Int J Thermophys (2016) 37:98
DOI 10.1007/s10765-016-2097-2

ICPPP 18

Statistical Analysis of Photopyroelectric Signals using Histogram and Kernel Density Estimation for differentiation of Maize Seeds

J. E. Rojas-Lima[1] · A. Domínguez-Pacheco[1] · C. Hernández-Aguilar[1] · A. Cruz-Orea[2]

Received: 10 October 2015 / Accepted: 4 July 2016 / Published online: 1 August 2016

Abstract Considering the necessity of photothermal alternative approaches for characterizing nonhomogeneous materials like maize seeds, the objective of this research work was to analyze statistically the amplitude variations of photopyroelectric signals, by means of nonparametric techniques such as the histogram and the kernel density estimator, and the probability density function of the amplitude variations of two genotypes of maize seeds with different pigmentations and structural components: crystalline and floury. To determine if the probability density function had a known parametric form, the histogram was determined which did not present a known parametric form, so the kernel density estimator using the Gaussian kernel, with an efficiency of 95 % in density estimation, was used to obtain the probability density function. The results obtained indicated that maize seeds could be differentiated in terms of the statistical values for floury and crystalline seeds such as the mean (93.11, 159.21), variance (1.64×10^3, 1.48×10^3), and standard deviation (40.54, 38.47) obtained from the amplitude variations of photopyroelectric signals in the case of the histogram approach. For the case of the kernel density estimator, seeds can be differentiated in terms of kernel bandwidth or smoothing constant h of 9.85 and 6.09 for floury and crystalline seeds, respectively.

This article is part of the selected papers presented at the 18th International Conference on Photoacoustic and Photothermal Phenomena.

✉ A. Domínguez-Pacheco
fartur@hotmail.com

J. E. Rojas-Lima
errojas@ipn.mx

[1] Instituto Politécnico Nacional, SEPI-ESIME, Zacatenco, Unidad Profesional “Adolfo López Mateos”, Col. Lindavista, C.P. 07738 Ciudad de Mexico, Mexico

[2] Departamento de Física, CINVESTAV-IPN, A. P. 14-740, C.P. 07360 Ciudad de Mexico, Mexico

Keywords Kernel density estimator · Maize seeds characterization · Photopyroelectric microscopy · Photopyroelectric signals · Photothermal techniques

1 Introduction

The application of photothermal (PT) techniques has been extended to different areas of science to study thermal and optical properties of a wide variety of materials including semiconductors, graphite, plant leaves, evaluation of foodstuffs, as well as the characterization of seeds [1–5]. PT techniques include among others photoacoustic spectroscopy (PAS) [6] and photothermal microscopy (PTM) which in turn includes photoacoustic microscopy (PAM) and photopyroelectric microscopy (PPEM) [7]. On the other hand, quality evaluation and control has become increasingly important in biology and agriculture [8]. In particular, the application of PTM techniques as alternative methods for the characterization of nonhomogeneous materials like maize grains, which also present great complexity in their structure, could be relevant in the food industry [9]. One of the main reasons is due to the fact that PTM has been applied with great advantage in biological studies to characterize materials with complex structures, since this technique allows obtaining optical and thermal images, through a nondestructive and noninvasive evaluation [7]. The PAM is an imaging technique that acoustically detects optical and thermal contrast via the photoacoustic effect. PAM can provide high-resolution images at desired maximum imaging depths up to a few millimeters. Most importantly, PAM can simultaneously image anatomical, functional, molecular, flow dynamic, and metabolic contrasts in vivo [10]. On other hand, the PPEM has been extensively applied to the study of the thermal properties of condensed matter samples. The major advantages of these techniques are their simplicity, high sensitivity, nondestructive character, and adaptation on experimental restrictions for theoretical requirements [1]. It has been shown that PAS in combination with mathematical analysis which consists of the first derivative and the mobile standard deviation approaches, when applied to the optical absorption coefficient obtained from the amplitude of the PA signal, allows distinguishing the maximum optical absorption peaks of maize grains with different pigmentations better [9]. As far as the time expended by these methods is concerned, the first derivative outperforms the mobile standard deviation when defining the maximum optical absorption peaks in the absorption spectrum. Also thermal images of coffee-seed germ were obtained with a difference of 12 h between them using the PAM technique. The differences observed between images provide information that reflects degradation due to the germ cells undergo changes as a function of time. Also, the differences observed in the thermal and optical properties of the embryo are due to the differences in moisture content and optical absorption coefficient [11]. Another application of PAS technique is the Fourier transform infrared photoacoustic spectroscopy (FT-IR-PAS) which has been applied to the problem of identifying pollen samples [12]. Construction of a spectral library allowed pollens in the set to be correctly identified as the first match on the basis of photoacoustic spectra at the species (99 %) or family (100 %) level. Even though a lot of research and development work has been done on food safety and quality, more needs to be done to find economic ways of monitoring food safety and quality [13].

It is well known that the use of statistical methods for analyzing recollected data from experimentation, with the purpose of decision making and scientific discovery when the available information is both limited and variable, has become a common practice in many scientific areas [14,15]. A fundamental concept in statistics is the probability density function (PDF) [16], due to fact that it completely specifies variables of random behavior, indicating how probabilities are distributed over the values that the random variable adopts [17]. In order to estimate the PDF from observed data for modeling the underlying probabilistic structure of the data [18], parametric and nonparametric approaches can be applied [19]. In the parametric approach, it is considered that collected data come from one of a known parametric family of distributions. The PDF underlying the data could then be estimated by finding estimates of the parameters from the data (i.e., the mean and variance) and substituting these estimates of parameters into the formula for the PDF [19]. This approach has the advantage that it is easy to apply and it yields stable estimates. The main disadvantage of the parametric approach is lack of flexibility [20]. To avoid restrictive assumptions regarding the form of PDF and to estimate it directly from the data, nonparametric approaches such as the histogram and the kernel estimators have also been proposed [19,21]. The nonparametric models allow great flexibility in the possible form of the PDF [18].

Considering the necessity of alternative approaches for characterizing nonhomogeneous materials as maize seeds, the objective of this work was to analyze statistically the random amplitude variations of PPE signals obtained from PPEM experiment [22], in order to characterize statistically two genotypes of maize seeds with different pigmentations and structural components. The genotype A had bluish coloration and floury structure and genotype B had yellowish coloration with crystalline structure. The PDF of the random amplitude variations of the PPE signals caused by the differences of the seed structural components was obtained using both the histogram and the kernel density estimator with the purpose of identifying the distribution of the collected data and characterize maize seeds in terms of the statistical averages such as the mean, variance, standard deviation, as well as the optimal value of the kernel bandwidth estimated from the time series of the collected data. To our knowledge, the statistical analysis presented in this work has not been reported previously in the literature.

2 Materials and Methods

2.1 Biological Materials

In the present research, the PPE signals obtained from PPEM experiment of two maize seed genotypes with different pigmentations and structural components have been analyzed statistically. The genotype A had bluish coloration and floury structure and genotype B had yellowish coloration with crystalline structure.

2.2 PPEM Experimental Setup

Figure 1 shows a schematic diagram of the experimental setup used for PPEM technique. In this setup, the PPE sensor was used in the inverse configuration, which is

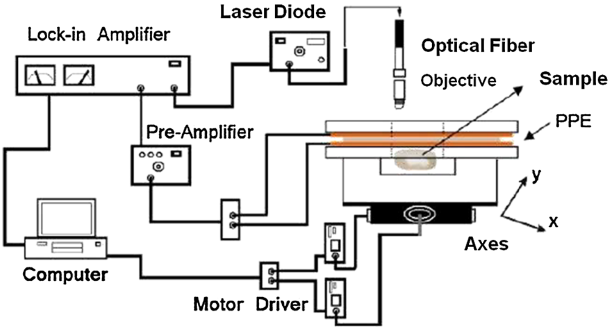

Fig. 1 Schematic diagram of the experimental setup used for PPEM technique

mounted on an x–y motorized stage, with spatial resolution of 70 µm and the scanned area was of 2.5mm × 2.5 mm. The excitation source is a fiber-coupled laser diode, at 650 nm wavelength and power of 100 mW, modulated at 1Hz of frequency by the reference oscillator of the lock-in amplifier. By using a microscope objective, the laser beam with a diameter of 40 µm was focused and periodically modulated on the pyroelectric sensor and below this the seed sample is attached to a sensor using thermal grease to obtain a better contact. This process generates a periodic heating that is transferred towards the sample. The PPE signal was preamplified and sent to the lock-in amplifier. A personal computer was used to control the scanning of the x–y stage and also to record, from the lock-in amplifier, the experimental PPE signal amplitude from each point of the scanned samples. The signal was recorded as a function of the position of the incident beam over the sample to obtain through the PPE data the thermal images of seeds, and additionally, the PPE recorded data (2809) were used to obtain the time series of the amplitude variations for each seed.

2.3 Nonparametric Approaches for Density Estimation

The histogram is the oldest and most widely used density estimator that is mainly used to identify the distribution of the data and decide whether a parametric or nonparametric approach shall be applied. The histogram has the main advantage of simplicity, but it also has drawbacks, such as lack of continuity and less accuracy in comparison to other nonparametric estimators [20]. To construct a histogram, assume that a sample of n real observations $X_1, X_2, \ldots, X_\mathrm{n}$, is given and whose underlying probability density function $f_x(x)$ is to be estimated. Select a left bound or starting point x_o and the bin width Δx. The bins of the histogram are defined as the intervals $[x_\mathrm{o} + (i-1)\Delta x, x_\mathrm{o} + i\Delta x), i = 1, 2, \ldots, m$. The histogram is defined as

$$\hat{f}_x(x) = \frac{Number\ of\ X_i\ as\ the\ same\ bin\ as\ x}{n\Delta x}. \tag{1}$$

Currently, one of the nonparametric approaches used for density estimation is the kernel density [18–21,23,24]. A kernel density is a statistical method used to estimate a continuous and smooth distribution from a finite set of observed points [25]. This method has been applied in a broad variety of applications including pattern recognition, computer vision, archeology, climatology, genetics, hydrology, economics [6,24,26], among others.

The idea underlying the kernel density estimate is that each data point $X_1, X_2, \ldots,$ X_n of a sample of size n from a random variable with a unknown probability density function $f_x(x)$ is replaced by a specified function $K(\cdot)$ called the kernel, centered at each data point X_i and with a scaling parameter h called bandwidth or smoothing constant. The kernel functions are added together and the resulting function is scaled to have a unit area, which is a smooth curve called the density estimate of $f_x(x)$ at the point x, given by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{2}$$

The most widely used kernel $K(z)$ is the Gaussian which is given by [21,26]

$$K(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right). \tag{3}$$

The efficient use of this method requires the optimal selection of the smoothing constant or the bandwidth of kernel [26]. It determines the amount of smoothing applied in estimating the PDF. The rule that is commonly used in practice to choose the optimal value of h is [21]

$$\hat{h}_{opt} = \frac{0.9\hat{\sigma}}{n^{1/5}}. \tag{4}$$

Usually $\hat{\sigma} = \min(\mathrm{s}, \mathrm{IQR}/1.34)$ is used, where s is the standard deviation of the sample and IQR is the interquartile range of the data. The value of h is called "optimal" in the sense that it minimizes the Asymptotic Mean Integrated Squared Error (AMISE) in density estimation, and for the case of the Gaussian kernel the efficiency is estimated to be 95.12 % [20].

3 Results

Figure 2 shows optical and thermal images obtained from PPEM for the two genotypes of maize seeds. According to the thermal images obtained from PPEM, it is possible to observe the differences in the structural components of nonhomogenous materials as maize seeds by using a color scale. Each color represents the different values of the PPE signal amplitude as a function of the scanned x–y coordinates. The differences of the PPE signal are caused by the inhomogeneous nature of the thermal properties of the structural components (i.e., endosperm, pericarp, germ, etc.). Although it is possible to observe differences in thermal images that allow identifying two genotypes of maize seeds that can be differentiated in terms of the thermal properties, it is also important to obtain a quantitative description from recollected data. Figure 3 shows the plots of the

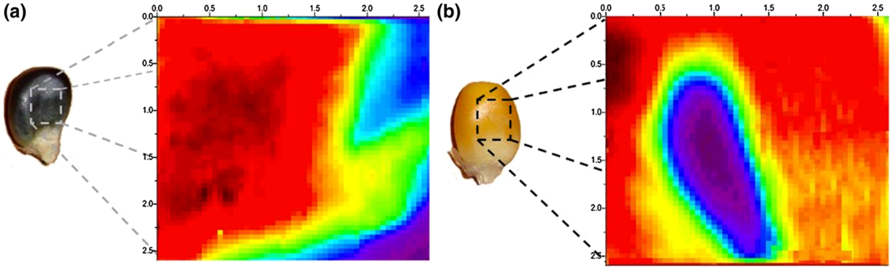

Fig. 2 Optical and thermal images obtained from PPEM (a) Genotype A: *bluish coloration* and floury structure (b) Genotype B: *yellowish coloration* with crystalline structure (Color figure online)

time series of the PPE signals (expressed in mV taken every 5 s) for the two genotypes of maize seeds. For the case of floury structure maize seed (Fig. 2a), the range of the amplitude variations is lower compared to the range of amplitude variations of crystalline structure maize seed (Fig. 2b). In the case of the bluish coloration seed, the pattern of behavior of the PPE signal tends to diminish with the time. In the case of the yellowish coloration, the pattern behavior of the PPE signal tends to diminish or to increase depending of the time interval. So, it is important to characterize them in statistical terms. The histograms with a normalized area, total area of 1, are shown in Fig. 4. It can be observed that the PDF obtained by histogram did not exhibit a known parametric form; however, it is possible to establish a difference between the ranges of amplitude variations of both signals.

According to the histogram shown in Fig. 4a, the amplitude variations of the PPE signal for the maize seed with floury structure (*i.e.*, bluish coloration) ranges from 1.92 mV to 158.77 mV, and for the case of maize seed with crystalline structure (*i.e.*, yellowish coloration) the amplitude variations of the PPE signal ranges from 73.5 mV to 232.93 mV (Fig. 4b). It indicates that the PPE signal obtained from a maize seed with crystalline structure is more intense when is compared with the PPE signal obtained from a maize seed of floury structure. So to obtain a more complete statistical characterization of the PPE signals the statistical averages such as the mean, variance, and standard deviation were obtained and are shown in Table 1. From the values shown in Table 1 and in accordance with the aforementioned, it can be noticed that the mean value of the signal amplitude of the bluish coloration seed is lower than the mean value obtained for the case of the signal from the yellowish coloration seed; this is due to the fact that the genotype B has a crystalline structure and according to the literature the values of thermal diffusivity and conductivity are higher in its pericarp compared with floury seeds [27]. Moreover, as it was described in the experimental setup section, the PPE sensor was used in the inverse configuration indicating that the obtained PPE signals depend only of the thermal properties of the sample [22]. Additionally, the detected signal current is proportional to the rate of temperature change [23].

Regarding the standard deviation, a lower value was obtained for the crystalline structure seed indicating thus a lesser variability in the PPE signal amplitude because the molecules of the crystalline seed variety are less amorphous this means that the diffusion of heat is better [27,28]. On the other hand, the floury seeds diffuse the heat into the seed with more difficulty due to their less structural molecular organization

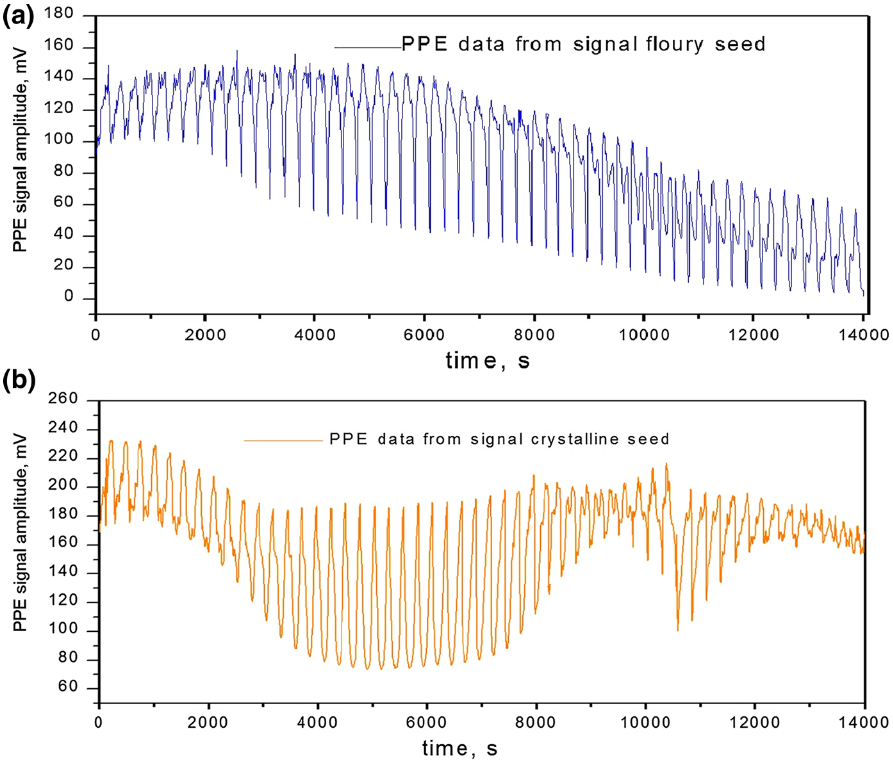

Fig. 3 Time series of PPE signals for (a) genotype A: *bluish coloration* and (b) genotype B: *yellowish coloration* (Color figure online)

and, in this way, produces more variability in the signal level, so that the increase in the temperature is higher at the surface. From the standard deviation values shown in Table 1, it can be determined that the amplitude levels tend to be more concentrated in the range of 52.575 mV to 133.665 mV for the case of floury seed, and for the crystalline seed case the amplitude values are more concentrated at the range of 120.736 mV to 197.685 mV.

As aforementioned, the histogram method has the advantage of being a simple method; however, it also has disadvantages such as lack of continuity and less accuracy in comparison to other nonparametric estimators [20], so that the kernel density estimation method based on the Gaussian kernel [21,26] was applied to obtain the PDF of the amplitude variations of the PPE signals for the two genotypes of maize seeds.

By using the kernel density estimation approach, the PDF of the amplitude variations of the PPE signal in terms of the Gaussian kernel for the bluish coloration maize seed can be described as

$$\hat{f}_{h-b}(x) = \frac{1}{nh_b}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_b}\right) = \frac{1}{nh_b\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{(x-X_i)^2}{2h_b}}, \tag{5}$$

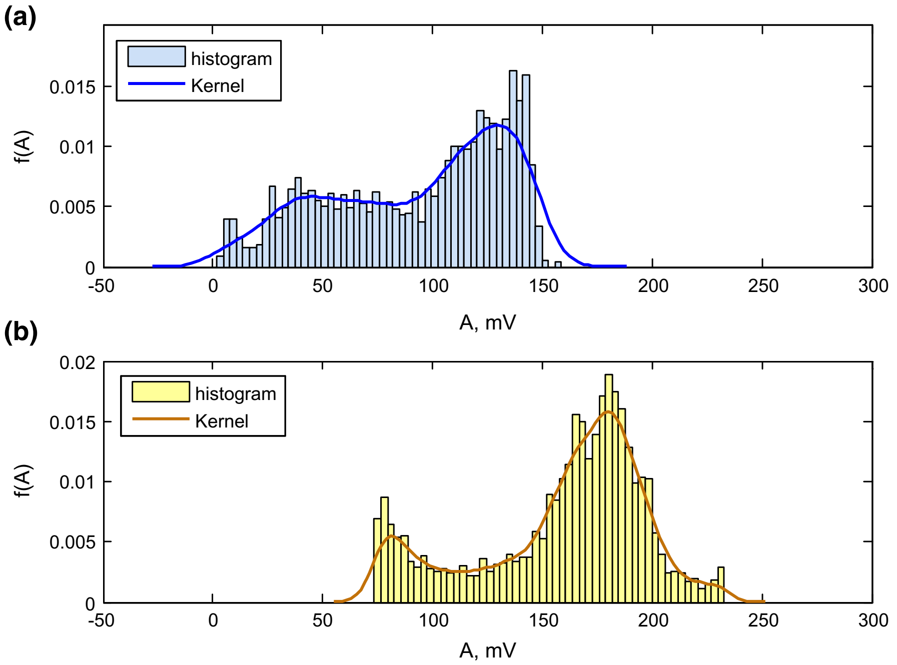

Fig. 4 Normalized histograms of the amplitude variations of the PPE signals and their comparison with the estimated PDF using the Gaussian kernel for two genotypes of maize seeds (a) genotype A: *bluish coloration* and (b) genotype B: *yellowish coloration* (Color figure online)

where $n = 2809$ and $h_b = 9.8582$ for the bluish coloration and floury structure seed.

In a similar way, the PDF of the amplitude variations of the PPE signal using the Gaussian kernel for the yellowish coloration maize seed is given by

$$\hat{f}_{h-y}(x) = \frac{1}{nh_y}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_y}\right) = \frac{1}{nh_y\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{(x-X_i)^2}{2h_y}}, \tag{6}$$

where $n = 2809$ and $h_y = 6.0943$ for the yellowish coloration and crystalline structure seed.

The estimated PDF using the Gaussian kernel for genotype A ($i.e.$, floury) and genotype B ($i.e.$, crystalline) as well as its comparison with the histograms are shown in Fig. 4.

From Fig. 4, it can be observed the smooth and continuous shape of the PDF estimated by means of the Gaussian kernel. According to the PDF estimated, the amplitude values for the genotype A maize seed (Fig. 4a) range from −27.658 mV to 183.346 mV, and for the genotype B (Fig. 4b) the amplitude values range from 55.253 mV to 251.212 mV, confirming in this way that the crystalline seed produces higher PPE signal amplitude values regarding to the PPE signal amplitudes obtained from a floury maize seed. It is noteworthy that from the determined smoothing factor or bandwidth of the kernel, it is also possible to identify differences between maize

Table 1 Statistical values obtained by the histogram and kernel density estimation approaches from PPE signals

Type of maize seed	Mean	Variance	Standard deviation	Smoothing constant
Genotype A: Floury structure bluish coloration	93.118	1.644×10^3	40.547	9.8582
Genotype B: Crystalline structure Yellowish coloration	159.210	1.480×10^3	38.475	6.0943

seeds with different structures. For the crystalline structure maize seed, the required bandwidth of the kernel is lower compared to the floury structure, finding that the PPE signal exhibits less variability in the amplitude levels requiring thus less smoothing for determining the PDF, confirming thus the feasibility of the use of statistical approaches that could be applied for differentiate seeds with different structures from PPE signals obtained from them. The main advantage of the kernel density estimation is because it allows great flexibility to estimate a continuous and smooth distribution from a finite set of observed points. It is also important to point out that it is possible to obtain a quantitative description from recollected data.

4 Conclusions

From the statistical analysis presented in this research, it has been demonstrated the feasibility of the statistical methods for the identification of seeds using the time series of the PPE signals, obtained from experimentation of two genotypes of maize seeds with different structures and pigmentations, by means of nonparametric approaches for estimation of the PDF from the signal amplitude variations. In the case of the histogram approach, due to its lack of continuity and less accuracy in comparison with other density estimators, it is important to determine the statistical averages such as the mean, variance, and standard deviation in order to have a more precise identification of seeds because it was not possible to identify a known parametric form of the PDF. From these values it is possible to obtain a quantitative differentiation of seeds expressed in terms of the mean, variance, and standard deviation values of 93.11, 1.64×10^3, 40.54 and 159.20, 1.48×10^3, 38.47 for the case of the floury and crystalline seeds, respectively. By using the kernel density estimation approach, we have observed that maize seeds with different structures and pigmentations can be identified in terms of the optimal kernel bandwidth or smoothing constant, which determines the performance of the kernel. For the case of the studied seeds, the value of smoothing constant was 9.8582 and 6.0943 for the floury and crystalline seeds, respectively.

Acknowledgments The authors thank the Instituto Politécnico Nacional, through the CONACYT, COFAA, EDI, and SIP Scholarship Projects. One of the authors (A. Cruz-Orea) is grateful for the economic support of CONACYT through Project 241330. Also, we thank Ing. Esther Ayala at the Physics Department of CINVESTAV-IPN for her technical support.

References

1. K. Strzałkowski, Mater Sci. Eng. B **184**, 80 (2014)
2. A. Mami, I. Mellouki, N. Yacoubi, Int. J. Eng. Sci. Innov. Technol. **3**(3), 608 (2014)
3. B.R. Briseño-Tepepa, J.L. Jiménez-Peréz, R. Saavedra, R. González-Ballesteros, E. Suaste, A. Cruz-Orea, Int. J. Thermophys. **29**(6), 2200 (2008)
4. S. Luterotti, D. Bicanic, K. Kljak, D. Grbesa, E.S.M. Martínez, R. Spruijt, Food Biophys. **6**(1), 12 (2011)
5. C. Hernandez-Aguilar, A. Cruz-Orea, R. Ivanov, A. Dominguez, A. Carballo, I. Moreno, R. Rico, Food Biophys. **6**(4), 481 (2011)
6. A. Rosencwaig, A. Gersho, J. Appl. Phys. **47**(1), 64 (1976)
7. A. Mandelis, M.M. Zver, J. Appl. Phys. **57**(9), 4421 (1985)
8. M. Suzuki, K. Miyamoto, Hoshimiya, Jpn. J. Appl. Phys. **44**, 6S (2005)
9. R.R. Molina, C.H. Aguilar, A.D. Pacheco, A. Cruz-Orea, M.A. Canseco, Int. J. Thermophys. **34**(8), 1540 (2013)
10. J. Yao, L.V. Wang, Laser Photon Rev. **7**(5), 758 (2013)
11. A. Dominguez–Pacheco, C. Hernández –Aguilar, A. Cruz-Orea, E.I. Alemán, E. Martínez-Ortíz, Int. J. Thermophys. **34**(8), 1499 (2013)
12. G. Parodi, P. Dickerson, J. Cloud, J. Appl. Spectrosc. **67**(3), 342 (2013)
13. J.T. Alexander, V. Bochko, B. Martinkauppi, S. Saranwong, S. Mantere, Int. J. Spectrosc. (2013). doi:10.1155/2013/341402
14. R.E. Walpole, R.H. Myers, S.L. Myers, K. Ye, *Probability & Statistics for Engineers & Scientist* (Pearson Education Inc, Prentice Hall, 2012), pp. 1–4
15. R. Ott, M. Lognecker, *An Introduction to Statistical Methods and Data Analysis*, 6th edn. (International ed.Books/Cole, Belmont, 2010)
16. I. Horová, Kernel Density Estimation. Encyclopedia of Environmetrics, 3, (2013)
17. T.T. Soong, *Fundamentals of Probability and Statistics for Engineers* (Wiley, New York, 2004)
18. A.Z. Zambom, R. Dias, Int. Econom. Rev. (IER) **5**(1), 20–42 (2013)
19. B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, vol. 26 (CRC Press, Boca Raton, 1986)
20. W. Zucchini, *Applied Smoothing Techniques, Part 1: Kernel Density Estimation* (Temple University, Philadephia, 2003)
21. T. Ledl, Austrian J. Stat. **33**(3), 267 (2004)
22. A. Domínguez-Pacheco, C. Hernández-Aguilar, R. Zepeda-Bautista, E. Martínez-Ortiz, A. Cruz-Orea, Superficies y Vacío **25**(2), 92 (2012)
23. Z.I. Botev, J.F. Grotowski, D.P. Kroese, Ann. Stat. **38**(5), 2916–2957 (2010)
24. S.J. Sheather, Stat. Sci. **19**(4), 588 (2004)
25. Y. Zheng, J. Jestes, J. M. Phillips, F. Li, in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, (June, 2013), pp. 433–444
26. V.C. Raykar, R. Duraiswami, in *SDM*, 524 (2006). doi:10.1137/1.9781611972764.53
27. W.J. Da Silva, B.C. Vidal, M.E.Q. Martins, H. Vargas, A.C. Pereira, M. Zerbetto, L.C. Miranda, Nature **362**, 417 (1993)
28. K.Y. Sastry, L. Froyen, J. Vleugels, E.H. Bentefour, C. Glorieux, Int. J. Thermophys. **25**(5), 1611 (2004)