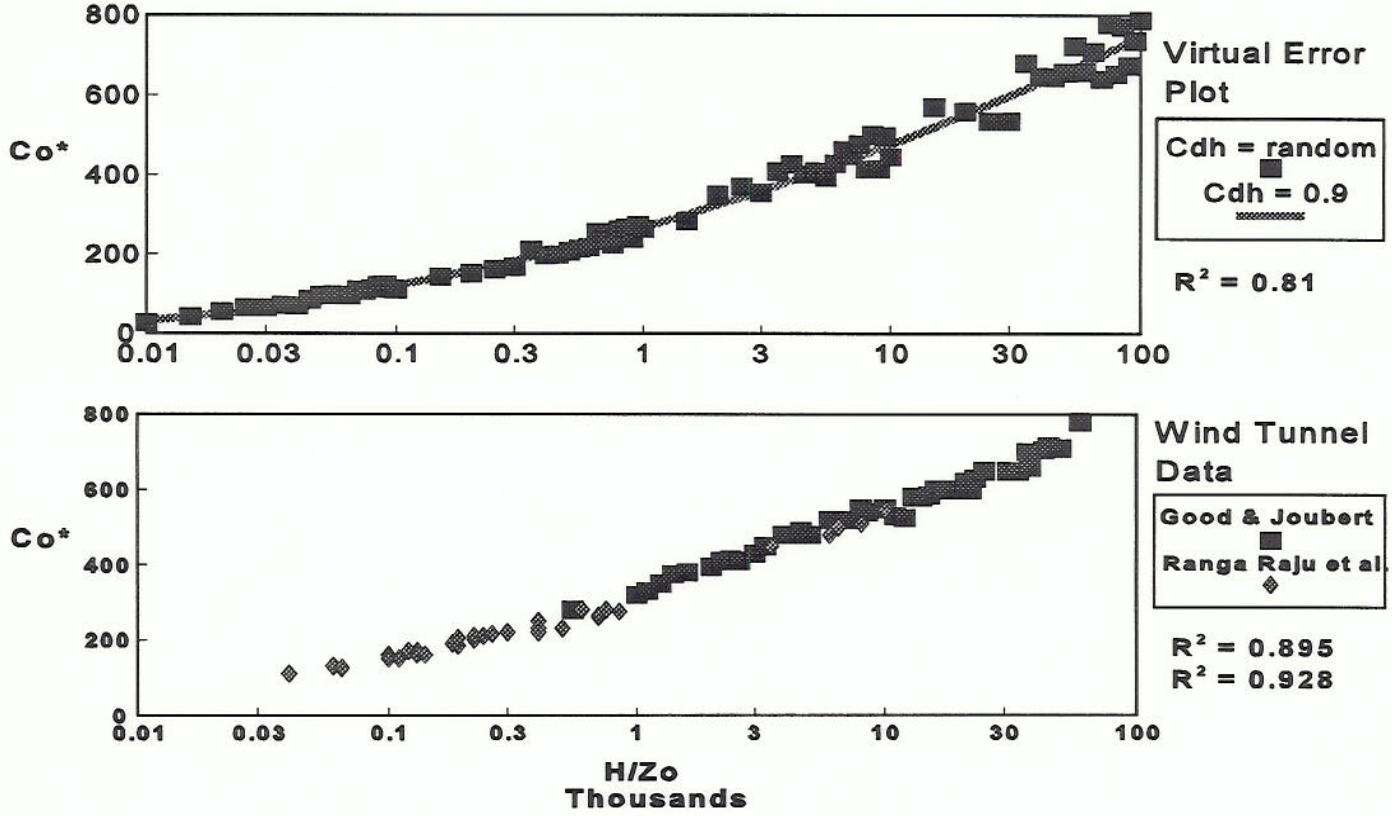

Figure 4: Log law correlation line for drag coefficient for a 2-d fence in a turbulent boundary layer from Ranga Raju et al., 1976, Figure 14.

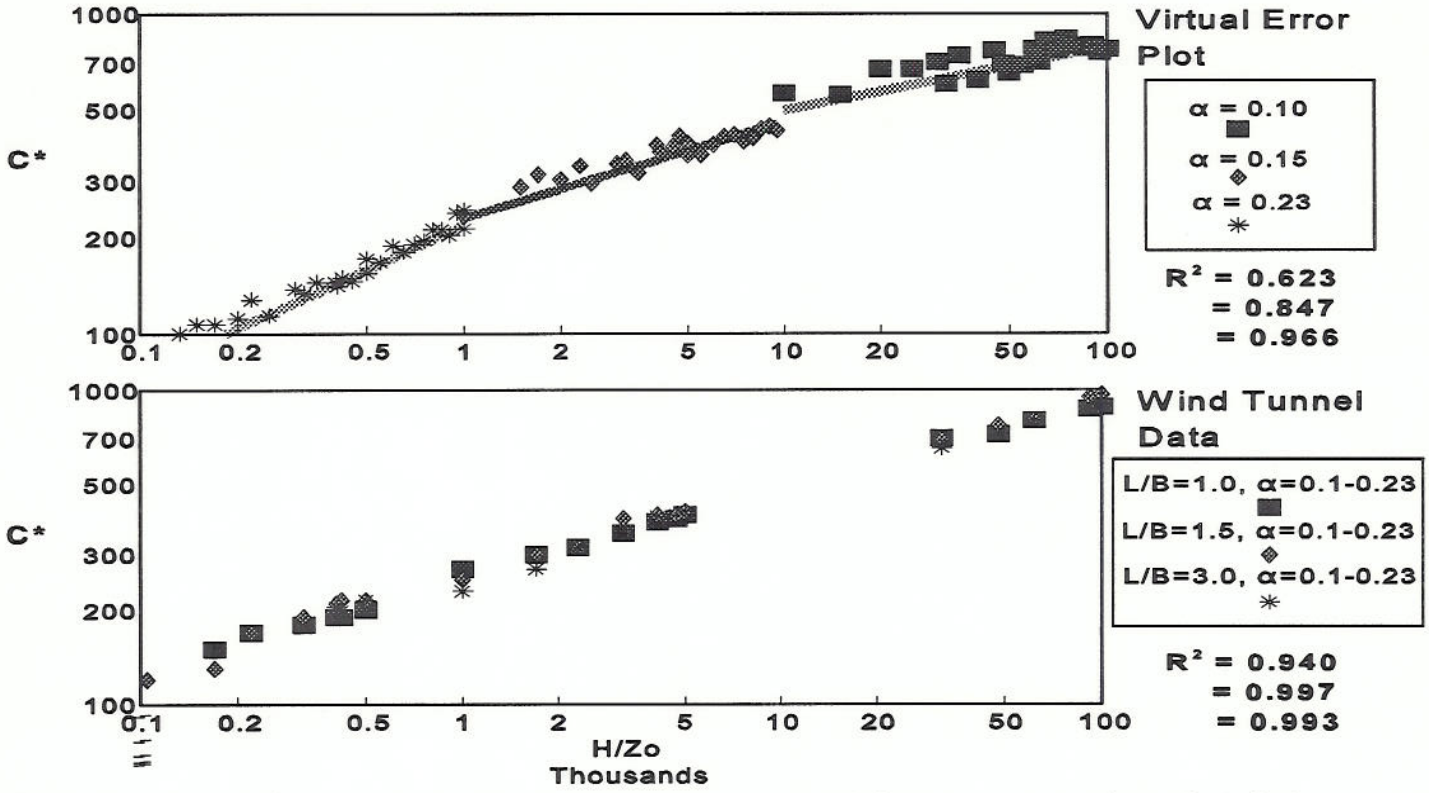

Figure 5: Power law correlation line for peak pressure coefficient data on a prismatic building model from Bachlin et al., 1982.