Educación Química (2015) **26**, 195–201

educación Química

www.educacionquimica.info

TEACHERS IN THE KNOW

The use of the *k* nearest neighbor method to classify the representative elements

João Elias Vidueira Ferreira*, **Clauber Henrique Souza da Costa, Ricardo Morais de Miranda, Antonio Florencio de Figueiredo**

Federal Institute of Education, Science and Technology of Pará, Pará State, Amazon, Brazil

Received 4 November 2014; accepted 27 December 2014
Available online 10 June 2015

KEYWORDS
Chemical periodicity;
Representative elements;
k-Nearest neighbor method;
Chemometrics

Abstract The use of Statistics in Chemistry has grown significantly with advances in computation. Nowadays it is easier to deal with a large data set and extract relevant chemical information. This paper describes the use of *k* nearest neighbor method to classify the representative elements as metal or nonmetal according to their periodic properties: atomic radius, ionization energy, electron affinity and electronegativity. The method requires a very simple mathematical background and can be easily performed and understood. The algorithm classifies an object into a distinct class when there are two or more groups of objects of known class and takes into account the distances among the objects. The pedagogical objective is to present an interdisciplinary activity in which Statistics can be used to make comparisons in Chemistry.

PALABRAS CLAVE
Propiedades periódicas;
Elementos representativos;
Método *k*-ésimo vecino más cercano;
Quimiometría

El uso del método del k-ésimo vecino más cercano en la clasificación de los elementos representativos en metales y no metales

Resumen El uso de la estadística en química ha aumentado mucho con el desarrollo de la computación. Hoy es más fácil trabajar con los datos y extraer importante información química. Este artículo describe el uso del método del k-ésimo vecino más cercano en la clasificación de los elementos representativos conforme sus propiedades periódicas: radio atómico, energía de ionización, afinidad electrónica y electronegatividad. El objetivo pedagógico es presentar

* Corresponding author.
E-mail address: joao.elias@yahoo.com.br (J.E. Vidueira Ferreira).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

http://dx.doi.org/10.1016/j.eq.2015.05.004

una actividad interdisciplinaria en la que la estadística se puede utilizar para hacer comparaciones en química.

Introduction

The increase in general availability of computing power represents the fastest and longest sustained technological advance in human history (Bedolla & Bermúdez, 2009). The maturing of hardware and software increased significantly our ability to statistically analyze data (Schlotter, 2013). Consequently advances in computation gave new horizons to chemistry research and education (Ferreira, Da Costa, De Miranda, Figueiredo, & Ginarte, 2014) because now interpretation of a great number of atomic/molecular properties is possible, helping chemists investigate simple and complex systems (Ferreira et al., 2009). However when you have a large set of data, it is not an easy task to analyze so many variables and extract useful information from them. Besides it seems that this process requires infinite patience (Ferreira, Figueiredo, Barbosa, & Pinheiro, 2012). Then what to do?

Whenever you have to find patterns by analyzing a huge data matrix, it is certainly difficult to make comparisons by visualizing the raw data. So application of a pattern recognition method is desirable. A very simple approach that can be easily performed and understood is the *k* nearest neighbor (*k*-NN) method. The algorithm in this method classifies an object into a distinct class when there are two or more groups of objects of known class. It is based on the concept of proximity and makes no assumption about the distribution in the classes (Miller & Miller, 2005). This similarity technique assumes that the closer objects lie in measurement space, the more likely they belong to the same category or they are similar with respect to the variables under study. In general, when the number of measurements is two or three, the classification of objects can be done by simple graphical approaches in two or three dimensions, respectively. But in the cases where there are more than three variables, as is usual in chemistry, it is necessary to extend the concept of distance to one in multidimensional space, each axis representing a variable. Although we cannot visualize more than three dimensions, computers can handle geometry in an indefinite number of dimensions, and the idea of distance is easy to generalize (Brereton, 2007).

In this work the *k* nearest neighbor (*k*-NN) method is used to classify the representative elements as metal or nonmetal according to their periodic properties. The properties considered are atomic radius (size), first ionization energy (the energy required to remove an electron from a gaseous atom), electron affinity (the energy change involved in adding an electron to a gaseous atom) and electronegativity (a measure of the tendency of an atom to attract electron in a chemical bond). One special aspect about the periodic properties is that generally they vary regularly as we move across a period or up to down in a family. The knowledge of periodicity helps understand chemical and physical properties of the elements and their compounds. The importance of the periodic table of the chemical elements is that it is the principal organisational feature of chemistry (Glasser, 2011).

The representative elements, which are also called main-group elements, comprise both metals and nonmetals. They are found, in the long-form periodic table, in groups numbered 1, 2 and 13 through 18 or, as an older classification, groups numbered 1 through 8 with each number followed by a letter A. The disposition of these columns is into two blocks of columns separated by the transition metals. Groups 1 and 2 are on the left side of the table and are called the *s*-block elements while groups 13 through 18 are on the right side and are called *p*-block elements. One common aspect about the representative elements is that every element in the group has the same valence electron configuration and shows distinct and fairly regular variations in their properties with changes in atomic number. For the transition elements, the variations are not so regular because electrons are being added to an inner shell. In this discussion the noble gases were not included because they are generally not listed in electron affinity/electronegativity tables. They have no affinity for electrons since they have eight electrons in their outermost shells (except for He, which has two), as a consequence any additional electron must be added to the next higher electron shell.

Metallic elements are typically solid (mercury is the only liquid), lustrous, malleable, ductile, conduct electricity and heat well and tend to lose electrons in chemical reactions. On the contrary, nonmetals may be solid, liquid (only bromine) or gaseous and generally are characterized as presenting low electrical conductance. We must remember that the periodic table of the elements does not exist in only a single form and different formats are possible as it is well discussed by Rich and Laing (2011). For example, the frontier between these two classes in the periodic table is not a definite matter. One of the common forms considers a division indicated by a ''staircase'' line that runs from the top of group 13 to the bottom of group 16. The metals lie to the left of this division (except hydrogen) whereas the nonmetals lie to the right.

The main pedagogical objective of this paper is to show how statistical method such as *k*-NN can be used in data analysis in Chemistry, particularly in making comparisons, as it is described in this example with the representative elements. So this work is a guide to the use of *k*-NN classification technique intended to teachers and advanced college students. Authors believe that chemometrics, the use of statistics in chemistry, is important to anyone who deals with

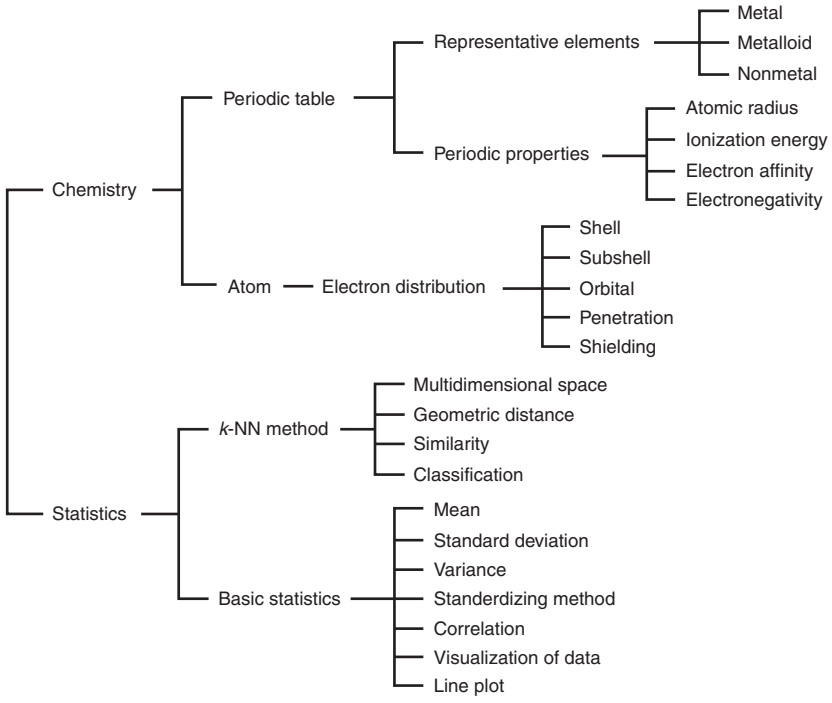

Figure 1 Topics involved in this activity.

chemical information. Chemometric methods enjoy an ever-increasing popularity and there is a need to introduce more graduate students to these research tools (Öberg, 2006). The topics involved that emerges from Chemistry and Statistics can be visualized in Fig. 1.

Methodology

The start point in this statistical analysis is the construction of the data matrix 38×4 (Table 1): 38 lines for chemical elements and 4 columns for periodic properties. Sometimes auto scaling data as preprocessing is necessary before running *k*-NN algorithm. Auto scaling a value is just subtracting mean followed by division by standard deviation. The results are scaled variables with zero mean and unit variance. This is required in order that the values have the same importance regarding the scale (Ferreira, 2002). Afterwards the chemical elements receive *a priori* their true classification (in this case, metal or nonmetal).

In the next step the geometric distance of an element to all other elements of the training set is calculated. Usually Euclidean distance is computed (Massart, Vandeginste, Deming, Michotte, & Kaufman, 2003). Euclidean distance is a common numerical measure of similarity used in multivariate analysis that is calculated through Eq. (1), where d is the distance between two points (chemical elements) in n dimensional space with coordinates (x_1, x_2, ..., x_n) and (y_1, y_2, ..., y_n). The distances are ranked in order. The class of an object is predicted according to the class of its k nearest neighbors, that is the origin of the name of the method. k assumes values ranging from 1 to a maximum k value (k_{max}), one less than the total number of objects in the training set, but as the size of k approaches the number of objects in the training set the comparisons are in fact made to far neighbors. In practice, one could try several values of k and use the one with the best error rate (Rencher, 2002).

$$d = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2]^{1/2} \quad (1)$$

The process of classification is quite similar to polling because every k closest object gives one vote for its class. Then the object is assigned to the class with the most votes. This method is self-validating since each object in the training set is compared to the others but not with itself. Fig. 2 shows an example of the process of classification of an unknown object, identified by a plus sign. If $k = 1$, only the closest neighbor and its class are taken into account and the unknown object is assigned to this class. In this case the shortest distance (d_1) is associated to an object of the square class, consequently the unknown object also belongs to this category. But if $k = 3$, the method considers the three closest objects. Thus the unknown object is classified as belong-

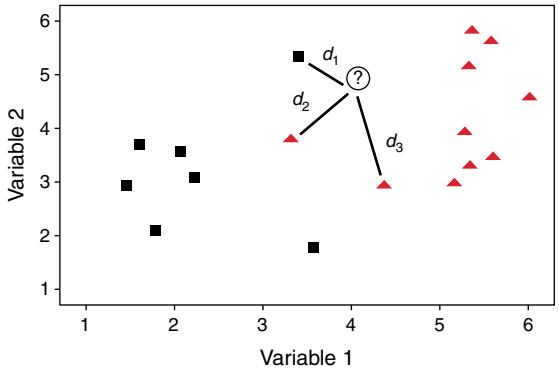

Figure 2 *k*-NN process of classification of unknown object. An example for $k = 3$, two classes and two variables.

Table 1 Elements and the four periodic properties[a] selected to run *k*-NN method.

Element	Atomic radius (pm)	Ionization energy (kJ mol^{-1})	Electron affinity[b] (kJ mol^{-1})	Electronegativity[c]	Group
Metal					
Li	152	519	60	1.00	1
Na	154	494	53	0.93	1
K	227	418	48	0.82	1
Rb	248	402	47	0.82	1
Cs	265	376	46	0.79	1
Fr	270	400	44	0.70	1
Be	113	900	−66[d]	1.60	2
Mg	160	736	−67[d]	1.30	2
Ca	197	590	2	1.30	2
Sr	215	548	5	0.95	2
Ba	217	502	14	0.89	2
Ra[e]	283	509	10	0.90	2
Al	143	577	43	1.60	13
Ga	122	577	29	1.60	13
In	163	556	29	1.80	13
Tl	170	590	19	2.00	13
Ge[f]	122	784	116	2.00	14
Sn	141	707	116	2.00	14
Pb	175	716	35	2.30	14
Sb[f]	141	834	103	2.10	15
Bi	155	703	91	2.00	15
Po[f]	167	812	174	2.00	16
Nonmetal					
H	30	1310	73	2.20	–
B[f]	88	799	27	2.00	13
C	77	1090	122	2.60	14
Si[f]	117	786	134	1.90	14
N	75	1400	−7	3.00	15
P	110	1011	72	2.20	15
As[f]	121	947	78	2.20	15
O	66	1310	141	3.40	16
S	104	1000	200	2.60	16
Se	117	941	195	2.60	16
Te[f]	137	870	190	2.10	16
F	58	1680	328	4.00	17
Cl	99	1255	349	3.20	17
Br	114	1140	325	3.00	17
I	133	1008	295	2.70	17
At	140[g]	1037	270	2.00	17

[a] Atkins and Jones (2009).
[b] The convention adopted in this work is the same in ref. a: positive electron affinity corresponds to an exothermic process while negative electron affinity corresponds to an endothermic process.
[c] Pauling scale.
[d] Lee (1996).
[e] Periodic Table (2014): http://www.rsc.org/periodic-table/element/88/radium.
[f] Metalloid.
[g] Zumdahl and Zumdahl (2007).

ing to the triangle class, because it receives two votes (d_2 and d_3) from the triangle class and only one vote (d_1) from the square class. In the specific case that there is a tie, the classification is based on the summed accumulated distances and the class with smallest accumulated distances is attributed to the object. This technique also can be used when the groups are not separated by a plane as occurs in Fig. 2. Perfect agreement between true and predicted class suggests a distinct separation of categories with respect to the variables under consideration. On the contrary, there may be overlap among classes or the presence of outliers or even objects of different categories close in distance.

Before *k*-NN algorithm is run, the elements had their true category (metal or nonmetal) assigned in the matrix

Table 2 Classification of the representative elements according to *k*-NN method. C for correct and I for incorrect classification.

Element	*k*1	*k*2	*k*3	*k*4	*k*5	*k*6	*k*7	*k*8	*k*9	*k*10
					Metal					
Li	C	C	C	C	C	C	C	C	C	C
Na	C	C	C	C	C	C	C	C	C	C
K	C	C	C	C	C	C	C	C	C	C
Rb	C	C	C	C	C	C	C	C	C	C
Cs	C	C	C	C	C	C	C	C	C	C
Fr	C	C	C	C	C	C	C	C	C	C
Be	C	C	C	C	C	C	C	C	C	C
Mg	C	C	C	C	C	C	C	C	C	C
Ca	C	C	C	C	C	C	C	C	C	C
Sr	C	C	C	C	C	C	C	C	C	C
Ba	C	C	C	C	C	C	C	C	C	C
Ra	C	C	C	C	C	C	C	C	C	C
Al	C	C	C	C	C	C	C	C	C	C
Ga	C	C	C	C	C	C	C	C	C	C
In	C	C	C	C	C	C	C	C	C	C
Tl	C	C	C	C	C	C	C	C	C	C
Ge[a]	I	I	C	C	C	C	I	C	I	C
Sn	C	C	C	C	C	C	C	C	C	C
Pb	C	C	C	C	C	C	C	C	C	C
Sb[a]	C	C	C	C	C	C	C	C	C	C
Bi	C	C	C	C	C	C	C	C	C	C
Po[a]	I	I	C	C	C	C	C	C	I	C
					Nonmetal					
H	C	C	C	C	C	C	C	C	C	C
B[a]	C	C	C	C	C	C	I	I	I	I
C	C	C	C	C	C	C	C	C	C	C
Si[a]	I	I	I	I	I	I	I	I	I	I
N	C	C	C	C	C	C	C	C	C	C
P	C	C	I	C	C	C	C	C	C	C
As[a]	C	C	I	C	C	C	I	C	C	C
O	C	C	C	C	C	C	C	C	C	C
S	C	C	C	C	C	C	C	C	C	C
Se	C	C	C	C	C	C	C	C	C	C
Te[a]	I	I	C	I	I	I	I	I	I	I
F	C	C	C	C	C	C	C	C	C	C
Cl	C	C	C	C	C	C	C	C	C	C
Br	C	C	C	C	C	C	C	C	C	C
I	C	C	C	C	C	C	C	C	C	C
At	C	C	C	C	C	C	C	C	C	C

[a] Metalloid.

in a column. The model built used auto scaled data and ten as k_{max}. This activity can be easily reproduced with any software that performs *k*-NN analysis. Authors employed Pirouette package, version 4.5 (2011).

Results

The main purpose of this work is to employ the *k*-NN method to classify the representative elements as metal or nonmetal taking as variables periodic properties as stated before. So the results can be summarized in Table 2, which lists the prediction for the elements according to a certain *k* value. Column number corresponds to *k* setting so that the first column of this matrix holds the class for each element when only one neighbor (the nearest) is polled whereas the last column holds the class for the elements when 10 neighbors are polled.

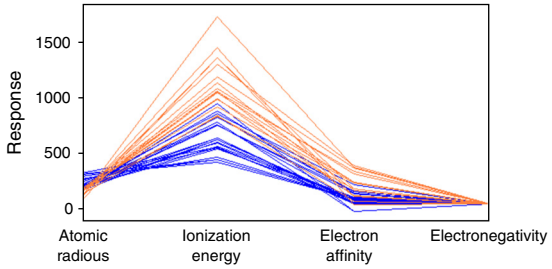

Figure 3 Line plot for the periodic properties. Overlap of lines for all periodic properties occurs and suggests no completely distinguishable range between metal and nonmetal.

The majority of elements received a correct classification for all 10-nearest neighbors, which is indicated by the letter C. However seven elements received a wrong classification, indicated by the letter I. Five nonmetals were classified as metals whereas two metals were classified as nonmetals for at least one *k* value. Silicon was misclassified for all ten values of *k*, tellurium, nine; boron and germanium, four; polonium, three; arsenic, two; and phosphorus only one. These elements are metalloids, except phosphorus. Table 3 presents the number of elements wrongly classified for each *k* value and the rate of error.

As stated before, the separation between elements typically metallic from those nonmetallic as traditionally drawn by a simple line in the periodic table is something relative to the property under consideration. When we consider the periodic properties the distinction is not so rigid. Actually a perfect classification would only be possible if the elements in both classes had distinct range for these properties, that is, there should be no overlap in multidimensional space, but this is not the case here.

Fig. 3 is a line plot of the data in Table 1, which exhibits response values for the periodic properties. This plot shows that metals (blue lines) and nonmetals (red lines) are categories not perfectly distinguishable considering these variables individually because it is apparent that there are overlaps of values as we analyze minimum and maximum values for each property. The range for the properties can be better visualized in Table 4. The elements listed in the table are those that received a wrong classification and the values in the parenthesis are for the respective periodic property. Overlap for silicon is present in all four periodic properties. By the way this element was given an incorrect class when all 10-nearest neighbors were analyzed. Tellurium, boron germanium, polonium and arsenic showed overlap for three properties and phosphorus only two. This last element received only incorrect prediction for $k = 3$.

Variations in periodic properties are mainly a result of *n* principal quantum number and shielding and penetration effects. Shielding is the reduction of the true nuclear charge felt by an electron resulting from repulsion forces on this electron that are caused by the presence of other electrons. Penetration is the presence of an electron closer to the nuclear charge. Consequently this electron feels a stronger attraction by the nucleus. The typical trends in the periodic

Table 3 Number of misclassifications and error rate for each *k* value.

Class	Number of elements	Number of elements misclassified									
		*k*1	*k*2	*k*3	*k*4	*k*5	*k*6	*k*7	*k*8	*k*9	*k*10
Metal	22	2	2	0	0	0	0	1	0	2	0
Nonmetal	16	2	2	3	2	2	2	4	3	3	3
% Error rate		10.5	10.5	7.9	5.3	5.3	5.3	13.2	7.9	13.2	7.9

Table 4 Ranges (minimum and maximum) for the periodic properties and the misclassified elements.

	Periodic property			
	Atomic radius (pm)	Ionization energy (kJ mol^{-1})	Electron affinity (kJ mol^{-1})	Electronegativity[a]
	Metals			
	Min: 113; Max: 283	Min: 376; Max: 900	Min: −67; Max: 174	Min: 0.70; Max: 2.30
Nonmetals misclassified as metals	– Si (117) – As (121) Te (137)	B (799) Si (786) – – Te (870)	B (27) Si (134) P (72) As (78) –	B (2.00) Si (1.90) P (2.20) As (2.20) Te (2.10)
	Periodic property			
	Atomic radius (pm)	Ionization energy (kJ mol^{-1})	Electron affinity (kJ mol^{-1})	Electronegativity[a]
	Nonmetals			
	Min: 30; Max: 140	Min: 786; Max: 1680	Min: −7; Max: 349	Min: 1.90; Max: 4.00
Metals misclassified as nonmetals	Ge (122) –	– Po (812)	Ge (116) Po (174)	Ge (2.00) Po (2.00)

[a] Pauling scale.

table are that atomic radius increases down a group and decrease across a period while ionization energy, electron affinity and electronegativity increase up a group and across a period.

However irregularities occur along the groups and across the periods and even within metals or nonmetals. For example the properties of elements in the second period usually differ significantly from those of other elements in their families, because second-period elements have no low-energy *d* orbitals. Going from groups 13 to 15, electrons are going singly into separate *p* orbitals, where they do not shield one another significantly. The general left-to-right increase in ionization energy for each period is interrupted by a dip between groups 2 and 13 because the 2*p* electrons are slightly higher in energy than the 2*s* electrons. The same dip is found between groups 15 and 16 but the explanation for this behavior is the fourth *p* electron in the group 16, which is paired with another electron in the same orbital, so it experiences greater repulsion than it would in an orbital by itself. This increased repulsion makes it easier to remove the *p* fourth electron in an outer shell for group 16 than the third *p* electron in an outer shell for group 15. The values of electron affinity usually become larger on moving across a period from left to right, but the trend is not smooth. The elements in groups 2 and 15 appear not to follow the general trend since the added electron would start a *p* subshell or would be paired with another electron in the *p* subshell, respectively (McMurry & Fay, 2003; Zumdahl & Zumdahl, 2007). As a numerical example involving irregularities we have the atomic radius for the nonmetal silicon (117 pm). This value is close to that for germanium (122 pm), a metal, but more distinct in size than that for carbon (77 pm), another nonmetal. The same irregularity occurs for electronegativity: silicon (1.90), germanium (2.00, metal) and carbon (2.60, nonmetal).

Besides the elements in groups 13 through 16 exhibit different classes within a column since the elements at the head of the group are nonmetals and those at the foot of the group are metals. For the elements in group 15 the change is so great that, with regard to monoatomic ions, elements at the top tend to form anions and those at the bottom tend to form cations. There are also allotropic variations in the sense that some elements exist as both metals and nonmetals. An example is group 15: nitrogen and phosphorus are nonmetals, but arsenic exists as nonmetal, metalloid, and metallic allotropes, and antimony and bismuth are metals. Elements in the *p* block typically form several allotropes (Atkins, Overton, Rourke, Weller, & Armstrong, 2010). All these aspects reinforce the presence of elements with intermediate characteristics between metals and nonmetals.

Conclusion

k-NN method of classification reveals that among all representative elements only seven elements received a wrong classification for at least one *k* value when 10-nearest neighbors are considered. Five nonmetals were classified as metals (boron, silicon, phosphorus, arsenic and tellurium) whereas two metals as nonmetals (germanium and polonium). Of course any classification of an element depends on the properties considered. In this work this classification took into account as properties the atomic radius, ionization energy, electron affinity and electronegativity. The reason for misclassifications lies in the irregularities in the periodic properties found along the groups and across the periods and even within metals or nonmetals. Consequently, the common division indicated by a ''staircase'' line that runs from the top of group 13 to the bottom of group 16 is really something relative to a certain point.

Conflict of interest

The authors declare no conflict of interest.

References

Atkins, P., & Jones, L. (2009). *Chemical principles: The quest for insight* (5th ed.). New York, USA: W. H. Freeman and Company.

Atkins, P. W., Overton, T. L., Rourke, J. P., Weller, M. T., & Armstrong, F. A. (2010). *Shriver and Atkins's inorganic chemistry* (5th ed.). New York, USA: W. H. Freeman and Company.

Bedolla, C. A., & Bermúdez, C. O. O. (2009). La química computacional en el salón de clase. *Educación Química*, *20*(2), 182–186.

Brereton, R. G. (2007). *Applied chemometrics for scientists*. West Sussex, England: John Wiley & Sons.

Ferreira, J. E. V., Da Costa, C. H., De Miranda, R. M., Figueiredo, A. F., & Ginarte, Y. M. A. (2014). An interdisciplinary study on monosubstituted benzene involving computation, statistics and chemistry. *Educación Química*, *25*(4), 424–430.

Ferreira, J. E. V., Figueiredo, A. F., Barbosa, J. P., & Pinheiro, J. C. (2012). Chemometric study on molecules with anticancer properties. In K. Varmuza (Ed.), *Chemometrics in practical applications* (pp. 185–200). Rijeka, Croatia: Intech.

Ferreira, J. E. V., Lobato, M., Figueiredo, A. F., Santos, M., Farias, M., Macedo, W., et al. (2009). Teaching computational chemistry by studying malaria. *Bulgarian Journal of Science Education*, *18*(6), 414–422.

Ferreira, M. (2002). Multivariate QSAR. *Journal of the Brazilian Chemical Society*, *13*(6), 742–753.

Glasser, L. (2011). Periodic tables on the world wide web. *Australian Journal of Education in Chemistry*, *71*, 3–4.

Lee, J. D. (1996). *Concise inorganic chemistry* (5th ed.). Oxford, England: Blackwell Publishing.

Massart, D. L., Vandeginste, B. M. G., Deming, S. N., Michotte, Y., & Kaufman, L. (2003). . *Chemometrics: A text book* (vol. 2) Amsterdam, The Netherlands: Elsevier.

McMurry, J., & Fay, R. C. (2003). *Chemistry* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Miller, J. N., & Miller, J. C. (2005). *Statistics and chemometrics for analytical chemistry* (5th ed.). Harlow, England: Pearson Prentice Hall.

Öberg, T. (2006). Introducing chemometrics to graduate students. *Journal of Chemical Education*, *83*(8), 1178–1181.

Periodic Table. (2014). *Royal society of chemistry*. Retrieved from http://www.rsc.org/periodic-table/element/88/radium

(2011). *Pirouette 4.5*. Woodinville, WA: Infometrix, Inc.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). New York, USA: John Wiley & Sons.

Rich, R. L., & Laing, M. (2011). Can the periodic table be improved? *Educación Química*, *22*(2), 162–165.

Schlotter, N. E. (2013). A statistics curriculum for undergraduate chemistry major. *Journal of Chemical Education*, *90*(1), 51–55.

Zumdahl, S. S., & Zumdahl, S. A. (2007). *Chemistry* (7th ed.). Boston, USA: Houghton Mifflin Company.