Available online at www.sciencedirect.com
ScienceDirect
LWT 41 (2008) 951–958
LWT
www.elsevier.com/locate/lwt

# Practical retention index models of OV-101, DB-1, DB-5, and DB-Wax for flavor and fragrance compounds[☆]

K.L. Goodner*

*USDA, ARS, Citrus & Subtropical Products Lab, Winter Haven, FL 33884, USA*

Received 24 April 2007; received in revised form 5 July 2007; accepted 16 July 2007

## Abstract

High-quality regression models of gas chromatographic retention indices were generated for OV-101 ($R = 0.997$), DB-1 ($R = 0.998$), DB-5 ($R = 0.997$), and DB-Wax ($R = 0.982$) using 91, 57, 94, and 102 compounds, respectively. The models were generated using a second-order equation including the cross product utilizing two easily obtained variables, boiling point and the log octanol-water coefficient. Additionally, a method for determining outlier data (the GOodner Outlier Determination (GOOD) method) is presented, which is a combination of several outlier tests and is less prone to discarding legitimate data.

Published by Elsevier Ltd. on behalf of Swiss Society of Food Science and Technology.

## 1. Introduction

Determination of unknowns in gas chromatography (GC) requires two independent forms of identification such as the retention time on two different chromatographic columns, retention time and mass spectral match, or retention time and aromatic match (Harris, 1987). Because retention times vary depending on the temperature programming of the GC, Kovat introduced a relative retention index scheme (Kovats, 1958). A relative retention index uses a series of standards, often *n*-alkanes, with other compounds referenced against the standards using the equation

$$I = 100\left[n + (N - n)\frac{\log t'_{\mathrm{r}}(\mathrm{unknown}) - \log t'_{\mathrm{r}}(n)}{\log t'_{\mathrm{r}}(N) - \log t'_{\mathrm{r}}(n)}\right],$$

where $n$ is the number of carbon atoms in the smaller alkane, $N$ the number of carbon atoms in the larger alkane, $t'_{\mathrm{r}}(n)$ the adjust retention time of the smaller alkane, and $t'_{\mathrm{r}}(N)$ is the adjusted retention time of the larger alkane (Harris, 1987).

There are compilations of retention indices on different stationary phase columns for many compounds, which are useful for identifying unknown compounds. However, these compilations are not complete and situations arise where one has a potential identification from a mass spectral match, but no information on retention index. Unless the compound is available or readily synthesized, it is unlikely that a positive identification using a standard can be determined. Another common situation is determining if co-eluting compounds on one type of column phase will be chromatographically separated on a different phase where the retention indices are not known. Additionally, knowing the retention indices of two target analytes a priori would ease method development. One potential aid to the researcher is a mathematical model of retention indices for the column stationary phase they are using.

There have been numerous mathematical models of GC retention indices on various stationary phases (Anker, Jurs, & Edwards, 1990; Bermejo & Guillén, 1987; Buydens & Massart, 1983; Gerasimenko & Nabivach, 1990; Hale et al., 1985; Hasan & Jurs, 1988; Héberger, Görgényi, & Sjöström, 2000; Jalali-Heravi & Fatemi, 2001; Ośmiałowski, Halkiewicz, & Kaliszan, 1986; Ośmiałowski, Halkiewicz, Radecki, & Kaliszan, 1985; Raymer, Wiesler, & Novotny, 1985; Robbat & Kalogeropoulos, 1990; Rohrbagh & Jurs,

---

*Tel.: +1 863 293 4133; fax: +1 863 299 8678.
*E-mail address:* goodner@citrus.usda.gov

0023-6438/$30.00 Published by Elsevier Ltd. on behalf of Swiss Society of Food Science and Technology.
doi:10.1016/j.lwt.2007.07.007

1985; Sabljić, 1985; Stanton & Jurs, 1989; Whalen-Pedersen & Jurs, 1981). One study used artificial neural networks to model retention indices on a Carbowax 20M stationary phase (Jalali-Heravi & Fatemi, 2001). There were 35 descriptors used in the neural network model that were based on electronic, geometric, topological, and a single physicochemical property resulting in a regression coefficient of the calculated and experimental indices of $R = 0.984$. Jurs and coworkers have used their in-house created ADAPT platform to generate descriptors for modeling retention indices (Anker et al., 1990; Hasan & Jurs, 1988; Rohrbagh & Jurs, 1985; Stanton & Jurs, 1989; Whalen-Pedersen & Jurs, 1981) with good results (generally $R > 0.99$).

Héberger et al. (2000) used a partial least-squares model of several physical properties to determine retention index and classify compounds as either ketones or aldehydes. The researchers used five terms and achieved good results, but were unable to classify compounds based solely on their retention data.

Many of these retention index models are quite good, providing excellent models with good predictive power. However, the drawback for the typical laboratory worker is their complexity. This report covers a series of models based on boiling point ($T_B$) and log of the octanol–water partition coefficient ($\log K_{ow}$, also called $\log P$) which are both easily obtained via reference books, the internet, and/or free programs. Additionally, a new algorithm for determining outlier data is presented and utilized in the analysis.

## 2. Materials and methods

### 2.1. Data collection

The data used in this report were collected from numerous sources. The Kovat's retention indices were collected from Acree's Flavornet (Acree, 2006), Rouseff's database (Rouseff, 2006), Adam's book (Adams, 1995), Kondjoyan and Berdagué's compilation (Kondjoyan & Berdagué, 1996), the NIST MS library (Stein et al., 1998), and/or Flavorworks (Flavometrics, NJ) for OV-101, DB-1, DB-5, and Carbowax. The boiling point ($T_B$) data was taken from the Handbook of Chemistry and Physics (Anonymous, 1990) and/or from internet searches. Retention indices and boiling points were averaged if available from multiple sources. Log $K_{ow}$ was determined using the Kowwin module in the Estimation Program Interface (EPI) suite (Anonymous, 2000). The data used for this study are provided in Table 1.

### 2.2. Data modeling

The data were then modeled in Statistica version 7 (Statsoft, Tulsa, OK) using the general linear model module and using $T_B$, $\log K_{ow}$, $T_B^2$, $(\log K_{ow})^2$, and $T_B \log K_{ow}$ for a second-order model with a cross product term. The residuals were examined for Potential outliers using the GOodner Outlier Determination (GOOD) Method, detailed below, and re-ran with outliers removed.

### 2.3. Outlier analysis

The data (both input data and results) were analyzed for outliers using the GOOD method which was programmed as a Microsoft Excel add-in (Microsoft, Redmond, WA) and is available from the author. This method is a combination of the Grubb's test, the Dixon's $Q$-test, and a modified "huge" rule. The Grubb's test is the basis for the ASTM E-178-02 (Anonymous, 2002) and is calculated by

$$T_1 = \frac{\bar{x} - x_1}{s}, \qquad T_n = \frac{x_n - \bar{x}}{s},$$

where the datapoints are arranged in increasing order from the smallest $x_1$ to the largest $x_n$. $T_1$ or $T_n$ is the data point in question (possible outlier) depending if the questionable data point is the smallest or largest value, $\bar{x}$ is the mean of all samples, and $s$ is the standard deviation of the data set. The calculated $T$-value is compared to a table of critical values and if the calculated statistic is larger than the critical value, then the data point is considered an outlier.

When you have a possible outlier, usually you have a cluster of data points along with an isolated (outlying) point. Dixon's $Q$-test relies on determining the ratio of the "gap" (distance from "possible outlier" to the main cluster of data) to the range and is dependent upon the number of data points in the sample (Dixon, 1951).

$$r = \frac{\text{gap}}{\text{range}},$$

If $3 \leqslant n \leqslant 7$

$$r_{10} = \frac{x_2 - x_1}{x_n - x_1} \quad \text{if smallest value is suspected}$$

$$r_{10} = \frac{x_n - x_{n-1}}{x_n - x_1} \quad \text{if largest value is suspected}$$

If $8 \leqslant n \leqslant 10$

$$r_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1} \quad \text{if smallest value is suspected}$$

$$r_{11} = \frac{x_n - x_{n-1}}{x_n - x_2} \quad \text{if largest value is suspected}$$

If $11 \leqslant n \leqslant 13$

$$r_{21} = \frac{x_3 - x_1}{x_{n-1} - x_1} \quad \text{if smallest value is suspected}$$

$$r_{21} = \frac{x_n - x_{n-2}}{x_n - x_2} \quad \text{if largest value is suspected}$$

If $14 \leqslant n \leqslant 30$

$$r_{22} = \frac{x_3 - x_1}{x_{n-2} - x_1} \quad \text{if smallest value is suspected}$$

$$r_{22} = \frac{x_n - x_{n-2}}{x_n - x_3} \quad \text{if largest value is suspected}$$

Table 1
Linear retention indices for OV-101, DB-1, DB-5, Wax columns along with boiling point ($T_B$) and log $K_{ow}$

| Compound | CAS | OV-101 | DB-1 | DB-5 | Wax | $T_B$ | log $K_{ow}$ |
|---|---|---|---|---|---|---|---|
| Pentane | 109-66-0 | 500 | 500 | 500 | 500 | 35.5 | 2.80 |
| Hexane | 110-54-3 | 600 | 600 | 600 | 600 | 69 | 3.29 |
| Heptane | 142-82-5 | 700 | 700 | 700 | 700 | 98 | 3.78 |
| Octane | 111-65-9 | 800 | 800 | 800 | 800 | 125 | 4.27 |
| Nonane | 111-84-2 | 900 | 900 | 900 | 900 | 151 | 4.76 |
| Decane | 124-18-5 | 1000 | 1000 | 1000 | 1000 | 174 | 5.25 |
| Undecane | 1120-21-4 | 1100 | 1100 | 1100 | 1100 | 196 | 5.74 |
| Dodecane | 112-40-3 | 1200 | 1200 | 1200 | 1200 | 216 | 6.23 |
| Tridecane | 629-50-5 | 1300 | 1300 | 1300 | 1300 | 234 | 6.73 |
| Tetradecane | 629-59-4 | 1400 | 1400 | 1400 | 1400 | 253 | 7.22 |
| Pentadecane | 629-62-9 | 1500 | 1500 | 1500 | 1500 | 269 | 7.71 |
| Hexadecane | 544-76-3 | 1600 | 1600 | 1600 | 1600 | 287 | 8.20 |
| Heptadecane | 629-78-7 | 1700 | 1700 | 1700 | 1700 | 302 | 8.49 |
| Octadecane | 593-45-3 | 1800 | 1800 | 1800 | 1800 | 317 | 9.18 |
| Nonadecane | 629-92-5 | 1900 | 1900 | 1900 | 1900 | 330 | 9.67 |
| Eicosane | 112-95-8 | 2000 | 2000 | 2000 | 2000 | 343 | 10.16 |
| Methanethiol | 74-93-1 | 483 |  |  | 683 | 6 | 0.78 |
| Acetaldehyde | 75-07-0 | 418 |  |  | 706 | 21 | −0.17 |
| Dimethyl sulfide | 75-18-3 | 527 |  | 510 | 734 | 38 | 0.92 |
| (E)-2hexenal | 6789-80-6 | 870 | 828 | 854 | 1148 | 148 | 1.58 |
| Butanal | 123-72-8 |  |  | 596 | 822 | 75 | 0.82 |
| Ethyl acetate | 141-78-6 | 595 | 600 | 614 | 898 | 77 | 0.86 |
| Ethanol | 64-17-5 | 500 |  |  | 928 | 78 | −0.14 |
| Butane-2,3-dione | 431-03-8 |  |  | 596 |  | 88 | −1.34 |
| Ethyl propanoate | 105-37-3 | 691 |  | 714 | 950 | 99 | 1.36 |
| Methyl butanoate | 623-42-7 | 705 |  | 724 | 990 | 102 | 1.36 |
| 2-methyl-1-propanol | 78-83-1 | 616 |  | 647 | 1103 | 108 | 0.77 |
| Dimethyl disulfide | 624-92-0 | 725 |  | 744 | 1078 | 110 | 1.87 |
| Methylbenzene | 108-88-3 | 756 |  | 769 | 1038 | 110 | 2.54 |
| Ethyl isobutanoate | 97-62-1 | 746 |  | 754 | 964 | 112 | 1.77 |
| 1-butanol | 71-36-3 | 655 |  | 675 | 1142 | 117 | 0.84 |
| Ethyl butanoate | 105-54-4 | 705 | 797 | 800 | 1038 | 121 | 1.85 |
| 3-Methylbutanol | 123-51-3 | 736 | 720 | 738 | 1208 | 130 | 1.26 |
| 2-Methylbutanol | 137-32-6 | 738 |  | 755 | 1208 | 128 | 1.26 |
| Hexanal | 66-25-1 | 780 | 792 | 800 | 1088 | 130 | 1.80 |
| Ethyl 2-methylbutanoate | 7452-79-1 | 837 | 835 | 846 | 1056 | 133 | 2.26 |
| 1-Pentanol | 71-41-0 | 756 |  | 768 | 1244 | 137 | 1.33 |
| 3-Methyl-2-buten-1-ol | 556-82-1 | 762 |  | 778 | 1127 | 140 | 1.17 |
| Ethyl pentanoate | 539-82-2 | 884 | 875 | 899 | 1138 | 144 | 2.34 |
| Heptanal | 111-71-7 | 883 |  | 901 | 1184 | 153 | 2.29 |
| α-Pinene | 80-56-8 | 926 | 929 | 939 | 1035 | 156 | 4.27 |
| 1-Hexanol | 111-27-3 | 858 | 846 | 865 | 1362 | 156 | 1.82 |
| (Z)-3-hexen-1-ol | 928-96-1 | 844 |  | 858 | 1392 | 156 | 1.61 |
| Camphene | 79-92-5 | 957 | 955 | 953 | 1080 | 159 | 4.35 |
| Furfural | 98-01-1 | 800 | 798 | 852 | 1432 | 162 | 0.83 |
| *β*-Pinene | 127-91-3 | 985 | 983 | 980 | 1118 | 164 | 4.35 |
| Methional | 3268-49-3 | 864 | 867 | 911 | 1468 | 166 | 0.41 |
| *β*-Myrcene | 123-35-3 | 990 | 991 | 990 | 1158 | 167 | 4.88 |
| Ethyl hexanoate | 123-66-0 | 983 | 982 | 1002 | 1231 | 168 | 2.83 |
| Hexyl acetate | 142-92-7 | 1012 | 994 | 1011 | 1268 | 169 | 2.83 |
| Octanal | 124-13-0 | 985 | 985 | 1004 | 1291 | 171 | 2.78 |
| α-Terpinene | 99-86-5 | 990 | 999 | 1015 | 1184 | 174 | 4.75 |
| 1,8-Cineole | 470-82-6 | 1015 | 1009 | 1031 | 1222 | 176 | 3.13 |
| Limonene | 5989-27-5 | 1013 | 1022 | 1030 | 1208 | 176 | 4.83 |
| *p*-Cymene | 99-87-6 | 1025 |  | 1026 | 1274 | 177 | 4.00 |
| Ibenzaldehyde | 100-52-7 | 943 |  | 964 | 1525 | 178 | 1.71 |
| *γ*-Terpinene | 99-85-4 | 1058 | 1022 | 1072 | 1249 | 182 | 4.75 |
| Ethyl heptanoate | 106-30-9 | 1083 |  | 1098 | 1397 | 188 | 3.32 |
| Phenylacetaldehyde | 122-78-1 | 1004 |  | 1047 | 1650 | 195 | 1.54 |
| 1-Octanol | 111-87-5 | 1061 | 1051 | 1071 | 1558 | 196 | 2.81 |
| *p*-Cresol | 106-44-5 | 1058 |  | 1074 |  | 202 | 2.06 |
| Benzyl alcohol | 100-51-6 |  | 1003 |  | 1879 | 204 | 1.08 |
| Guaiacol | 90-05-1 | 1056 |  | 1088 | 1872 | 205 | 1.34 |

Table 1 (*continued*)

| Compound | CAS | OV-101 | DB-1 | DB-5 | Wax | $T_B$ | $\log K_{ow}$ |
|---|---|---|---|---|---|---|---|
| Citronellal | 106-23-0 | 1134 | 1138 | 1157 | 1492 | 207 | 3.53 |
| Ethyl octanoate | 106-32-1 | 1185 | | 1198 | 1497 | 207 | 3.81 |
| Decanal | 112-31-2 | 1188 | | 1207 | 1500 | 208 | 3.76 |
| Octyl acetate | 112-14-1 | | 1137 | 1212 | 1478 | 211 | 3.81 |
| Terpinen-4-ol | 2438-10-0 | 1182 | 1160 | 1180 | 1606 | 212 | 3.33 |
| Nonyl acetate | 143-13-5 | | | | 1583 | 212 | 4.30 |
| 1-Nonanol | 143-08-8 | 1161 | 1155 | 1162 | 1654 | 215 | 3.30 |
| α-Terpineol | 10482-56-1 | | 1184 | 1192 | 1700 | 218 | 3.33 |
| Phenethyl mercaptan | 4410-99-5 | | | 1182 | 1688 | 218 | 2.98 |
| Naphthalene | 91-20-3 | 1197 | | 1196 | 1718 | 218 | 3.17 |
| Linalyl acetate | 115-95-7 | 1244 | | 1259 | 1566 | 220 | 4.39 |
| α-Terpinyl acetate | 80-26-2 | | | | 1681 | 220 | 4.34 |
| Citronellol | 106-22-9 | 1229 | 1225 | 1235 | 1768 | 225 | 3.56 |
| 4-Methylacetophenone | 122-00-9 | 1174 | | 1186 | | 226 | 2.22 |
| Ethyl phenylacetate | 101-97-3 | 1235 | 1213 | 1248 | 1785 | 229 | 2.57 |
| Carvone | 99-49-0 | 1236 | 1208 | 1248 | 1765 | 229 | 3.07 |
| Geraniol | 106-24-1 | 1258 | 1238 | 1266 | 1852 | 229 | 3.47 |
| 1-Decanol | 112-30-1 | 1263 | | 1268 | 1767 | 231 | 3.79 |
| 2-Undecanone | 112-12-9 | 1276 | | 1294 | | 232 | 3.69 |
| Citronellyl acetate | 150-84-5 | 1335 | | 1356 | 1639 | 240 | 4.56 |
| Ethyl decanoate | 110-38-3 | 1379 | | 1396 | 1636 | 245 | 4.79 |
| α-Cubebene | 17699-14-8 | 1328 | 1360 | 1348 | 1472 | 246 | 6.73 |
| Eugenol | 97-53-0 | 1381 | 1335 | 1360 | 2164 | 254 | 2.73 |
| Methyleugenol | 93-15-2 | 1390 | | 1404 | | 254 | 3.03 |
| Geranyl acetone | 3796-70-1 | 1431 | | 1416 | 1840 | 256 | 4.36 |
| α-Farnesene | 502-61-4 | 1518 | 1456 | 1504 | | 260 | 7.10 |
| α-Ionone | 127-41-3 | 1408 | | 1424 | 1840 | 261 | 4.29 |
| Valencene | 4630-07-3 | | 1497 | 1490 | 1746 | 274 | 6.30 |
| Tetradecanol | 112-72-1 | | | 1676 | 2177 | 289 | 5.75 |
| 2-Octanol | 123-96-6 | 988 | | 990 | 1332 | 179 | 2.73 |
| 1,4-Cineole | 470-67-7 | 1001 | 994 | 1017 | 1188 | 173 | 3.13 |
| Linalool | 78-70-6 | 1085 | 1098 | 1100 | 1544 | 199 | 3.38 |
| Nerol | 106-25-2 | 1234 | 1212 | 1230 | 1792 | 225 | 3.47 |
| Nonanal | 124-19-6 | 1087 | 1091 | 1106 | 1397 | 195 | 3.27 |
| α-Phellandrene | 99-83-2 | 1009 | | 1006 | 1170 | 172 | 4.62 |
| Sabinene | 3387-41-5 | 970 | | 974 | 1125 | 164 | 4.69 |
| β-Phellandrene | 555-10-2 | 1035 | 985 | 1042 | 1245 | 172 | 4.70 |
| Acetic acid | 64-19-7 | 710 | | | | 118 | 0.09 |
| Propanoic acid | 79-09-4 | | | | 1525 | 141 | 0.58 |
| 1-Butanoic acid | 107-92-6 | | | | 1628 | 162 | 1.07 |
| Pentanoic acid | 539-82-2 | | | | 1698 | 186 | 1.56 |
| Hexanoic acid | 142-62-1 | | | | 1797 | 203 | 2.05 |
| Heptanoic acid | 111-14-8 | | | | 1900 | 223 | 2.54 |
| Octanoic acid | 124-07-2 | | | | 2065 | 237 | 3.03 |
| Nonanoic acid | 112-05-0 | | | | 2185 | 268 | 3.52 |
| 2-Methylbutanoic acid | 116-53-0 | | | | 1672 | 176 | 1.49 |

The calculated "$r$" value is compared to a table of critical values and if the calculated statistic is larger than the table value, the data point is considered an outlier.

The huge rule (Marascuilo, 1971) is determined by calculating the mean and standard deviation of the data set with the questionable data point excluded and determining how many standard deviations the questionable data point would be using

$$M = \frac{(|x_i - \bar{X}'|)}{S'},$$

where $\bar{X}'$ and $S'$ are the mean and standard deviation calculated excluding the data point in question, $x_i$. The rule is that if the data point is more than four standard deviations away from the reduced data set, then it is an outlier. It is extremely unlikely ($p<0.00005$) that data points would be more than four standard deviations away from the mean assuming a normal distribution. The GOOD method uses a modified test using the $t$-distribution (using an $\alpha/10$ confidence level) instead of the normal distribution and adds confidence levels. The calculation is performed similarly, but the critical value, $C_t$, is no longer 4, but rather determined by calculating the $t$-value at $n-1$ degrees of freedom for the confidence level desired.

$$C_t = t_{\alpha/10}.$$

The GOOD method performs all three outlier tests and determines the data point to be an outlier only if all three of the tests return a positive result. This results in a more conservative test than the three individually. As is noted in ASTM-E178, repeated application of outlier tests reduces the significance level and one would use $\alpha/k$ for $k$ tests to approximate an overall significance level of $\alpha$. However, by requiring that at all three of the tests return a positive value, the overall significance level of the GOOD method is less than $\alpha$. While this is a laborious technique for determining if datasets have possible outliers, the Excel add-in introduces a new function that performs all three tests and returns a “yes”/”no” answer making the method readily usable to test all data sets.

The GOOD method was analyzed as compared to its constituent tests using randomly generated Gaussian data sets. Eight thousand 15-point data set was generated in Statistica and analyzed in Excel. The data set was analyzed as both a 15 datapoint set and as 3 individual data sets of 5 points each were analyzed. The results are tabulated in Table 2.

## Results and discussion

The resulting equations from the regression modeling are

$$\begin{aligned}
\mathrm{RI}_{\text{OV-101}} &= 1.1623T_{\mathrm{B}} + 0.0064T_{\mathrm{B}}^2 + 35.2861K_{\mathrm{OW}} \\
&\quad - 9.0068K_{\mathrm{OW}}^2 + 0.2756T_{\mathrm{B}}K_{\mathrm{OW}} + 421.4386, \\
\mathrm{RI}_{\text{DB-1}} &= 1.5901T_{\mathrm{B}} + 0.0040T_{\mathrm{B}}^2 + 14.0606K_{\mathrm{OW}} \\
&\quad - 7.7975K_{\mathrm{OW}}^2 + 0.3396T_{B}K_{\mathrm{OW}} + 436.2845, \\
\mathrm{RI}_{\text{DB-5}} &= 1.7273T_{\mathrm{B}} + 0.0049T_{\mathrm{B}}^2 + 17.2140K_{\mathrm{OW}} \\
&\quad - 7.0459K_{\mathrm{OW}}^2 + 0.2663T_{B}K_{\mathrm{OW}} + 422.8297, \\
\mathrm{RI}_{\text{DB-Wax}} &= 2.8662T_{\mathrm{B}} + 0.0199T_{\mathrm{B}}^2 - 58.2317K_{\mathrm{OW}} \\
&\quad + 3.0049K_{\mathrm{OW}}^2 - 0.4975T_{B}K_{\mathrm{OW}} + 630.7004.
\end{aligned}$$

The regression results for the four column stationary phases is presented in Table 3. The models include the cross product of $T_{\mathrm{B}}$ and $\log K_{\mathrm{ow}}$, which has not previously been used. The multiple $R$, which is a measure of how well the model fits the data set, is excellent for the OV-101, DB-1, and DB-5 stationary phases with $R \geqslant 0.997$ and the DB-Wax multiple $R$ was 0.982. Additionally, various quality measures are provided such as the average percent error and the percentage of the samples that are within 3% and 5% of the model. Also provided is the percentage from the model needed to account for 75% and 95% of the data set. For example, 95% of the data set for DB-5 are within 5.7% of the model. Examining Table 2, one can see that approximately about 80% of the cases were within 3%

Table 3
Regression models for different phase columns

| Model term | Column phase | | | |
|---|---|---|---|---|
| | OV-101 | DB-1 | DB-5 | DB-Wax |
| Intercept | 421.4386 | 436.2845 | 422.8297 | 630.7004 |
| $T_{\mathrm{B}}$ | 1.1623 | 1.5901 | 1.7273 | 2.8662 |
| $T_{\mathrm{B}}^2$ | 0.0064 | 0.0040 | 0.0049 | 0.0199 |
| $K_{\mathrm{ow}}$ | 35.2861 | 14.0606 | 17.2140 | −58.2317 |
| $K_{\mathrm{ow}}^2$ | −9.0068 | −7.7975 | −7.0459 | 3.0049 |
| $T_{\mathrm{B}} \times K_{\mathrm{ow}}$ | 0.2756 | 0.3396 | 0.2663 | −0.4975 |
| Multiple $R$ | 0.997 | 0.998 | 0.997 | 0.982 |
| $n$ | 91 | 57 | 94 | 102 |
| Average % error | 2.11 | 1.41 | 1.90 | 4.41 |
| ±3% | 74.7 | 89.5 | 79.8 | 38.2 |
| ±5% | 91.2 | 96.5 | 94.7 | 63.7 |
| 75% cases± | 3.3 | 2.0 | 2.7 | 6.2 |
| 95% cases± | 6.5 | 4.4 | 5.7 | 11.2 |

*Note*: The table is interpreted as follows for OV-101: 74.7% of the cases are with 3% of the reported values and 75% of the cases are within 3.3% of reported values.

Table 2
Performance of the GOOD method for determining outlying data points as compared to its constituent tests

| Dataset size | Number of datasets | Confidence level | Datapoint in question | Outlier method | | | |
|---|---|---|---|---|---|---|---|
| | | | | ASTM | *Q*-test | “Huge” | GOOD method |
| 15 | 8000 | 95% | Maximum | 388 | 323 | 397 | 241 |
| | | | Minimum | 430 | 326 | 439 | 230 |
| | | 99% | Maximum | 69 | 72 | 80 | 31 |
| | | | Minimum | 69 | 62 | 75 | 32 |
| | | 99.5% | Maximum | 27 | 39 | 33 | 19 |
| | | | Minimum | 34 | 32 | 38 | 13 |
| 5 | 24000 | 95% | Maximum | 1217 | 1227 | 940 | 939 |
| | | | Minimum | 1199 | 1200 | 919 | 916 |
| | | 99% | Maximum | 234 | 280 | 281 | 227 |
| | | | Minimum | 214 | 268 | 254 | 212 |
| | | 99.5% | Maximum | 116 | 125 | 160 | 111 |
| | | | Minimum | 114 | 117 | 151 | 110 |

error and about 95% are within 5%. Thus, 3% and 5% were chosen as reasonable limits based on variations of reported retention indices. OV-101 and DB-1 are very similar stationary phases, 100% polydimethylsiloxane, but the DB-1 is bonded and crosslinked, which apparently led to differences in retention indices reported for numerous compounds with the average difference being 1.7%.

The models for OV-101 and wax stationary phases were tested by comparing the results of these models with those previously published by Anker et al. (1990). In this paper, approximately 115 compounds were modeled using a linear regression of 7 variables. The average absolute percent error was 0.98% and 1.16% for OV-101 and Wax stationary phases, respectively. The models presented here and applied to the data presented in Anker's paper showed a 2.9% and 3.6% average absolute error for OV-101 and Wax stationary phases, respectively. However, it was noted that there was a systematic overestimation for both data sets. Applying a correction factor reduced the average absolute error to 1.9% and 2.7%, respectively.

Fig. 1 is a multiple plot of the observed versus predicted retention indices. The line indicates a theoretical perfect fit. The difference between the predicted value and the actual value is the residual. The residuals were analyzed for potential outliers using the GOOD method. These potential outliers were inspected to determine if there was a logical reason to exclude from the model. Some compounds had suspicious input parameters such as a highly unlikely reported boiling point (e.g., a reduced pressure $T_B$ reported as an atmospheric $T_B$). A few potential outliers were merely transcription errors (e.g., inputting 218 °C instead of 128°C as the $T_B$) which were included in the model after correction. Some systematic compounds were removed from the models. For example, carboxylic acids performed very poorly on the non-wax columns and were excluded from the models. This is not unexpected as carboxylic acids often perform poorly on non-polar columns, resulting in poor peak shape with severe tailing.

Table 4 is a comparison of this work with other similar reports from the literature. While some of the models appear to perform better than the one presented here, none

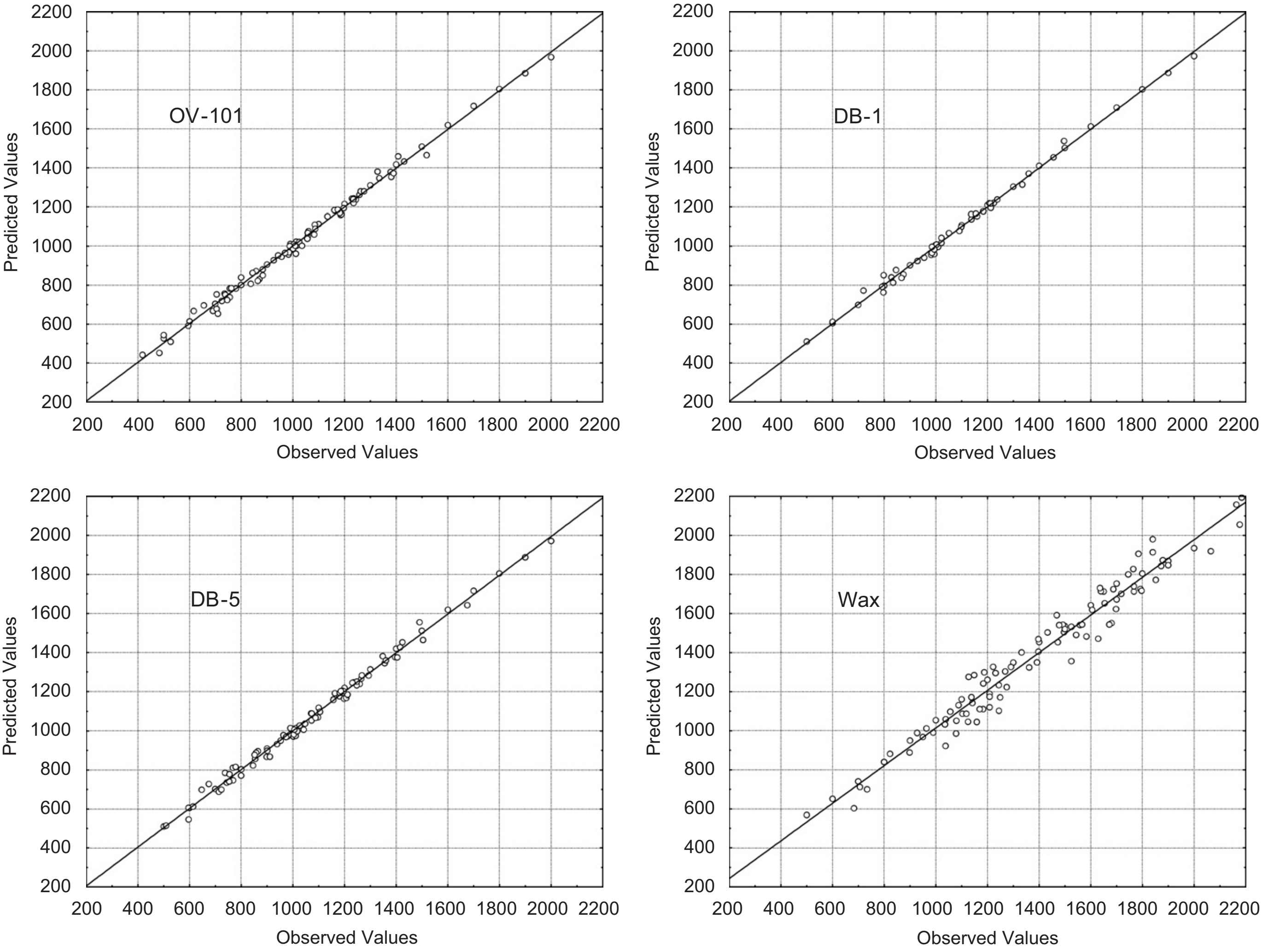

Fig. 1. Graphical plots of observed versus predicted retention indices for the (a) OV-101, (b) DB-1, (c) DB-5, and (d) DB-Wax column types.

Table 4
A comparison of various models from literature

| Reference | Phases | Number of variables | Variables | $R$-range |
|---|---|---|---|---|
| Héberger et al. (2000) | HP-1, HP-50, DB-210, HP-Wax | 5 | $T_B$, $V_M$, $W_m$, $R_m$, $\log K_{ow}$ | 0.980–0.995 |
| Stanton and Jurs (1989) | OV-101, Wax | 6 | ADAPT | 0.993–0.997 |
| Rohrbagh and Jurs (1985) | PONA, DB-1, DB-5 | 2–4 | $T_B$, $1/R_m$, $\log K_{ow}$, $I_m$, $W_m$ | 0.996–0.998 |
| Anker et al. (1990) | OV-101, Wax | 7 | ADAPT | 0.997–0.999 |
| Bermejo and Guillén (1987) | SE-30, Dioctyl Phthlate, Oronite Niw, Pluronic F88, Carbowax 1000 | 7 | $T_B$, $V_M$, $R_m$, $V_W$, $D$, $N_c$, $^1X$ | 0.994–0.999 |
| Raymer et al. (1985) | SE-30 | 1–3 | $X$ | 0.946–0.988 |
| Gerasimenko and Nabivach (1990) | HPE, OV-101, SE-30 | 4 | $V_W$, $X$, $T_B$, $R_M$, $V_M$ | 0.997–1.000 |
| Ośmiałowski et al. (1985) | OV-101 | 2–3 | $E_T$, $\Delta$, $E_{HOMO}$ | 0.93–0.96 |
| This report | OV-101, DB-1, DB-5, Wax | 2 | $T_B$, $\log K_{ow}$ | 0.982–0.998 |

$T_B$, boiling temperature; $V_M$, molar volume; $W_m$, molecular weight; $R_m$, molar refraction; $I_m$, moment of inertia; $\log K_{ow}$, log of octanol–water partition coefficient; ADAPT, various variables from modeling program; $X$, molar connectivity; $V_W$, Van der Waals' volume; $E_T$, total energy; $\Delta$, polarity parameter; $E_{HOMO}$, energy of HOMO.

combine the simplicity and easily accessible input data. From the table it is clear that several descriptors are fairly common in the modeling of retention indices. Boiling point, molar refraction, $\log K_{ow}$, and molar volume seem to be some fairly common descriptors used. Of those, boiling point, $\log K_{ow}$, and molar volume are likely the easiest to be found in literature. From a user's perspective, the fewer number of descriptors needed, coupled with the ease of finding or determining the values of the descriptors, is ideal. Boiling point is readily available from numerous resources, including the CRC Handbook of Chemistry and Physics, material safety data sheets (MSDS), and internet searches. $\log K_{ow}$ is also easily obtainable from several sources. One source is by using the EPI Suite software, which is currently freely downloadable from the Environmental Protection Agency (http://www.epa.gov/opptintr/exposure/pubs/episuitedl.htm), and also available through some internet webpages (http://www.logp.com, http://146.107.217.178/lab/alogps/start.html, and http://www.syrres.com/esc/est_kowdemo.htm).

As an example of using these models in the lab consider DB-Wax co-eluting compounds ethyl butanoate and isopropyl butanoate. If the retention indices were not known for these two compounds, or one of them, then this method could indicate whether they would be co-eluting on a different column phase. The boiling points and $\log K_{ow}$ for ethyl butanoate and isopropyl butanoate are 121 °C, 1.85 and 131 °C, 2.26, respectively. Using the equations above the retention indices for ethyl butanoate and isopropyl butanoate are 753 and 800 for OV-101, 763 and 806 for DB-1, and 771 and 815 for DB-5. These average 45 unit separation with errors of reported values of 3.5%. These values indicate that you could use any of those columns to achieve separation of the compounds.

By using the models presented here, one would be able to narrow a list of potential unknown identifications quickly and easily to a more manageable list for further analysis, which would be beneficial to many researchers around the world who have limited funds for purchasing standards or access to synthesis equipment. One could also determine a priori if co-eluting compounds would be resolved on a different stationary phase, aid in determining which column phase would be best suited for separating analytes of interest, and aid in the selection of internal standards. The GOOD method is a general purpose algorithm which can be used for determining outliers for normally distributed data sets.

## References

Acree, T. (2006). Flavornet 〈http://www.flavornet.org/flavornet.html〉.

Adams, R. P. (1995). *Identification of essential oil components by gas chromatography/mass spectroscopy*. Carol Stream, IL: Allured Publishing Corp.

Anker, L. S., Jurs, P. C., & Edwards, P. A. (1990). Quantitative structure–retention relationship studies of odor-active aliphatic compounds with oxygen-containing functional groups. *Analytical Chemistry*, *62*, 2676–2684.

Anonymous. (1990). *CRC handbook of chemistry and physics*. Boca Raton, FL: CRC Press, Inc.

Anonymous. (2000). EPI Suite. Washington, DC, USA: Office of Pollution Prevention Toxics and Syracuse Research Corporation. U.S. Environmental Protection Agency.

Anonymous. (2002). E178-02. Standard practice for dealing with outlying observations. *Annual book of ASTM standards*, 14.02, pp. 56–73.

Bermejo, J., & Guillén, M. D. (1987). Prediction of Kovats retention index of saturated alcohols on stationary phases of different polarity. *Analytical Chemistry*, *59*, 94–97.

Buydens, L., & Massart, D. L. (1983). Prediction of gas chromatographic retention indexes with topological, physicochemical, and quantum chemical parameters. *Analytical Chemistry*, *55*, 738–744.

Dixon, W. J. (1951). Ratios involving extreme values. *Annals of Mathematical Statistics*, 68–78.

Gerasimenko, V. A., & Nabivach, V. M. (1990). Relationships between gas chromatographic retention indices and molecular structure of aromatic hydrocarbons. *Journal of Chromatography*, *498*(2), 357–366.

Hale, M. D., Hileman, F. D., Mazer, T., Shell, T. L., Noble, R. W., & Brooks, J. J. (1985). Mathematical modeling of temperature programmed capillary gas chromatographic retention indexes for polychlorinated dibenzofurans. *Analytical Chemistry*, *57*, 640–648.

Harris, D. C. (1987). *Quantitative chemical analysis*. New York: W.H. Freeman and Company.

Hasan, M. N., & Jurs, P. C. (1988). Computer-assisted prediction of gas chromatographic retention times of polychlorinated biphenyls. *Analytical Chemistry*, *60*, 978–982.

Héberger, K., Görgényi, M., & Sjöström, M. (2000). Partial least squares modeling of retention data of oxo compounds in gas chromatography. *Chromatographia*, *51*(9–10), 595–600.

Jalali-Heravi, M., & Fatemi, M. H. (2001). Artificial neural network modeling of Kovats retention indices for noncyclic and monocyclic terpenes. *Journal of Chromatography A*, *915*(1–2), 177–183.

Kondjoyan, N., & Berdagué, J. L. (1996). *A compilation of relative retention indices for the analysis of aromatic compounds*. Theix, France: Laboratoire Flavour.

Kovats, E. (1958). Gaz-chromatographische Charakterisierung organishcher Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helvetica Chimica Acta*, *41*, 1915–1932.

Marascuilo, L. A. (1971). *Statistical methods for behavioral science research*. New York, NY: McGraw-Hill.

Ośmlałowski, K., Halkiewicz, J., & Kaliszan, R. (1986). Quantum chemical parameters in correlation analysis of gas–liquid chromatographic retention indices of amines. II. Topological electronic index. *Journal of Chromatography*, *361*, 63–69.

Ośmlałowski, K., Halkiewicz, J., Radecki, A., & Kaliszan, R. (1985). Quantum chemical parameters in correlation analysis of gas–liquid chromatographic retention indices of amines. *Journal of Chromatography*, *356*, 53–60.

Raymer, J., Wiesler, D., & Novotny, M. (1985). Structure-retention studies of model ketones by capillary gas chromotography. *Journal of Chromatography*, *325*, 13–22.

Robbat, A., Jr, & Kalogeropoulos, C. (1990). Prediction of gas chromatographic retention indexes for polychlorinated dibenzofurans. *Analytical Chemistry*, *62*, 2684–2688.

Rohrbagh, R. H., & Jurs, P. C. (1985). Prediction of gas chromatographic retention indexes of selected olefins. *Analytical Chemistry*, *57*, 2770–2773.

Rouseff, R. L. (2006). Flavor Database 〈http://www.lal.ufl.edu/crec_websites/Rouseff/Website2002/Subpages/database_f_Frameset.htm〉.

Sabljić, A. (1985). Calculation of retention indices by molecular topology chlorinated benzenes. *Journal of Chromatography*, *319*, 1–8.

Stanton, D. T., & Jurs, P. C. (1989). Computer-assisted prediction of gas chromatographic retention indices of pyrazines. *Analytical Chemistry*, *61*, 1328–1332.

Stein, S. E. D., Mikaya, A., Ausloos, P. J., Clifton, C., Lias, S. G., Zaikin, V., et al. (1998). NIST/EPA/NIH Mass Spectral Library—NIST 98 Version National Institute of Standards and Technology.

Whalen-Pedersen, E. K., & Jurs, P. C. (1981). Calculation of linear temperature programmed capillary gas chromatographic retention indices of polycyclic aromatic compounds. *Analytical Chemistry*, *53*, 2184–2187.