Revista Colombiana de Estadística
Current Topics in Statistical Graphics
Diciembre 2014, volumen 37, no. 2, pp. 399 a 417
DOI: http://dx.doi.org/10.15446/rce.v37n2spe.47945

Visualization of Skewed Data: A Tool in R

Visualización de datos sesgados: una herramienta en R

Raydonal Ospina[1,a], Antonio Marcos Larangeiras[2,b], Alejandro C. Frery[2,c]

[1]Departamento de Estatística, Universidade Federal de Pernambuco, Recife, Brazil

[2]Laboratório de Computação Científica e Análise Numérica, Universidade Federal de Alagoas, Maceió, Brazil

Abstract

After discussing the main characteristics of the histogram and of a number of variations in the boxplot, this work presents a visualization tool specifically tailored to deal with skewed data. The idea is to use various types of boxplots (the classical one, which is tuned for skewness of the data, the shifting boxplot, and the box-percentile plot), the violin plot, and the histogram with a nonparametric estimate of the density overlay. The plots are presented in such a way that they facilitate the extraction of additional information from each one. We show that a good deal of information can be extracted from the inspection of the output using example data from synthetic aperture radar images. We provide an implementation in `R` based on functions already available.

***Key words*:** Exploratory Data Analysis, Skewed Data, Boxplot, Violin Plot, Visualization.

Resumen

Después de discutir las principales características del histograma y de un número de variables en el boxplot, se presento una herramienta de visualisación específicamente diseñada para el tratamiento de datos. La idea es usar varios tipos de boxplots (el clásico, el cual es adaptado para la consideración de sesgo de los datos, el boxplot trasladado, y el gráfico de cajas de percentiles), el gráfico violin, y el histograma con un estimador no paramétrico de la densidad. Los gráficos son presentados de forma que faciliten la extracción de información adicional. Se muestra como una buena cantidad

[a]Professor. E-mail: rayospina@gmail.com
[b]MSc Candidate. E-mail: amlarangeiras@gmail.com
[c]Professor. E-mail: acfrery@gmail.com

de información que puede ser extraída a través de ejemplos de imágenes de radar de apertura sintética. Se presenta su implementacón en `R` basada en funciones actualmente disponibles.

***Palabras clave*:** análisis exploratorio de datos,boxplot, datos sesgados gráficos de violin, visualización.

1. Introduction

Tukey's (1977) work set the basis for Exploratory Data Analysis (EDA), which is the art of seeking relevant information from the data with the least possible distributional assumptions about the underlying process. Such a quest is frequently based on graphical representations.

Among the schematic plots that survived or emerged since the advent of powerful personal computers, one should mention, for univariate data, the scatterplot, the histogram (defined and discussed in Pearson 1895), the boxplot (Tukey 1977), the adjusted boxplot (Hubert & Vandervieren 2008), the shifting boxplot (Marmolejo & Tian 2010), the violin plot (Hintze & Nelson 1998), and their many variations.

The histogram and the boxplot are the most used plots which convey information about the shape of the underlying distribution. They work in the same fashion; they extract and display key quantifiers from the data. These quantifiers can be tuned for specific situations as, for instance, the choice of the bins in the histograms (the Freedman-Diaconis, Sturges, and Scott options in the `hist` function available in `R`).

A key point to note when using more than a single graphical presentation of the same dataset is to clearly convey the same or complementary information. A common mistake is simply showing several summaries side-by-side, but the precision and extent of enhancement that such a juxtaposition provides is arguable. Since no single plot is able to provide all the relevant information in every conceivable case, a possible solution for this problem consists in presenting the plots with clear visual clues of their same origin: the dataset. Visualization techniques are often used to drive important decisions. If the information conveyed by the graphical summaries is flawed, decisions may be biased or completely wrong.

For instance, Doulgeris, Anfinsen & Eltoft (2011) present a segmentation procedure for Polarimetric Synthetic Aperture Radar (PolSAR) imagery which, albeit automatic, exhibits the quality of the product at each iteration in the form of histograms overlapped with fitted densities; the closer the fit, the better the result. When the data are overly asymmetric, the automatic presentation is hard to grasp as the abscissas span a huge interval.

In this work, we present a tool for the visual display of skewed data developed in R that is freely available. The tool is based on the integration and coordination of several graphical representations, some of which are tailored to this kind of data. We test the tool on PolSAR data, which exhibit intense asymmetry.

Section 2 presents the graphical summaries that will be integrated in our visualization tool. Section 3 presents the data, which is from different types of land covers as retrieved by Synthetic Aperture Radar (SAR) sensors. These kinds of data are prone to presenting extreme deviations from the Gaussian hypothesis, as they are heavily skewed. Finally, section 4 concludes the paper with further suggestions. The Appendix provides details about the implementation and instructions to obtain the code and the data.

2. The Summaries and Their Coordination

We start this section with a brief presentation of the data that will be used to illustrate the graphical summaries.

One of the main goals of remote-sensing is to capture and analyze information scenes concerning the Earth surface. The PolSAR technology has achieved an important position among the remote-sensing modalities (Mott 2007). A polarimetric radar transmits two orthogonal waves, in either horizontal (H) or vertical (V) polarization, and receives backscattered waves in either H or V polarization, yielding four resultant combinations of complex signals: HH, VV, HV, and VH. Polarimetric SAR systems employ *coherent illumination* and, as a consequence, their resulting images are contaminated with fluctuations on its detected intensity called "speckle". Speckle significantly degrades the perceived image quality, as much as the ability of extracting information from the data.

In the remainder of this section we comment on the advantages and disadvantages of some commonly used graphical summaries of data. These summaries are illustrated with the same dataset: a sample from a PolSAR image of the Niigata area; see Figure 5(c). As presented in Table 1, the VV polarization is the one with the strongest asymmetry; so the data henceforth presented come from this band. More information about this and other images is given in Section 3.

2.1. Histograms and Kernels

Estimates of the underlying probability density function are one of the main tools to extract information about the distribution of the underlying population in EDA. The estimated density may reveal patterns and features representative of the targeted object for data modeling, analysis, and decision-management.

Generally speaking, the problem of density estimation can be defined as the processes of estimating the unknown distribution by means of the also unknown density function f defined on the observations of a random sample of size n, namely $\mathbb{X} = \{X_1, \ldots, X_n\} \subset \mathcal{X}$ drawn from the target distribution.

The following features from the data play an important role in EDA: the minimum and maximum values (the difference between them is the range), the first, second (median), and third quartiles, labeled Q_1, Q_2, Q_3, and the interquartile range $IQR = Q_3 - Q_1$.

The histogram is a basic form of nonparametric density estimator where the region covered by $\mathcal{X}$ is usually divided into equal-sized bins whose height is proportional to the count of hits within that bin. This estimator depends on the choice of the bin width h and on the starting points of the bins. These two values determine how the data will be grouped, i.e., to which bin each observation will belong to.

The number of bins k is related to the bin width h, for instance, $k = \text{range}/h$. Different rules to choose h are available. For example, Scott (1979) proposed $h = 3.5\widehat{\sigma}/n^3$ with $\widehat{\sigma}$ the sample standard deviation, while Sturges (1926) proposed $k = 1 + \log_2(n)$. Freedman & Diaconis (1981) proposed $h = 2IQR/n^3$. These methods usually render different graphical summaries of the data. Additionally, all histograms present the inconvenient feature of nonsmoothness (Silverman 1986), which, as noted by Cassetti, Gambini & Frery (2013), may make them unsuitable for parameter estimation techniques based on stochastic distances.

Thus, Rosenblatt (1956) and Parzen (1962) developed the kernel density estimator that is smooth, controls bin boundary effects, and (under very mild conditions) also converges to the true density, but faster than the histogram. A "kernel" is any smooth function (generally, a symmetric probability density) that depends on the bandwidth parameter h which controls both the spread and the orientation. As in histograms, h determines the smoothness of the estimation. In practice, the choice of the kernel is less important than that of h. For example, a small value of h will lead to under-smoothing and masking important features of the data, such as skewness and multimodality. On the other hand, rough curves produced by larger values of h yield smoother estimates but might dodge significant peaks or other important structures (Silverman 1986). Generally speaking, the choice of the bandwidth is based on the idea of minimizing the Mean Square Error (MSE) and many of these methods are discussed by Mugdadi & Ahmad (2004). We use `R` to produce high-quality and fully customizable histograms and kernel density estimates with the `hist` and `density` functions, respectively.

The effect of different choices of the bandwidth is illustrated in Figure 1. The curves are the Gaussian kernel estimates of the density function of the intensity VV data of Niigata, with $h = 0.008, 0.01, 0.05, 0.1, 0.15$, where h is the standard deviation of a zero-mean Gaussian density. We see how sensitive the estimate $\widehat{f}$ is in relation to h. Note that the plots reveal the strong data asymmetry. Smaller values of h produce far more fluctuations than larger ones. In particular, due to the combined effect of asymmetry and large data spread, the bulk of the information is confined to a small region, approximately in the interval $[0, .7]$, while there is little to visualize in the remaining area of the plot, which spans in $[.7, 2.5]$.

2.2. Boxplots and Variants

Tukey (1977) introduced the boxplot to analyze univariate datasets by graphically displaying important core statistics of the data. The plot is based on a few descriptive values and shows information about location, spread, skewness as well as the tails of the distribution.

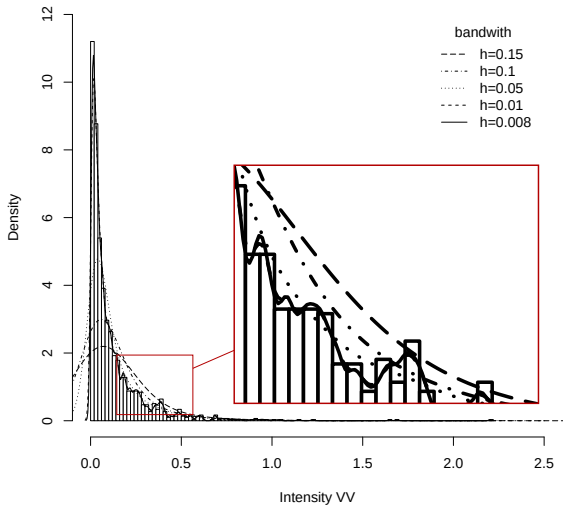


FIGURE 1: Histogram with kernel density estimation for the intensity VV of Niigata dataset for different values of the bandwidth h.

A vertical boxplot, as implemented in `R`, comprises the following elements:

1. A horizontal line at Q_2,
2. Horizontal lines close to or at Q_1 (lower hinge), and Q_3 (upper hinge); these lines form a box,
3. Compute the lower and upper fences $F_L = Q_1 - 1.5IQR$ and $F_U = Q_3 + 1.5IQR$; these are bounds that help in identifying outliers. Draw vertical lines from the upper and lower quartile up to the highest and lowest observations that are within the fences; these are the whiskers.
4. All observations beyond the fences are individually marked.

Horizontal boxplots are obtained simply by rotating these graphical elements.

Boxplots look quite symmetric when data come from a symmetric distribution (the median to the middle of the two other quartiles, and the box in the middle of the two whiskers). They do not give information about the number of observations on which they are based.

Boxplots may also give information on the tails of the distribution. On the one hand, if the box is well separated from the two whiskers, we can deduce that the tails are not very short. On the other hand, if the box is close to the whiskers, it is an indication of short tails.

A variation of the boxplot is the notched boxplot (McGill, Tukey & Larsen 1978) which is useful for determining whether two samples were drawn from the same population in terms of their median values. The notch displays a 95% approximate confidence interval around the median based on the Gaussian hypothesis:

$Q_2 \pm 1.57 \cdot IQR/\sqrt{n}$. According to Chambers, Cleveland, Kleiner & Tukey (1983), although not a formal test, if two box notches do not overlap, there is "strong evidence" that their medians differ.

`R`'s default graphical tools include the `boxplot` function which has the option `notch=TRUE` to add a notch to the box. Boxplots in `R` are vertical by default, but the optional argument `horizontal=TRUE` renders horizontal boxplots.

In many situations, as with skewed data, the boxplot may erroneously identify as outlier values which exceed the whiskers. To correct this distortion, Hubert & Vandervieren (2008) proposed an adjusted boxplot for skewed distributions. The main idea is the inclusion of the *medcouple* introduced by Brys, Hubert & Struyf (2004) as a robust measure of skewness in the determination of the whiskers. The medcouple (MC) is defined as a scaled median difference of the left and right half of distribution, and hence not based on the third moment as the classical skewness. The Adjusted Boxplot can be useful as a fast tool for automatic outlier detection, without making any assumption about the distribution of the data. Lower outliers are below $\exp\{-3.5MC\}$, while upper outliers are above $\exp\{4MC\}$. The function `adjbox` of the `R` package `robustbase` can be used to produce this graphical representation.

Figure 2 illustrates these three types of boxplots with the Niigata VV dataset. Note that the classical and notched boxplots look alike; cf. Figure 2(a) and 2(b), respectively. This is typical of large samples, $n = 4,446$ in this case, for which the width of the notch becomes negligible.

The difference between both classical and notched boxplots and the adjusted boxplot shown in Figure 2(c) is noticeable. While the former two identify numerous observations as outliers, the latter only considers a few as surprising data. This last graphical representation is more adequate and may lead to better informed decisions than the former ones.

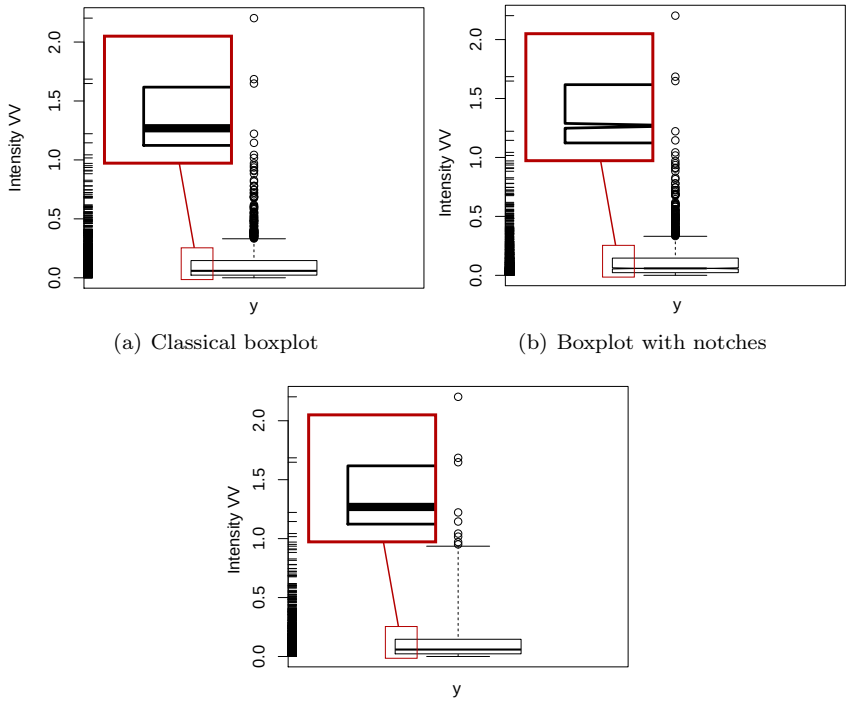
Esty & Banfield (2003) proposed the box-percentile plot as a variant of the boxplot which allows the sides of the plot to convey more information, presenting details about the distribution of the data. The width of the box is not fixed, but is proportional to the number of data. In this way, the box-percentile plot summarizes more than the histogram, but shows more details than the boxplot. The `HMisc` package in `R` computes and displays this graphical summary with the function `bpplot`.

Recently, Marmolejo & Tian (2010) provided a comprehensive literature review on boxplots, and proposed a variant, the shifting boxplot. This graphical summary incorporates the mean as basilar information instead of the median. The methodology supports conducting parametric tests.

A vertical shifting boxplot is a compound of nine quantities about a normally distributed batch of data (Marmolejo & Tian 2010):

(1) Smallest outlying observation,

(2) Minimum value within the $\pm 2s$ range, where s is the sample standard deviation,



(a) Classical boxplot

(b) Boxplot with notches

(c) Adjusted boxplot

FIGURE 2: Boxplots for the intensity Niigata VV dataset, with zooms of the notches.

(3) Mean of the first half of the data ($Q_{1\bar{x}}$),

(4) Lower 95% confidence interval (CI) limit for the mean,

(5) Mean ($Q_{2\bar{x}}$),

(6) Upper 95% CI limit for the mean,

(7) Mean of second half of the data ($Q_{3\bar{x}}$),

(8) Maximum value within the $\pm 2s$ range,

(9) Largest outlying observation.

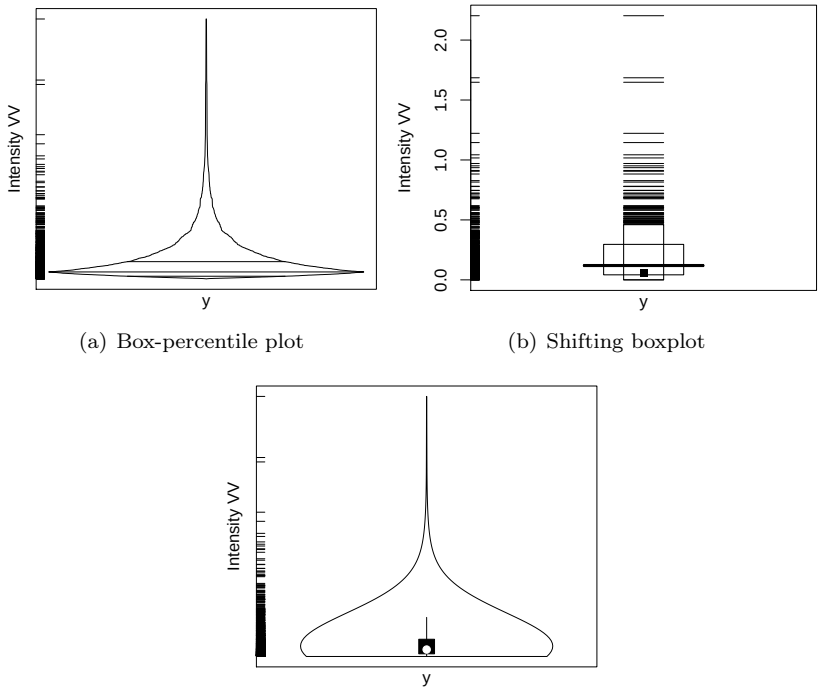
Observations that fall below $-2s$ and above $2s$ are represented by dashes. Observations that are equal to or above $-2s$ and equal or below $2s$ are enclosed by the innermost box, which covers approximately 95% of the data and is similar to the whiskers in the traditional boxplot. Observations that fall between the mean of the first half of the data ($Q_{1\bar{x}}$) and the mean of the second half of the data ($Q_{3\bar{x}}$) are enclosed by the middle box. This box covers approximately 50% of the observations that are closest to the mean.

The thick horizontal line joining the borders of the outermost box represents the mean of the data set, and the horizontal limits of this box represent the lower and upper limits of the confidence interval. The Shifting boxplot is displayed as three boxes and as many dashes as outlying observations exist.

Modifications to the shifting boxplot incorporate more accurate confidence intervals and robust estimators of central tendency to explore non-normal data (Marmolejo & Tian 2010).

Finally, Hintze & Nelson (1998) presented the violin plot; a combination of a boxplot and a (doubled) kernel density plot. The violin plot does not include the individual points, but it displays the median and a box indicating the interquartile range. It is useful when comparing multiple groups and with large datasets. The violin plot reveals important information about the tails, multimodality and stability of the smoothed data. There is also a `vioplot` package in `R`.

Figure 3 presents these last three graphical representations of the Niigata VV dataset. Although visual complexity is somewhat increased from the plots presented in Figure 2, the amount of visual information is also enhanced.



(a) Box-percentile plot

(b) Shifting boxplot

(c) Violin plot

FIGURE 3: Modified boxplots for the intensity Niigata VV dataset.

The problem is how to efficiently exploit the information conveyed by these types of graphical representations. We propose a solution in the next section.

2.3. The Proposed Tool

We discussed a number of graphical summaries in the previous section. Each one conveys an important aspect of the properties of the sample. We are interested in producing a graphical representation of the data able to help decision processes as, for instance, the choice of a specific distribution for the data. In particular, PolSAR image interpretation is not easily carried out except by specialists and such a representation should help in this task.

To adequately assess highly asymmetric data, and to extract as much information and features as possible from the dataset, we propose the simultaneous use of the histogram enhanced with density estimation, the boxplot with notches (B-N), the violin plot (V-P), the shifting boxplot (S-P), the adjusted boxplot (A-B), and the box-percentile plot (B-P).

All graphical summaries are coordinated with respect to a location parameter estimate, for example, the median. The kernel density estimation employs the Gaussian kernel and the bandwidth is chosen by the "rule-of-thumb" suggested by Silverman (1986).

The proposed tool follows the guidelines proposed by Tufte (2001) for quality visual display of quantitative information in particular, the ratio information/ink was maximized by showing only graphical elements, which convey the essential information.

The median is shown as a vertical line connecting all plots. A rug plot is displayed in the middle of the summaries to reinforce the position of the data, and to visually emphasize the notion that the data are common to all plots.

Figure 4 presents the result of computing the proposed visualization on the Niigata VV dataset.

3. Application to SAR Data

SAR sensors provide information about the target which complements the one provided by optical sensors. Their use has proven valuable for as diverse applications as the mapping of the surface of Venus by the Magellan and Venera missions (Arvidson, Schulte, Kwok, Curlander, Elachi, Ford & Saunders 1988), the unearthing of lost Maya ruins (Adams, Brown & Culbert 1981), and the 4D (space and time) monitoring of the environment (Moreira, Prats-Iraola, Younis, Krieger, Hajnsek & Papathanassiou 2013).

The main characteristics making the data provided by SAR sensors valuable are (i) their ability to produce images with high spatial resolution independently from daylight, cloud coverage, weather and environmental conditions like fog, smoke, smog, rain, etc., and (ii) the fact that their return is the result of complex interactions between the incident signal and the target, which complement the information available in the visible and near-visible spectrum. The recent tutorial by Moreira et al. (2013) is an excellent starting point for the reader interested in this field.

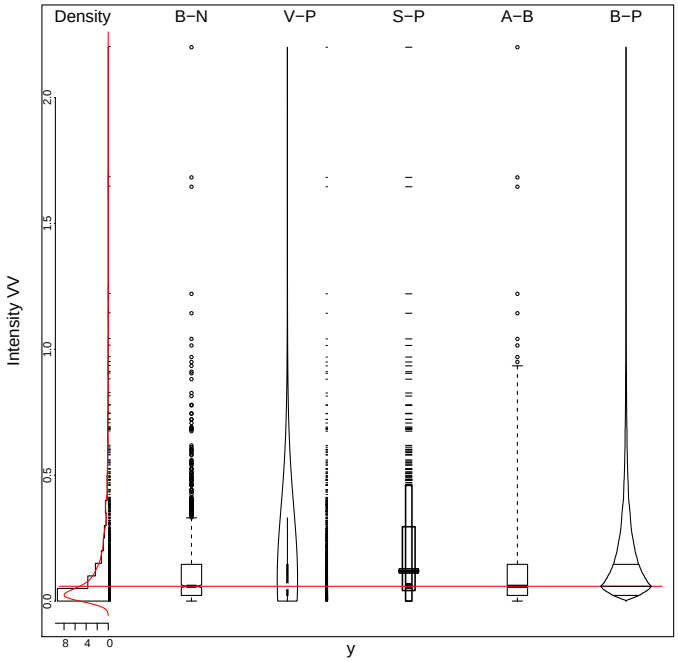


FIGURE 4: Synchronized graphs in the median of intensity VV of Niigata dataset.

The simplest form that SAR data adopt is the intensity. Mejail, Jacobo-Berlles, Frery & Bustos (2003) presented evidence that the $\mathcal{G}_I^0$ model is able to describe many types of target textures, from textureless (such as crops) to extremely textured (as, for instance, urban areas), but including areas with moderate texture (e.g. forests).[1] This model was proposed by Frery, Müller, Yanasse & Sant'Anna (1997), and was later extended to the full polarimetric case by Freitas, Frery & Correia (2005). Depending on a relationship between the last two parameters, the rth-order moment of this distribution is infinite. The visualization of both the density and data from this distribution can be demanding, since extreme observations are expected.

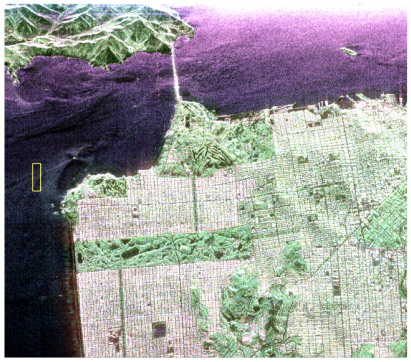
Among the works that present qualitative analyses of SAR data, Frery, Correia & Freitas (2007) and Doulgeris et al. (2011) make critical decisions based on such information. The former decides which joint distribution will be used for the classification, while the latter forms a stopping rule for an iterative segmentation procedure.

[1] The $\mathcal{G}_I^0$ intensity model, denoted by $Z \sim \mathcal{G}_I^0(\alpha, \gamma, L)$, is characterized by the density function

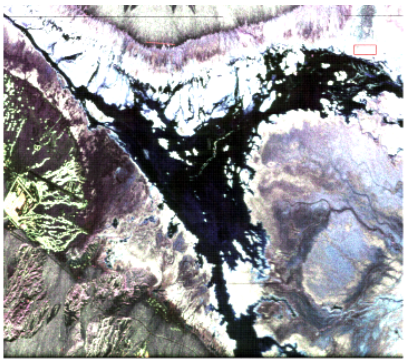
$$f_Z(z) = \frac{L^L \Gamma(L-\alpha)}{\gamma^\alpha \Gamma(L)\Gamma(-\alpha)} \frac{z^{L-1}}{(\gamma + Lz)^{L-\alpha}}, \qquad L \geq 1, -\alpha, \gamma, z > 0. \tag{1}$$

The parameters that index such a distribution are: (i) the number of looks L, which is a measure of the signal-to-noise ratio, (ii) the scale parameter γ, which is related to the relative strength between the incident and reflected signals, and (iii) the roughness parameter α, which relates to the target texture.

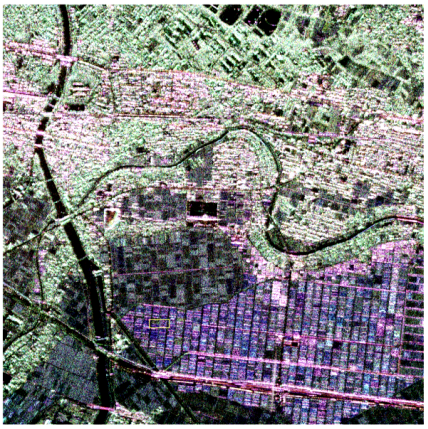
Below, we present the data that will be analyzed with the proposed technique. Figure 5 presents the color composites of three images from different PolSAR sensors and areas. Figure 5(a) is from the San Francisco bay, and the area under analysis is highlighted in yellow; it is a textureless sample from the sea. Figure 5(b) is from Death Valley, and the sample in red has moderate texture. Figure 5(c) is from Niigata, and the sample in yellow has extreme texture since it is from an urban area.



(a) San Francisco, sample in yellow



(b) Death Valley, sample in red



(c) Niigata, sample in yellow

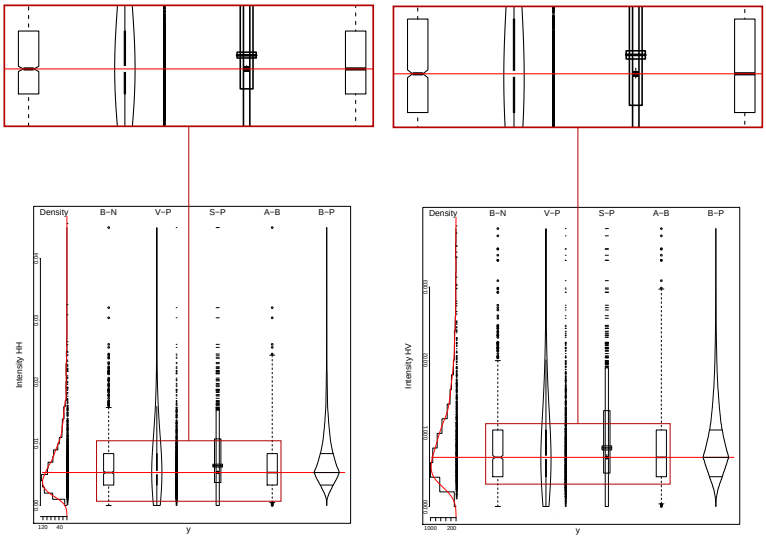
FIGURE 5: Color composites and samples

Table 1 presents quantitative summary information about these datasets. As expected, regardless the range, the image and the polarization of the data, there is intense skewness and kurtosis in all cases. The sample skewness and kurtosis are estimators for the third and fourth central moments by using the method of moments: $\hat{\eta}_3 = \hat{\mu}_3/\widehat{\sigma}^3$, and $\hat{\eta}_4 = \hat{\mu}_4/\widehat{\sigma}^4$, where, $\hat{\mu}_3 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^3$, $\hat{\mu}_4 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^4$ and $\hat{\sigma}^2 = s^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

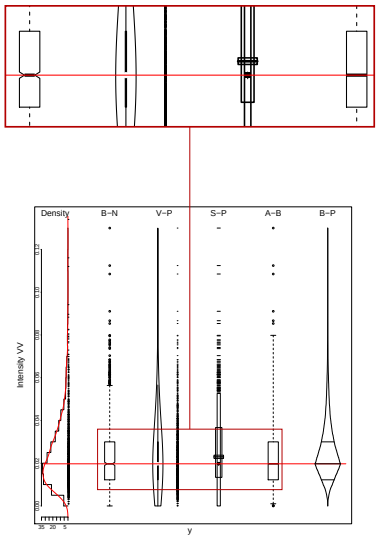
Table 1: Summary statistics.

	San Francisco $n = 4,320$			Death Valley $n = 4,446$			Niigata $n = 4,446$		
	HH	HV	VV	HH	HV	VV	HH	HV	VV
Min	1.45×10^{-4}	1.04×10^{-4}	8.92×10^{-4}	0.72	0.03	0.27	1.15×10^{-4}	5.37×10^{-7}	9.22×10^{-5}
1st Quartile	3.48×10^{-3}	5.04×10^{-4}	0.01	0.44	0.13	0.92	0.02	1.97×10^{-3}	0.02
Median	5.53×10^{-3}	7.64×10^{-4}	0.02	0.56	0.18	1.18	0.05	4.54×10^{-3}	0.06
Mean	6.64×10^{-3}	8.91×10^{-4}	0.02	0.58	0.18	1.23	0.10	0.01	0.12
3rd Quartile	8.62×10^{-3}	1.13×10^{-3}	0.03	0.69	0.22	1.49	0.11	0.01	0.15
Max	0.04	3.83×10^{-3}	0.13	1.45	0.48	3.15	1.42	0.04	2.20
Skewness	1.84	1.56	1.57	0.71	0.66	0.62	3.74	1.92	4.13
Kurtosis	9.34	6.60	7.49	3.59	3.77	3.64	23.37	7.78	32.12

Figures 6, 7 and 8 show the results of applying the proposed technique to each polarization of the samples from Figures 5(a), 5(b), and 5(c), respectively. All figures include a zoomed area, which enhances a particular region of the graphical summary. As can be seen from the values presented in Table 1, these datasets are comparable within each image, but the ranges differ widely among images.



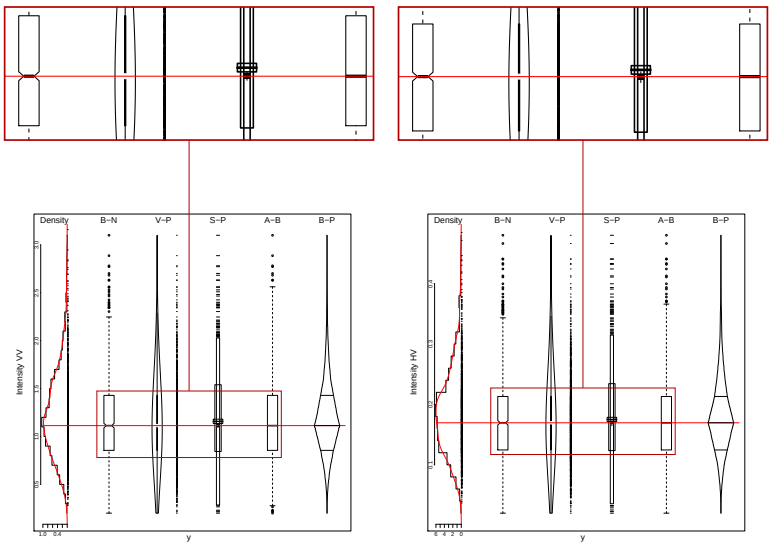
(a) Intensity HH (b) Intensity HV



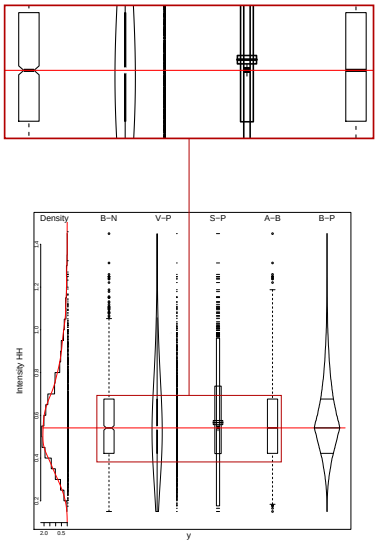
(c) Intensity VV

FIGURE 6: Synchronized graphs in the median of intensity bands of San Francisco dataset.

The samples presented in Figure 6 look alike when comparing the histograms and fitted densities, but important differences arise in the violin and shifting boxplots. In these graphical summaries, it is clearer that the HV band has more spread than the other two. The extent can be visually quantified by the shifting boxplots.



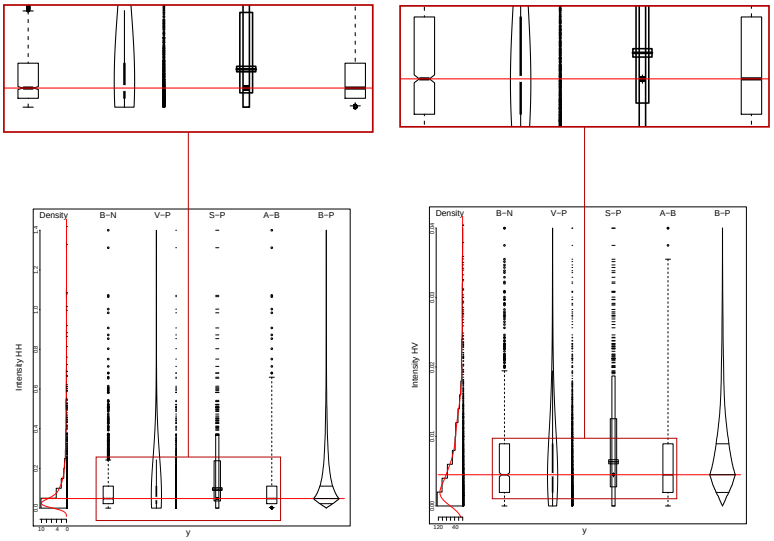
(a) Intensity HH (b) Intensity HV



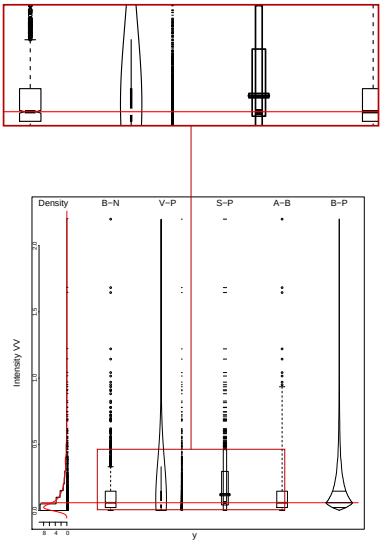
(c) Intensity VV

FIGURE 7: Synchronized graphs in the median of intensity bands of Death Valley dataset.

The data from the Death Valley image, summarized in Figure 7, are the most symmetric; this confirms the values presented in Table 1. Nevertheless, one observes in the boxplots that there are outliers to the right of the three samples. If asymmetry is assumed, the adjusted boxplot also detects outliers to the left of two of the three polarizations, namely in Figure 7(a) and 7(c).



(a) Intensity HH (b) Intensity HV



(c) Intensity VV

FIGURE 8: Synchronized graphs in the median of intensity bands of Niigata dataset.

Figure 8 shows at a glance the different behaviors among polarizations, with the HV channel exhibiting more spread than the other two; cf. the box-percentile and violin plots. Although HH and VV polarizations behave alike, the adjusted boxplot reveals that the former has more outliers than the latter.

4. Conclusions and Future Work

As presented in the examples, the use of synchronized graphical summaries promoted the discovery of information conveyed by the data. If only one type of plot had been used, some of these features would not have been identified. Synchronization is essential for retaining the ability to compare graphical representations of the same dataset. A loose presentation of two or more of the plots would not allow the discovery of such information. More customizable options are being added to the tool as, for instance, the ability to interactively choose the order in which the plots appear. In its current version, the function does not return any object. This can be easily customized, for instance using lists.

Acknowledgments

The authors thank Dr. Fernando Marmolejo-Ramos for providing the `R` codes of the shifting boxplot. They also thank three anonymous referees for their thoughtful comments, suggestions and criticisms. The study was supported partially by CNPq and Fapeal grants, from Brazil.

[Recibido: mayo de 2014 — Aceptado: octubre de 2014]

References

Adams, R. E. W., Brown, W. E. & Culbert, T. P. (1981), 'Radar mapping, archeology, and ancient Maya land use', *Science* **213**(4515), 1457–1468. doi: 10.1126/science.213.4515.1457.

Arvidson, R., Schulte, M., Kwok, R., Curlander, J., Elachi, C., Ford, J. P. & Saunders, R. (1988), 'Construction and analysis of simulated Venera and Magellan images of Venus', *Icarus* **76**(1), 163–181. doi: 10.1016/0019-1035(88)90149-2.

Brys, G., Hubert, M. & Struyf, A. (2004), 'A robust measure of skewness', *Journal of Computational and Graphical Statistics* **13**(4), 996–1017. doi: 10.1198/106186004X12632.

Cassetti, J., Gambini, J. & Frery, A. C. (2013), Parameter estimation in SAR imagery using stochastic distances, *in* 'Proceedings of The 4th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR)', Tsukuba, Japan, pp. 573–576.

Chambers, J., Cleveland, W., Kleiner, B. & Tukey, P. (1983), 'Graphical methods for data analysis', *The Wadsworth Statistics/Probability Series. Boston, MA: Duxury* .

Doulgeris, A. P., Anfinsen, S. N. & Eltoft, T. (2011), 'Automated non-Gaussian clustering of polarimetric synthetic aperture radar images', *IEEE Transactions on Geoscience and Remote Sensing* **49**(10), 3665–3676.

Esty, W. W. & Banfield, J. D. (2003), 'The box-percentile plot', *Journal of Statistical Software* **8**(17).

Freedman, D. & Diaconis, P. (1981), 'On the histogram as a density estimator: L2 theory', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **57**(4), 453–476.

Freitas, C. C., Frery, A. C. & Correia, A. H. (2005), 'The polarimetric G distribution for SAR data analysis', *Environmetrics* **16**(1), 13–31.

Frery, A. C., Correia, A. H. & Freitas, C. C. (2007), 'Classifying multifrequency fully polarimetric imagery with multiple sources of statistical evidence and contextual information', *IEEE Transactions on Geoscience and Remote Sensing* **45**(10), 3098–3109.

Frery, A. C., Müller, H.-J., Yanasse, C. C. F. & Sant'Anna, S. J. S. (1997), 'A model for extremely heterogeneous clutter', *IEEE Transactions on Geoscience and Remote Sensing* **35**(3), 648–659.

Hintze, J. L. & Nelson, R. D. (1998), 'Violin plots: A box plot-density trace synergism', *The American Statistician* **52**(2), 181.

Hubert, M. & Vandervieren, E. (2008), 'An adjusted boxplot for skewed distributions', *Computational Statistics & Data Analysis* **52**(12), 5186–5201. doi: 10.1016/j.csda.2007.11.008.

Marmolejo, R. F. & Tian, T. S. (2010), 'The shifting boxplot: A boxplot based on essential summary statistics around the mean', *International Journal of Psychological Research* **3**(1), 37–45.

McGill, R., Tukey, J. W. & Larsen, W. A. (1978), 'Variations of boxplots', *The American Statistician* **32**(1), 12–16.

Mejail, M. E., Jacobo-Berlles, J., Frery, A. C. & Bustos, O. H. (2003), 'Classification of SAR images using a general and tractable multiplicative model', *International Journal of Remote Sensing* **24**(18), 3565–3582.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I. & Papathanassiou, K. P. (2013), 'A tutorial on synthetic aperture radar', *IEEE Geoscience and Remote Sensing Magazine* **1**(1), 6–43.

Mott, H. (2007), *Remote Sensing with Polarimetric Radar*, Wiley-IEEE Press, USA.

Mugdadi, A. R. & Ahmad, I. A. (2004), 'A bandwidth selection for kernel density estimation of functions of random variables', *Computational Statistics & Data Analysis* **47**(1), 49–62.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *The Annals of Mathematical Statistics* **33**(3), 1065–1076.

Pearson, K. (1895), 'Contributions to the mathematical theory of evolution II: Skew variation in homogeneous material', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **186**(0), 343–414.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org/

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *The Annals of Mathematical Statistics* **27**(3), 832–837.

Scott, D. W. (1979), 'On optimal and data-based histograms', *Biometrika* **66**(3), 605–610.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Sturges, H. A. (1926), 'The choice of a class interval', *Journal of the American Statistical Association* **21**(153), pp. 65–66.

Tufte, E. R. (2001), *The Visual Display of Quantitative Information*, 2 edn, Graphics Press.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley, USA.

Appendix. Implementation

The tool for producing synchronized plots was written in the `R` programming language (R Core Team 2013). The code involves functions from other freely available packages: `Hmisc`, `robustbase`, `vioplot`, `bootstrap`, `MASS`, `lfstat`, `graphics`, `gplots` and some new implementations for horizontal histogram, line density estimates in violin plots, and boxplot percentile.

The proposed new `R` function, termed `SyncronizedPlot`, joins in coordinated form the histogram enhanced with a density estimation, the boxplot with notches (B-N), the violin plot (V-P), the shifting boxplot (S-P), the adjusted boxplot (A-B), and the box-percentile plot (B-P). The main argument that must be supplied is `data`, a numeric vector or an `R` object which is coercible to one by `as.vector(x, "numeric")`

The time required to produce an output is negligible.

The code and data used in this work are freely available from `http://www.de.ufpe.br/~raydonal/SynchronizedPlots/SynchronizedPlots.zip`. To try the tool, the user must load the `R` scripts, and issue the following `R` commands:

```
# Data
x=rgamma(100, shape=0.5)
# Plot
SyncronizedPlot(x)
```