Computers & Chemistry, Vol. 5, pp. 33–39, 1981
Printed in Great Britain.

0097-8485/81/010033-07$02.00/0
Pergamon Press Ltd.

STATISTICAL CORRELATION OF MOLECULAR STRUCTURE WITH BOILING POINTS OF N-HETEROCYCLIC COMPOUNDS: MULTIPLE LINEAR REGRESSION ANALYSIS

J. F. OGILVIE
Research School of Chemistry, Australian National University, Canberra, A.C.T. 2600, Australia

and

M. A. ABU-ELGHEIT†
Department of Chemistry, Kuwait University, P.O.Box 5969 Safat, Kuwait

(*Received* 27 *February* 1980; *in revised form* 2 *June* 1980)

Abstract—After a summary review of statistical theory in matrix form, the computer program MULNRG for multiple linear regression analysis is outlined. Then multiple correlation analysis is applied to the boiling points of two collections of chemical substances, alkanes (for comparison with previous results) and N-heterocyclic compounds. The results of fitting show a general accuracy of about 2°C, but outliers are recognised in the latter category. The general problems of multiple linear regression analysis are discussed with references to the specific results.

1. INTRODUCTION

The correlation of bulk properties of chemical compounds with their molecular structural features is an enduring problem. In general, qualitative trends are readily established for many properties, but their quantitative reproduction by means of analytic functions remains an elusive goal. One category of compounds which success has relatively favoured is the hydrocarbon collection, especially alkanes. For the latter substances, physical properties have been claimed to be well represented according to linear functions of numbers of carbon atoms and bond types (Jessup, 1937; Rossini, 1940; Tilicheyev & Iogansen, 1951; Laidler, 1956). Further development of these procedures has led to their application to higher alkanes, other classes of hydrocarbons, thiols and thiophenes, for which extensive data are now available (American Petroleum Institute, 1974).

N-heterocyclic compounds have remained neglected in this regard. Our interest in the properties of these compounds is due to their occurrence in petroleum and heavy petroleum distillates in molar fractions that although small are still sufficient to act as permanent poisons to acidic catalysts in catalytic cracking processes. It is desirable to be able to represent characteristic physical properties of these organic nitrogen compounds by analytic functions for three reasons. First, providing that such a fucntion can be found, there is thus a check on the experimental accuracy of the measured properties. Second, there would be some predictive capacity for compounds for which the particular properties had not been measured. Third, significant deviations between measured and predicted properties may indicate, if not simply an inapplicability of the analytic function for reasons of poor choice of parameters, the presence of either experimental error or intermolecular effects that cannot be encompassed explicitly in an analytic function constructed purely on intramolecular parameters. Although the specific property of interest in this article is the normal boiling point (the temperature at which the equilibrium vapour pressure is 101,325 N m^{-2}) of pure compounds, we suppose that analogous procedures could be applied to such other properties as heat of vaporisation, molar volume and vapour pressure (Tatevskii *et al.*, 1961) or molar thermal capacities (Akhmedov, 1979), for which experimental data are both less abundant and less easily obtained.

In the first attempts at extensive use of quantitative correlations, calculations were performed manually for lack of a better method. With the recent general availability of electronic digital computers, not only can the same calculations be conducted much more rapidly, with a full statistical analysis included, but also it is easy to test many more molecular parameters as known variables, combinations of parameters, and collections of compounds, in order that the best or most useful conditions to apply to any desired objective may be found. Furthermore, testing of methods with synthetic data or archetypal compounds is particularly convenient, again with the reliability of the process being assessed at each stage by means of standard statistical procedures.

In this work we have developed a new computer program, MULNRG, for multiple linear regression analysis, written in interactive BASIC language for implementation on small computers. The interactive nature of BASIC is especially convenient for testing as well as production runs, because the actual computations are neither extremely time-consuming nor requiring of huge core memory. Moreover, the availability of BASIC interpreters containing several useful matrix functions and operating with more than ten significant digits ensures economy of programming effort and relative efficiency of execution.

In this work, we have applied MULNRG to the boiling points of N-heterocyclic compounds found in petroleum or having related structures. First we have tested

†Permanent address: Department of Chemistry, Faculty of Science, Alexandria University, Alexandria, Egypt.

MULNRG on synthetic data and on alkanes in order to gauge the accuracy of the procedure and also to obtain a comparison with previous results (Tatevskii *et al.* 1961). Then the program has been applied to a collection of the nitrogen compounds and extensive tests of possible molecular parameters as known variables, in both linear and power forms, have been made. The results demonstrate the effectiveness of the method and its applicability to other problems in which a complicated dependence of some property on several parameters may be suspected. In what follows, we first outline the mathematical and statistical theory, then describe briefly its implementation in MULNRG; finally we present the results for our exemplary problem and discuss the general applicability of the method.

2. MATHEMATICAL BASIS OF THE ALGORITHM

We suppose that boiling point T, the dependent known variable, of a member of a collection of chemical compounds may be expressed as a function of structural parameters of constituent molecules as formally independent variates or regressor variables, such as the numbers of some kinds of bonds N_k. Thus we can write

$$T = c_0 + c_1N_1 + c_2N_2 + \ldots + c_kN_k + \ldots + c_mN_m$$

for each compound as a function of m regressors. Such a relation is supposed to apply to each compound i, with boiling point T_i and bond parameters N_{ki} as known variables or variates, for all the n compounds in the collection. The regression coefficients c_k that become the unknown parameters remain to be determined statistically, according to the principle of least squares (Kendall & Stuart, 1979).

We write the normal equations (Yamane, 1973) in the matrix form

$$((E))\ \ ((C)) = ((D)),$$

where

$$((E)) \equiv ((e_{kl})) \equiv \begin{bmatrix} n\Sigma N_{1i}\Sigma N_{2i} & \ldots\ \Sigma N_{ki} & \ldots\ \Sigma N_{mi} \\ \Sigma N_{1i}^2\Sigma N_{1i}N_{2i} & \ldots\ \Sigma N_{1i}N_{ki} & \ldots\ \Sigma N_{1i}N_{mi} \\ \Sigma N_{2i}^2 & \ldots\ \Sigma N_{2i}N_{ki} & \ldots\ \Sigma N_{2i}N_{mi} \\ & \vdots & \vdots \\ \text{(symmetric)} & \Sigma N_{ki}^2 & \ldots\ \Sigma N_{ki}N_{mi} \\ & & \vdots \\ & & \Sigma N_{mi}^2 \end{bmatrix}$$

Thus the E matrix is a square symmetric two-dimensional array of order $m+1$, where m is the number of regression coefficients (in addition to the intercept c_0) to be determined, whereas the C and D matrices are column arrays or vectors of order $m+1$. In each case the symbol Σ indicates summation over all the n compounds in the selected collection.

We define the sum of squared errors (or deviations between measured and calculated quantities) to be

$$\mathrm{SSE} \equiv \sum_{i=1}^{n} [T_i - (c_0 + c_1N_{1i} + c_2N_{2i} + \ldots + c_kN_{ki} + \ldots + c_mN_{mi})]^2$$

which can also be expressed as

$$\mathrm{SSE} = \Sigma T_i^2 - ((C))'\,((D))$$

in which $((C))'$ is a row matrix, the transpose of the column matrix $((C))$. The regression coefficients c_k, $(0 \leq k \leq m)$, are determined in the least-squares sense, so as to *minimise* SSE. The *multiple correlation coefficient* R is a measure of the strength of the linear relationships between the bond parameters N_{ki} and boiling point T_i:

$$R^2 = 1 - \frac{\mathrm{SSE}}{\Sigma T_i^2 - d_0^2/n}$$

and $0 \leq R^2 \leq 1$, $d_0 = \Sigma T_i$. A value of $|R|$ near unity indicates that a strong correlation is present. In order to test whether the multiple correlation coefficient is significantly *different* from zero, we can use the F-statistic, defined as

$$F = \frac{(n-m-1)}{m} R^2 \frac{(\Sigma T_i^2 - d_0^2/n)}{\mathrm{SSE}};$$

in this case F signifies (Kendall & Stuart, 1979) the F-distribution with m and $(n-m-1)$ degrees of freedom: a large value of F indicates that correlation is present, but more refined use of the F-statistic is also possible (Yamane, 1973). We require also a measure of the significance of each regression coefficient c_k, specifically its variance (i.e. the square of its standard deviation). The variance of coefficient c_k can thus be designated $\sigma_k^2 = g_{kk}\sigma^2$, in which g_{kk} are the diagonal elements of the G matrix, itself the inverse of the E matrix:

$$((G)) = ((E))^{-1};$$

σ^2, an unbiased estimator of the error variance of regression, is related to SSE by

$$\sigma^2 = \frac{\mathrm{SSE}}{n-m-1}.$$

$$((C)) \equiv ((c_k)) \equiv \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_k \\ \vdots \\ c_m \end{bmatrix}, \text{ and } ((D)) \equiv ((d_k)) \equiv \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_k \\ \vdots \\ d_m \end{bmatrix} \equiv \begin{bmatrix} \Sigma T_i \\ \Sigma N_{1i}T_i \\ \Sigma N_{2i}T_i \\ \vdots \\ \Sigma N_{ki}T_i \\ \vdots \\ \Sigma N_{mi}T_i \end{bmatrix}.$$

Finally we can use as a simple indicator of goodness of fit the average magnitude of deviation ΔT_{av} between calculated (on the basis of the regression coefficients) and observed boiling points; thus

$$\Delta T \equiv T_{calc} - T_{obs}, \text{ and } \Delta T_{av} \equiv \frac{1}{n}\sum_{i=1}^{n} |\Delta T_i|.$$

In this work we assume that the best criterion for selection of parameters, disregard of outliers etc., is the F-statistic or F-value, rather than the multiple correlation coefficient R or the average magnitude of deviation ΔT_{av}, although in practice we find (as expected) that a large value of F corresponds to relatively small values of σ_k and ΔT_{av}, and values of $|R|$ approaching unity. However, the use of one of these criteria of goodness of fit does not in general correspond to an extremal property of another; the reasons for this are presumably the only moderate number of data and their lack of exact conformity to any particular theoretical distribution function. Nevertheless the differences between the extremal values and the values consistent with, for instance, a maximum in F are relatively small.

Another useful quantity in error analysis is the dispersion matrix ((V)) (Kendall & Stuart, 1979), a square array formed by scalar multiplication of the G matrix by the error variance of regression:

$$((V)) = \sigma^2((G)) \text{ where } ((V)) = ((v_{rs})).$$

Apart from the diagonal elements of this matrix that are the squares of the standard deviations of the regression coefficients, $v_{kk} = \sigma_k^2$, the off-diagonal elements yield information about the correlation of the nominally independent known variables or regressors. For this reason the dispersion matrix may also be called the variance-covariance matrix. The use of the covariance (off-diagonal) elements occurs in assessing the variance of an unknown boiling point from a given set of known variables (from structural considerations of the particular molecule) and the already determined regression coefficients:

$$\text{var}(T) = \sum_{j=0}^{m} N_j^2 \,\text{var}(c_j) + 2\sum_{0<r<s\leq m}\sum N_r N_s \,\text{cov}(c_r, c_s).$$

In terms of the V matrix, we may rewrite this as

$$\text{var}(T) = \sum_{j=0}^{m} N_j^2 v_{jj} + 2\sum_{0<r<s\leq m}\sum N_r N_s v_{rs},$$

and the standard deviation of the predicted boiling point is just the square root of the variance var (T).

3. IMPLEMENTATION OF ALGORITHM

The newly prepared program MULNRG has been designed to take advantage of present small computers employing BASIC interpreters including several matrix operations. Because these circumstances are not now uncommon and will become even less so in the near future, our procedure is of general applicability. In BASIC, data are conveniently stored in DATA statements within a program, where they are readily modified interactively, or in data files maintained on tape cartridge or disc units. We have employed the former practice, because of adequate core memory. Thus the first block of the program consisted of sufficient DATA statements to supply all test and production runs. The second block had statements to read and select the desired data, including a character string as compound identifier (chemical name), the boiling point, and a group of molecular parameters specifying its structure as a basis of multiple regression. The next block of program was intensive numerical computation, according to the mathematical formalism outlined in the previous section. Despite the powerful applicability of our straightforward approach, there were as few as thirty-two brief statements in (one operating version of) this computation section of MULNRG. The final block was directed to printing of results. Typical execution times on a Hewlett–Packard 9845S computer were 200 s, including inversion of a tenth-order matrix and printing 80 lines of alphanumeric text (input data, statistical results, and table of residuals); on the PDP 11/45 computer used in time-shared mode, the total time was greater because of a slower printer although matrix inversion was quicker.

The program was tested by generation of synthetic data by means of a DEF FNT function generator statement, that for a given set of ten N_{ki} data (simulated bond parameters arbitrarily chosen within the range 0–20), for thirty different cases, formed a T_i value according to a defined relation including coefficients set to have the range of magnitudes 0–50 annd varied signs. Also within this generated function was included a (pseudo-) random number generator that provided a "noise" addendum in the range -0.5–0.5x; the noise amplitude x was varied in the range 0.5–20. Under these conditions, the program MULNRG determined the eleven regression coefficients c_k to be within one (in 9 cases) or 1.5 (in the other two cases) estimated standard deviations of the preset coefficients in the function generator. Furthermore the estimated standard deviations σ_k and average differences ΔT_{av} were in every case less than the maximum noise amplitude, actually less than ~80% of the maximum amplitude for the standard deviations (as expected for a normal distribution function), and 40% correspondingly for the average magnitude of deviation. The individual ΔT_i were always in the range expected from the noise amplitude. We therefore concluded that the program MULNRG operated efficiently within the conditions of testing, which were designed to simulate the usage for real data for alkanes and N-heterocyclic compounds.

4. SELECTION OF MOLECULAR PARAMETERS, AND RESULTS

The molecular parameters to be applied to boiling-point computation as known variables were constrained to be those derivable by inspection of the structural formula, i.e. number of atoms of a given type or numbers of bonds between functional positions, and not quantitative measures of molecular geometry such as internuclear separations. Many parameters were tested, including some generated from others, such as squares or cubes of one simple parameter or linear combinations of simple parameters. Despite the evident relationships between these composite regressors, the analysis procedure is kept rigorously linear in the regression coefficients. The progress of testing was guided by the sensitivity of the results, according to the F-value, to inclusion of various compounds incorporating a range of structural features. Many of the more useful parameters are listed in Table 1 for alkanes and Table 2 for N-heterocyclic compounds.

The alkanes were included for test purposes because of existing results (Tatevskii *et al.*, 1961) that indicated

Table 1. Structural parameters for alkane molecules

No.	Symbol	Definition of Variate
1	n_1	No. of primary hydrogen atoms
2	n_2	" " secondary " "
3	n_3	" " tertiary " "
4	$n_{1,2}$	No. of C-C bonds between primary and secondary carbon atoms
5	$n_{1,3}$	" " " " primary and tertiary carbon atoms
6	$n_{1,4}$	" " " " primary and quaternary carbon atoms
7	$n_{2,2}$	" " " " secondary carbon atoms
8	$n_{2,3}$	" " " " secondary and tertiary carbon atoms
9	$n_{2,4}$	" " " " secondary and quaternary carbon atoms
10	$n_{3,3}$	" " " " tertiary carbon atoms
11	$n_{3,4}$	" " " " tertiary and quaternary carbon atoms
12	$n_{4,4}$	" " " " quaternary carbon atoms
13	n_2^2	square of parameter 2
14	n_2^3	cube of parameter 2
15	$n_{1,4}^2$	square of parameter 6
16	$n_{1,4}^3$	cube of parameter 6
17	$\sqrt{n_2}$	square root of parameter 2

Table 2. Structural parameters for N-heterocyclic compounds

No.	Symbol	Definition of Variate
1	$n_{2,2}$	No. of C-C bonds between secondary carbon atoms
2	$n_{2,3}$	" " " " " secondary and tertiary carbon atoms
3	$n_{3,3}$	" " " " " tertiary carbon atoms
4	$n_{2,N}$	No. of C-N " " nitrogen and secondary carbon atoms
5	$n_{3,N}$	" " " " " nitrogen and tertiary carbon atoms
6	n_{Csc}	No. of ring carbon atom bearing the side chain
7	$n_{N,Hr}$	Summation of bonds between nitrogen and all hydrogen atoms of the ring
8	$n_{N,Hsc}$	Summation of bonds between nitrogen and all hydrogen atoms of the side chain
9	n_C	No. of carbon atoms
10	n_H	No. of hydrogen atoms
11	n_{Me}	No. of methyl groups
12	$n_{1,3}$	No. of C-C bonds between primary and tertiary carbon atoms
13	$n_{1,N}$	No. of C-N bonds between nitrogen and primary carbon atoms
14	$n_{N,Me}$	No. of bonds between nitrogen atom and methyl groups
15	$n_{N,Hr}^2$	Square of parameter 7
16	$n_{2,2}^2$	Square of parameter 1
17	MS	Molar mass
18	n_C^2	Square of parameter 9
19	$n_{2,3}^2$	Square of parameter 2
20	$n_{6,7}$	Summation of parameters 6 and 7
21	$n_{6,7,16}$	Summation of parameters 6, 7 and 16

the feasibility of relating physical properties, such as boiling points, to structural features of these compounds. The test set of fifty hydrocarbons included unbranched alkanes from hexane to eicosane and branched derivatives of butane to nonane, boiling in the range 60–343°C. Eight known variables sufficed to produce the best fit for these alkanes, yielding an F-value of 3185. Statistical results from a sample of seven runs are summarised in Table 3. Within the arbitrarily selected range of representative compounds, as described above, no alkane was excluded (as an outlier) because its calculated value of boiling point deviated unacceptably from the measured value. The relation producing the maximum F-value was:

$$T/°C = (-131.7 \pm 4.5) + (15.5 \pm 0.66)\,N_1 + (9.447 \pm 0.41)N_2 + (4.09 \pm 2.1)N_3 + (6.968 \mp 0.95)N_4 + (9.648 \pm 1.7)\,N_6 + (7.679 \pm 0.78)\,N_7 + (-0.00224 \pm 0.00011)\,N_{14} + (0.0833 \pm 0.02)\,N_{16};$$

with each regression coefficient in this relation appears the corresponding standard deviation; the subscripts on the formal parameters N_i indicate the actual known variable according to Table 1. The average magnitude of deviation in the best fits was about 1.9°C, and the maximum deviation ΔT was only 4.8°C. Because only fifty compounds from a usable group exceeding one hundred were utilised in the analyses, the remaining compounds within the same range may be employed to test the predictive capacity of the above relation; the average magnitude of differences between predicted and observed boiling points was less than 2°C in the representative sample tested. In previous work (Tatevskii *et al.*, 1961), the same average deviation of about 2°C was obtained on the entire set of the same alkanes.

For the forty N-heterocyclic compounds for which data were available, selection of molecular parameters as regressor variables was more difficult than for alkanes. Eventually, more than twenty known variables— some simple, some composite— were tested. The best fit required ten regressors, according to Table 4 which presents results from a sample of runs. The relation that produced these results was:

$$T/°C = (49.3 \pm 22) + (59.98 \pm 5.3)\,N_1 + (68.7 \pm 4.6)\,N_2 + (74.48 \pm 3.1)\,N_3 + (32.74 \pm 4.7)\,N_4 + (43.28 \pm 3.5)\,N_5 + (-58.75 \pm 6.0)\,N_9 + (20.09 \pm 3.7)\,N_{10} + (3.207 \pm 0.77)\,N_{14} + (-0.612 \pm 0.26)\,N_{19} + (-1.40 \pm 0.33)\,N_{21};$$

the subscripts of N_i refer to the known variables defined in Table 2. The various indicators of success of fit in this case were:

standard deviation of fit = 3.5°C;
root mean square of residuals = 2.8°C;
magnitude of multiple correlation coefficient = 0.999197;
average magnitude of deviation = 2.3°C;
F-value = 1306.

Table 3. Sample statistical results of regression analyses for alkanes

Run No.	Parameter	ΔT_{av}/°C	F	R
1	1-4,7,8	5.68	388	0.9909
2	1-4,7,8,17	5.16	453	0.9934
4	1-4,7,8,13	2.89	1638	0.9982
7	1-4,6,7,13,14	2.29	2273	0.9989
9	1-4,6,7,14	2.34	2608	0.9989
11	1-4,6,7,13,14,16	1.89	2854	0.9992
13	1-4,6,7,14,16	1.94	3185	0.9992

Table 4. Sample statistical results of regression analyses for N-heterocyclic compounds

Run No.	Parameter	ΔT_{av}/°C	F	R
1	1-3,5-7,9,16	4.27	529	0.9973
8	1-7,9,10,14,15	3.26	608	0.9985
10	1-7,9,10,15,18	3.21	651	0.9986
12	1-7,9,10,15,16,18,19	2.53	777	0.9991
14	1-7,9,10,16,18	2.94	822	0.9989
16	1-7,9,10,14,15,16,18,19	2.16	852	0.9993
19	1-7,9,10,14,16	2.86	876	0.9990
21	1-7,9,10,14,16,19	2.32	947	0.9992
23	1-5,9,10,14,16,19,20	2.33	1086	0.9992
24	1-5,9,10,14,19,21	2.30	1306	0.9992

Of the total group of compounds, eight had to be rejected as outliers, according to the unbiased criterion of diminution of F-value if any was included. For the remaining thirty-two compounds, deviations were in the ranges 0.1–1.0°C for ten compounds, 1.4–1.9°C for six, 2.2–2.6°C for five, 3.2–3.9 for eight, and 4.7, 5.8, 6.1°C for the remaining three; this distribution is approximately normal.

Of the outliers, four have large deviations from the above regression relation: imidazole ($\Delta T = -124.9$°C), pyrazole (-67.0°C), indazole (-23.6°C) and pyridazine (-94.6°C). In the first three cases, strong hydrogen bonding between donor and acceptor sites on different molecules provides a qualitative explanation of the deviation, but such an explanation is invalid for pyridazine. A partial explanation for the latter deviation may however be found in its large dipole moment. The dipole moments (in units of 10^{-30}C m) of the cyclic azines (McClellan, 1974) are pyridine, 7.305; pyrazine, 0; pyrimidine, 7.785; pyridazine, 14.08. For this last compound, the moment expected according to a simple bond moment model (on the basis of pyridine) would be 14.1, practically the same as the observed; the coincidence may be taken to indicate that no special intramolecular interactions result from the proximity of two nitrogen atoms in the molecular ring. The other two diazines, pyrimidine ($\Delta T = 1.8$°C) and pyrazine (11.7°C), cannot be considered particularly anomalous with respect to their conformity to this boiling point correlation. The other three compounds are 7-methylisoquinoline ($\Delta T =$ 15.5°C, quinoxaline (20.1°C) and 2-methylquinoxaline (18.4°C); in these cases the causes of the deviations also remain obscure, so these compounds can simply be considered outliers with respect to the boiling-point correlation, resulting from an unsolved deficiency of the model. Presumably in the cases of pyrrole and its derivatives, the strength of the hydrogen bonding, which in the pure substance can occur only between nitrogen atoms each of which carried a hydrogen atom, is insufficient to disrupt the correlation with other structural aspects.

Table 5. Table of residuals for N-heterocyclic compounds with best fitting parameters

Compound	T_{calc}/°C	ΔT/°C
Pyrrole	129.1	-0.6
Pyrrole, 1-methyl	115.2	2.4
Pyrrole, 2-methyl	143.6	3.9
Pyrrole, 3-methyl	147.0	3.6
Pyridine	115.4	0.1
Pyridine, 2-methyl	132.6	3.2
Pyridine, 3-methyl	140.2	-3.9
Pyridine, 4-methyl	143.4	-1.9
Indole	253.7	-0.3
Indole, 2-methyl	267.3	-4.7
Indole, 3-methyl	270.0	3.7
Indole,1,3-dimethyl	256.1	-2.4
Quinoline	237.1	-0.5
Quinoline, 2-methyl	253.5	5.8
Quinoline, 3-methyl	260.2	0.6
Quinoline, 4-methyl	259.5	-6.1
Quinoline, 5-methyl	264.5	2.3
Quinoline, 6-methyl	265.2	0.2
Quinoline, 7-methyl	259.2	1.5
Quinoline, 8-methyl	249.3	1.4
Isoquinoline	241.8	-1.4
Isoquinoline, 1-methyl	251.7	3.7
Isoquinoline, 3-methyl	252.8	-0.3
Isoquinoline, 4-methyl	256.8	0.8
Isoquinoline, 6-methyl	266.5	1.0
Isoquinoline, 8-methyl	255.8	-2.2
Carbazole	357.4	2.6
Acridine	345.1	-0.4
Phenanthridine	344.3	-3.7
Pyrimidine	125.3	1.8
Pyrimidine, 2-methyl	134.3	-3.7
Pyrazine, 2-methyl	138.4	1.9

5. CONCLUSIONS

We have demonstrated the feasibility of correlating bulk properties, such as the normal boiling point, of classes of compounds to the topological features of their molecular structures, even though the structures of the N-heterocyclic compounds are quite complicated relative to alkanes (unbranched, and methyl derivatives). In so doing, we have proceeded quite empirically: there is no attempt to construct a theory to encompass these results. Indeed, the only "theory" that might be generally ascribed to our procedure is that of additivity—the properties of a molecule might be constituted as the sum of contributions of its parts (atoms, groups, bonds etc.). Such "theory" has long been known to be generally wrong in application to macroscopic properties that depend not only on intramolecular (or intrinsic) effects but also strongly on intermolecular (or environmental) interactions. In fact, we expect that correlations of bulk properties with structures will *not* generally apply, but within restricted collections of compounds approximate correlations might be found. Thus, although we have demonstrated correlations of boiling points with structural features of alkanes and N-heterocyclic compounds separately, we would not expect a satisfactory correlation to apply to both collections simultaneously.

The evident advantage of including non-linear terms (cubic terms for the alkanes and square terms for the N-heterocyclic compounds—see Tables 1 and 2 in relation to the regression equations and Tables 3 and 4) in the best-fitting regression equations might be taken to suggest that a simple description of compounds in terms of one-dimensional variables (corresponding to bond types or other topological features) is inadequate to account for the three-dimensional intermolecular interactions implicit in the nature of the evaporation or boiling process. Any such direct inference from the form of the non-linear terms may be misleading because of the empirical means by which these terms were generated and because all the terms implicitly incorporate some component of intermolecular interaction. It should be noted also that even for the unbranched alkanes a fifth-order polynomial in carbon number is required to produce a reasonable correlation of boiling-points between methane and tetracontane, but there is no evident physical interpretation of the magnitude and sign of all the regression coefficients.

This analysis of boiling points in relation to molecular indices is a valid application of the least-squares method provided that the variance of the boiling point is independent of this quantity. A plot of ΔT against T for the two sets of compounds proves this condition to be

valid. Transforming boiling point (easily done in BASIC with a DEF FNT statement) to either a logarithmic function or an exponential function of the negative reciprocal of temperature produced only smaller F-values and worse other indicators of goodness of fit, by comparison with the linear correlation.

What then is the use of such correlations? They may have some predictive power in the case in which measurements are not practicable or convenient. For instance, our check of boiling points of alkanes not taken as members of the basis collection was satisfactory; with regard to nitrogen compounds, we can predict the boiling point of 5-methylisoquinoline to be (264.6 ± 1.7)°C (not reported). Furthermore, a deviation from such correlations may indicate (if very large) the presence of a significant additional effect, such as hydrogen bonding, or an unusual interaction, as may apply in the case of pyridiazine, or it might suggest the need for another measurement of the property. Other correlations have already uncovered errors in this way (Laub & Pecsok, 1978).

How reliable is this type of correlation achievable by means of MULNRG or similar programs? The factors affecting the reliability are the appropriateness of the model, the accuracy and precision of the input data which form the basis of the correlation, and the precision of the computations. With at least ten significant digits carried through the computations on both computers, the latter factor was proved to be negligible despite the fact that the E matrix may be somewhat ill-conditioned (large variation in size of matrix elements); repeated inversion of a typical E matrix, or multiplication of E by its inverse G, proved that in both cases numerical precision was adequate. The accuracy and precision of input data are a significant source of potential error; the accuracy of a boiling-point measurement depends upon sample purity, atmospheric pressure (even if some correction is applied), and accuracy of thermometric measurement. We found that in some cases boiling points determined from vapour-pressure measurements (Boublik *et al.*, 1973) were expressed to 0.001°C but some such values differed by 2–3°C from other accepted values (from commonly consulted reference collections): this situation illustrates strikingly the distinction between accuracy and precision. For this reason also we have made no attempt to include weighting factors in our statistical analysis, although program modification for this purpose would be minimal. For both the alkanes and N-heterocyclic compounds, we estimate the general accuracy of the input data to be ± 2°C, and should expect on that basis a similar reliability of predicted values. Use of the correct relation, that through variance given at the end of Section 2 above, for estimation of the standard deviation of a predicted boiling point gives indeed a similar magnitude (because of the presence of negative covariance matrix elements), despite the much larger standard deviations of the regression coefficients. The fallacy of molecular additivity, with respect to boiling points, is a serious hazard in regard to predictivity, less in the case of alkanes than for N-heterocyclic compounds because there is less variation in the strengths and types of intermolecular interactions of the former compounds than for the latter collection. However, intelligent use of such correlations, with due consideration of the possibility of strong hydrogen bonding in some cases for instance, can help to prevent grossly misleading predictions solely on the basis of the structural formula of a particular molecule.

Multiple linear correlation has traditionally been little used by chemists. The contrast with the use of bivariate analysis, involving the relation of one dependent known variable to a single independent regressor, is marked. For instance, a relation between the logarithm of differences of boiling points of members of simple homologous series has been found with the wavenumbers of certain vibrational transitions of alkanes and chloroalkanes (Lielmezs, 1968). The reasons for the disparity in usage between single and multiple known regressors, despite the general difficulty of isolating one independent parameter from all others, are mainly the additional mathematical complications involving several variables, and the concomitant problems of assessing errors. Really the accessibility of modern electronic digital computers with high-level compilers or interpreters such as BASIC entirely eliminates these factors. The procedure that we have outlined and discussed above is generally applicable to many chemical problems, and the implementation of MULNRG or similar algorithms can yield an improved awareness of the multiple dependence or interrelation of many physical parameters with regard to a particular measured quantity. The book by Kendall & Stuart (1979) provides examples (non-chemical) that show how oversimplification of an analysis of a dependence by neglect of some factors leads to erroneous conclusions. For use with small computers with limited accessible core, the present algorithm may be preferable to that of Albritton, *et al.* (1976) that requires 1.5–2.1 times the array storage requirement, but in other respects the algorithms are equivalent. If theory or a model for testing indicates a non-linear dependence, then transformation of known variables or inclusion of composite regressors, as we have done, may simply remove the formal intractability. Well-defined methods of error analysis in such cases can equally be applied to complete the determination of the correlation (Clifford, 1973). We hope that this work has indicated the potential effectiveness of this approach.

Acknowledgements—We thank Profs. A. R. Katritzky and S. Broadbent for helpful comments. J. F. O. is indebted to Dr. D. J. Daley and Prof. V. R. Cane for advice on statistical matters.

REFERENCES

Akhmedov, A. G. (1979), *Russ. J. Phys. Chem.* **53**, 1366.
Albritton, D. L., Schmeltekopf, A. L. & R. N. Zare (1976) *in* Molecular Spectroscopy: Modern Research, p. 30, K. N., Ed., New York, Academic Press.
American Petroleum Institute (1974), Research Project 44, Texas A & M University, College Station, Texas.
Boublik, T., Fried, V. & Hala, E. (1973), The Vapour Pressures of Pure Substances, Amsterdam, Elsevier.
Clifford, A. A. (1973), Multivariate Error Analysis, Barking, U.K., Applied Science Publishers.
Jessup, R. S. (1937), *J. Res. NBS* **18**, 115.
Kendall, M. & Stuart, A. (1979), The Advanced Theory of Statistics, London, Griffin.
Laidler, K. J. (1956), *Can. J. Chem.* **34**, 626.
Laub, R. J. & Pecsok, R. L. (1978), Physicochemical Applications of Gas Chromatography, New York, Wiley.
Lielmezs, J. (1968), *Ind. Engng Chem. Fundam.* **7**, 315.
McLellan, A. L. (1974), Tables of Experimental Dipole Moments, El Cerrito, U.S.A., Rahara Enterprises.
Rossini, F. D. (1940), *Chem. Rev.* **27**, 1.
Tatevskii, V. M., Benderskii, V. A. & Yarovoi, S. S. (1961), Rules and Methods for Calculating the Physicochemical Properties of Paraffinic Hydrocarbons, Oxford, Pergamon Press.
Tilicheyev, M. D. & Iogansen, A. V. (1951), *Zh. fiz. khim.* **25**, 1295.
Yamane, T. (1973), Statistics, 3rd Edn, New York, Harper & Row.