Graphic modeling of radioecological data using Violinplot

Modelagem gráfica de dados radioecológicos usando Violinplot

DOI: 10.55905/oelv21n9-197

Recebimento dos originais: 28/08/2023
Aceitação para publicação: 25/09/2023

Cleomacio Miguel da Silva
PhD in Energy and Nuclear Technologies
Institution: Universidade de Pernambuco (UPE)
Address: Rua Amaro Maltês de Farias, s/n, Nazaré da Mata - PE
E-mail: cleomacio@hotmail.com

Lívia de Souza Alexandre
PhD student in Computer Science
Institution: Universidade Federal de Pernambuco (UFPE)
Address: Avenida Jornalista Anibal Fernandes, s/n, Cidade Universitária, Recife
Email: isa5@cin.ufpe.br

Marcos Filipe Silva Lino
Bachelor in Biological Sciences
Institution: Universidade de Pernambuco (UPE)
Address: Rua Amaro Maltês de Farias, s/n, Nazaré da Mata - PE
E-mail: marcoshit9@hotmail.com

ABSTRACT
Data collection of environmental radioactive chemical contaminants is information of fundamental importance for decision-making in environmental planning and management. In this case, data analyses should be performed with a high level of significance within robust statistical planning. The objective of this study was to utilize the Violinplot as a graphical modeling tool for statistical analysis of radioecology data. To achieve this, a computational algorithm was developed in the Python language. The findings clearly demonstrate that the Violinplot is an outstanding data analysis tool for robust statistical planning in environmental management.

Keywords: computational mathematics, applied statistics, radioactive elements, outliers, environmental planning, environmental management.

RESUMO
Coleta de dados de contaminantes químicos radioativos ambientais são informações de fundamental importância para a tomada de decisão no planejamento e gestão ambiental. Neste caso, as análises dos dados devem ser realizadas com elevado nível de significância dentro de um planejamento estatístico robusto. Sendo assim, e dentro deste contexto, o
objetivo do presente trabalho foi realizar modelagem gráfica usando o Violinplot como ferramenta de análise estatística de dados em radioecologia. Para tanto, foi desenvolvido um algoritmo computacional na linguagem Python. Os resultados obtidos mostraram que o Violinplot é uma excelente ferramenta de análise de dados em planejamento estatístico de gestão ambiental.

Palavras-chave: matemática computacional, estatística aplicada, elementos radioativos, outliers, planejamento ambiental, gestão ambiental.

1 INTRODUCTION

In Brazil, the inclusion of strategic monitoring plans for natural and artificial radionuclides in environmental management projects conducted by public and private agencies has been significantly overlooked. Brazil experiences a high level of intensity in mining and nuclear medicine activities, that’s why there’s a need to establish radioecological monitoring planning. Also, Brazil has several natural radioactive anomalies that also need to be monitored, as they make radionuclides available for the food chain.

The planning of radio-environmental management should contain specific objectives to avoid the degradation of the natural environment, establish control and monitoring guidelines, and depending on the case, develop a strategy for the recovery of existing liabilities. The Sustainable Development Goals (SDGs) are part of the so-called “2030 Agenda”. It’s a global pact signed during the United Nations Summit in 2015 by 193 member countries. This agenda consists of 17 ambitious and interconnected goals, broken down into 169 targets with focus on overcoming the key development challenges faced by people in the world, promoting global sustainable growth by 2030. In Goal 15 (SDG 15) is written that “terrestrial life: protect, recover and promote the sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, halt and reverse land degradation and halt biodiversity loss” (UNITED NATIONS, 2023). The important phrase “halt and reverse the degradation of the Earth” must also be understood from the point of view of minimizing the radio-environmental impact caused by the presence of natural and artificial radioactive chemical contaminants in biotic and abiotic media.
The radiotoxic effects caused by natural or artificial radionuclides in the food chain should be very well explained in the radioecological planning of environmental management. The specialized literature is abundant in describing the toxic effects caused by various environmental radioactive chemical contaminants that are efficiently transferred to the food chain (VOGELTANZ-HOLM; SCHWARTZ, 2018; BULUBASA et al., 2021). Depending on the concentration level of a given radioactive chemical contaminant present in the human body, the risk of teratogenic and carcinogenic effects is significantly increased (VOGELTANZ-HOLM; SCHWARTZ, 2018; BULUBASA et al., 2021). Due to the presence of radioactive chemical contaminants in different compartments of the environment, developing environmental monitoring programs is a task of great responsibility for public and private agencies (WIERSEMA, 2017). In this case, the success in the decision-making of each agency will depend on how the Environmental Management Commission will develop its plans of goals. On the other hand, the excellence that must be achieved in an environmental monitoring program will depend on excellent specific and strategic statistical planning to analyze the data obtained.

Data sets of values of concentrations of environmental chemical contaminants have a strong asymmetry on the right, due to the presence of outliers or anomalous values (SINGH; SINGH; ENGELHARDT, 1997). As there’s no control over the concentration levels of a contaminant chemical element in the natural environment, the set of experimental data obtained from a given locality can vary from below the detection limit to anomalous values (SINGH; SINGH; ENGELHARDT, 1997). Thus, there’s great variability in the data set, which makes it difficult to find a measure of central tendency that is robust to represent the data obtained. As a result, the median is the most appropriate measure of central tendency to represent the most likely value in the analysis of environmental data (FRIGGE; HOAGLIN; IGLEWICZ, 1989). In this case, the graphic design of the data could be represented by the boxplot which is a statistical tool widely used by environmental scientists, especially by radioecologists.

However, it’s important to acknowledge the limitations of the boxplot as it does not provide a visual representation of the probability density function for the distribution of experimental data (PARZEN, 1962). In contrast, the Violinplot not only allows for all
the operations performed by the boxplot but also provides valuable insights into the probability density function of the distribution of experimental data. Thus, the Violinplot emerges as a powerful statistical tool for comprehensive data analysis.

Since a Violinplot is the combination of a boxplot and a probability density function per kernel (KDE), its geometric representation is a graphic design of great importance in the statistical analysis of data. Every numerical representation, as far as possible, becomes more scientifically elegant when the data acquires geometric configurations that help in its interpretation. Knowing how to interpret the conclusions and insights that each geometric configuration represents, it will help in making the right decision. In this context, the importance of thoughtful and effective data visualization becomes apparent. The graphical design of the Violinplot can serve as a valuable ally in the communication process. Thus, as the definition of the problem to be solved is the first step before starting the preparation and analysis of the data, choosing the type of graphic design is fundamental in building good storytelling of the results. In addition to helping in the interpretation of the data, the graphic beauty of the Violinplot can stimulate the cognitive load of the reader, increasing his interest in the results presented.

Size, color, and position on the page define the pre-attentive attributes of a Violinplot chart. These three attributes must be allied when performing geometric visualization of data using Violinplot. Used strategically and appropriately, Violinplot graphic design helps direct the reader's attention to what is needed, thus creating a visual hierarchy of components. This directs your receiver to the information you want to communicate and the way you want it to be processed. Size, for example, denotes relative importance: the larger, the more important. Color is not just a resource for coloring the chart, it is one of the most powerful tools to draw attention to specific data. It’s worth, for example, choosing a basic color for the scale (black) and highlighting specific data with a striking color (blue). Blue is a joker: it avoids color blindness problems and prints well in black and white if printing is needed.

Thus, and within this context, the objective of the present work was to use the graphic modeling using the Violinplot as a statistical tool for data analysis of natural radionuclides for decision-making in environmental management.
2 METHODOLOGY

2.1 CONSTRUCTION OF THE GRAPHIC DESIGN OF VIOLINPLOT

The graphic design of Violinplot was built through an algorithm developed in the computer programming language Python (Figure 1). The internal structure of the Violinplot consisted of a boxplot and a probability density function per kernel. Figure 2 shows an example of how to integrate a boxplot and a Violinplot in the statistical analysis of data (HAMERMESH, 1994).

```python
import matplotlib.pyplot as plt
import pandas as pd

data = pd.read_csv("YOUR DATASET")

fig, ax = plt.subplots()

violin = ax.violinplot(data, positions=None, vert=True, width=0.5, showmeans=False, showextrema=True, showmedians=True, quantiles=None, points=100, bw_method=None, data=None)

box = ax.boxplot(data, positions=[1], widths=0.2, showfliers=False)

for pc in violin['bodies']:
    pc.set_facecolor('#00BFD7')
    pc.set_edgecolor('black')
    pc.set_alpha(1)

ax.set_ylabel('Y LABEL')
plt.xticks([1], ["LABEL"])
plt.title( 'TITLE')
plt.show()
```

Source: the authors.

The boxplot is an important statistical analysis tool because it has as measures the first quartile ($Q_1$), the median that is the second quartile ($Q_2$), the third quartile ($Q_3$), and the interquartile range (IQR). Because it has the median as a measure of central tendency, the boxplot is widely used to analyze data sets with discrepant values, helping in the identification of outlier or anomalous values. In this case, since the median is not significantly influenced by outliers, it becomes the best measure of central tendency to represent the data set. Quartiles have great relevance for locating possible outliers in the
dataset. If $Q_1$ and $Q_3$ are, respectively, the first and second quartiles, then it’s possible to obtain a rule to determine lower and upper extreme values of locating outlier values. In this case, one should consider the interquartile range (I IQ), where: $I IQ = Q_3 - Q_1$. The lower anomalous value is calculated by $Q_1 - k(Q_3 - Q_1)$; while the upper anomalous value is determined by $Q_3 + k(Q_3 - Q_1)$, $k = 1.5$ (which is normally accepted). In addition, a change in the value of “k” can have a significant impact on the criteria for locating anomalous values. Figure 3 shows the statistical measurements of a boxplot (FRIGGE; HOAGLIN; IGLEWICZ, 1989).

Figure 2. Integration between a boxplot and a Violinplot.

Source: Modified from Hamermesh (1994).
Figure 3. Configuration of a typical boxplot.


A density estimator per kernel (KDE), according to Parzen (1962), is obtained by superimposing kernel functions, as described in Equation 1, centered on each of the elements of the dataset obtained $x_i (i = 1, 2, \cdots N)$ from the samples.

Thus, the density estimate $\hat{f}(x_t)$ on point $x_t$ depends only on the spatial relationship between $x_t$ and sample elements $x_i (i = 1, 2, \cdots N)$, quantified by the metric built-in into the kernel function. In general, Equation 2 describes an univariate density estimator per kernel (PARZEN, 1962).

\[
k(x_i, \frac{1}{\sqrt{2\pi}}x_j) = e^{-\left(\frac{x_i-x_j}{h}\right)^2}
\]

(1)

Where:

“h” is the radius or standard deviation of the Gaussian function and $k(x_i,x_j) = a_{ij}$.
\[ \hat{f}(x_t) = \frac{1}{Nh} \sum k(x_t, x_i) \]  

(2)

Where:

\( N \) is the number of samples, \( h \) is the smoothing parameter of the kernel, and \( k(x_t, x_i) \) is the kernel operator, whose integral \( \int k(u) du \) must be unitary. The argument of the function \( k(\cdot) \) is, actually, the point where one wants to do the estimation, because the data obtained from the samples \( x_i (i = 1, 2, \ldots, N) \) are fixed and provided beforehand.

The width “\( h \)” of the kernel is the only parameter to be determined in KDE, which is responsible for smoothing the curve chosen to perform the estimation. The value of “\( h \)” has great importance in the form of the distribution of the sample data. Hence, your choice must be carried out quite carefully. The problem of kernel width estimation is unsupervised since the function generating the data is not known in advance and, therefore, it would not be possible in principle to minimize an error function to obtain this parameter.

Despite this, the method of estimation of the width presented by Silverman (1986), is based on the minimization of the integrated mean square error (MISE), to find the value of “\( h \)” that obtains the best for the generating function “\( f \)”. MISE was calculated using Equation 3. \( \hat{f}(x) \)

\[
\text{MISE}(\hat{f}(x)) = \int E[\{\hat{f}(x) - f(x)\}^2] dx = \int \{E\hat{f}(x) - f(x)\}^2 dx + \int \text{var}\hat{f}(x) dx
\]  

(3)

According to Equation 3, the MISE value can be expressed in terms of the sum of the integrated bias and the integrated variance. According to Silverman (1986), bias can be represented by \( \frac{1}{2} h^2 f'''(x) k_2 \) and variance by \( \frac{1}{nh} \int k(t)^2 dt \). Equation 3 is therefore rewritten in such a way that Equation 4 was obtained.
\[ \text{MISE}(\hat{f}(x)) = \frac{1}{4} h^4 k^2 \int f''(x)^2 \, dx + \frac{1}{nh} \int k(t)^2 \, dt \] (4)

Where:

“h” is the width of the kernel, what is a constant from the second term of the Taylor series expansion, \( f''(x) \) is the second derivative of the generating function, “n” is the sample size of the data, and “k” is the kernel function used. Silverman (1986) proposes that the optimized “h” value would be the one that minimizes MISE. However, since such a calculation depends on the second derivative of the data-generating function that is unknown and the type of kernel function used, the author assumes that the data was generated by a normal distribution and that the kernel is Gaussian. Thus, after some algebraic manipulations, the proposed value of “h” was calculated using Equation 5.

\[ h_1 = 1.06 \sigma n^{-\left(\frac{1}{5}\right)} \] (5)

When the data are normalized, Equation 6 is proposed as equivalent to Equation 5.

\[ h_{11} = \left(\frac{4}{n+2}\right)^{\frac{1}{n+4}} \left(\frac{N-1}{n}\right)^{\frac{-1}{n+4}} \] (6)

Where:

“n” is the number of dimensions and “N” is the number of samples.

Assuming that the generating function of the data is a Gaussian curve, the value given by \( h_1 \) will provide a good estimation of the density of the data. However, this statement is not always true. In this case, Silverman (1986) suggested that for a more robust scattering measure, the variance of the data should be replaced by the interquartile range (IIQ), as shown in Equation 7.

\[ h_2 = 0.79(\text{IIQ}) n^{-\left(\frac{1}{3}\right)} \] (7)

Where:
IIQ = Q₃ − Q₁ is the difference between the third and first quartile. However, the h₂ value further softens the estimate of bimodal distributions, thus generating the third kernel width proposal, as shown in Equation 8.

\[ h₃ = 0.9A_n^{-\left(\frac{1}{3}\right)} \]  \hspace{1cm} (8)

Where:

\[ A = \min\left(\sigma, \frac{IIQ}{1.34}\right) \]

Within the context analyzed, the kernel widths proposed by Equations 5, 6, 7, and 8 depend on two variables: the scattering measure (standard deviation or interquartile range) and the sample size. It is safe to say that the value of “h” will always be closely linked with the scattering of the data. Hence, the importance of knowing how to properly choose the value of this parameter.

Scott (2015) presented another proposal for calculating the width “h”, also based on minimizing the error of the estimated density function concerning the generating function of the unknown data. In this case, we used the asymptotic mean square error (AMISE), composed of the sum of the integrated variance and the square of the integrated bias, as shown in Equation 9.

\[ \text{AMISE} = \frac{R(K)}{nh} + \frac{1}{4} \sigma^4 h^4 R\left( f' \right) \]  \hspace{1cm} (9)

Where:

\[ R(g) = \int_{-\infty}^{\infty} g(u)^2 du \] is the roughness of function. Thus, the ideal “h” would be the one that minimizes the AMISE, whose calculation depends on the previous knowledge of the generating function. After some algebraic manipulations, Scott (2015) proposed using Equation 10.

\[ h₄ = 3 \left[ \frac{R(K)}{35\sigma^4_k} \right]^{\frac{1}{5}} \sigma_n^{-\left(\frac{4}{5}\right)} = 1.144\sigma_n^{-\left(\frac{4}{5}\right)} \]  \hspace{1cm} (10)
Where:

\[ R(K) = \frac{0.5}{\sqrt{\pi}} \text{ and } \sigma_k^2 = 1 \] assuming the Gaussian kernel. At less than one constant, Equation 10 equals the others, being composed of a scattering measure and a function of the sample size.

### 2.2 DATA

To ensure the integrity of data authorship and comply with data protection laws, the developed algorithm was tested using concentration values of natural radioactive elements derived from the authors’ own database. Specifically, the data used in this study was originated from research conducted by Silva (2000, 2019) in the field of radioecology, exploring various transfer pathways in the food chain. Although the present work specifically addresses radioecological data, it’s worth noting that the graphical modeling approach utilizing the Violinplot can be applied to analyze the probability distribution behavior of any dataset related to environmental contaminants.

### 3 RESULTS AND DISCUSSION

According to Wiersma (2007), environmental monitoring is important to determine and evaluate the levels of chemical contaminants present in the natural environment. In this case, before performing any data analysis, it’s necessary to understand the environmental information of each locality. This is key to optimizing investments, evaluating new possibilities, and connecting with stakeholders. Data analysis contributes to the construction of sustainability indicators. Thus, returning to Wiersma (2007), in a summarized way, it’s possible to say that environmental monitoring refers to the collection, recording, and analysis of environmental data in a systematic way to identify changes (whether positive or negative, natural or caused by human activities) in the quality of the environment of a given region.

For Vandecasteele (2004), environmental monitoring is a fundamental tool to support strategic, effective, and sustainable decision-making in radioecology, as it allows the impact caused by natural or artificial radionuclides in a given locality to be evaluated, adopting comprehensive actions and, in this way, carrying out corrective measures, in an attempt to mark the effects caused on the environment and consequently in humans.
To obtain quality data, Pentreath (2009) stated that radioecologists use many environmental monitoring technologies that include sensors and measurement devices, data collection systems, geographic information systems, data analysis platforms, and environmental modeling software. These technologies allow the collection of accurate and reliable information about the quality of air, water, soil, and vegetables, among others. Radioecological data collection involves the installation of sensors and measuring devices at specific locations for automated recordings. It’s also possible that these data are collected in loco, by specialized technical teams (PENTREATH, 2009).

In radioecology, data are sent to platforms, where they are processed and stored for further analysis (PENTREATH, 2009). In this context, different statistical methods of data analysis are available to radioecologists (STEIGER 2008; ACEVEDO 2012; OFUNGWU 2014). Thus, the introduction of Violinplot as a statistical tool for the analysis of radioecological data will contribute significantly to improving the graphical representation of data sets, as well as to provide robustness in the choice of a measure of central tendency and beauty in the design format.

Silva (2000) conducted important studies in radioecology on the presence of natural uranium, radon-222 ($^{222}\text{Rn}$), and polonium-210 ($^{210}\text{Po}$) in the public water supply of the uranium-phosphate region of the State of Pernambuco - Brazil. This region has high levels of uranium and thorium in rocks and soil, being considered by radioecological standards, as a typically anomalous site of natural occurrence of radionuclides (AMARAL, 1987). The concentrations of natural uranium, $^{222}\text{Rn}$ and $^{210}\text{Po}$ ranged from 32 to 282 mBq L$^{-1}$ (n = 14); 12 to 43 Bq L$^{-1}$ (n = 14), and 22 to 63 mBq L$^{-1}$ (n = 16), respectively, in samples of groundwater from the public supply. On the other hand, the concentration of $^{210}\text{Po}$ in surface water (river water) of public supply ranged from 22 to 52 mBq L$^{-1}$ (n = 16) (SILVA, 2000). These data were used to test the algorithm that was built in the Python computer language (Figure 1). According to Ramalho (2022), it’s fundamental and necessarily enough to test an algorithm to demonstrate its effectiveness and robustness in the operationalization and execution of the numerical and graphical modeling task. Figures 4, 5, and 6 shows, respectively, the graphical modeling of the
Violinplot for data on concentrations of natural uranium, $^{222}$Rn, and $^{210}$Po in groundwater. Figure 7 shows the graphical representation of the $^{210}$Po concentration data in river water.

Figure 4. Violinplot of natural uranium concentration in well water.

Source: the authors.
Figure 5. Violinplot of $^{222}$Rn concentration in well water.

Source: the authors.

Figure 6. Violinplot of $^{210}$Po concentration in well water.

Source: the authors.
Figures 4, 5, 6, and 7 show the probability distributions of the data with the respective statistical measures of the first and second quartile and median. The graphical representations and statistical quantities obtained showed that the algorithm constructed was efficiently robust to provide results for decision-making. Thus, Violinplot is a graphical tool that can be used in the analysis of environmental numerical data. The data presented in this work were obtained from groundwater and surface water samples from localities of anomalous natural radioactive occurrence that have high levels of uranium and thorium. According to Singh, Singh, and Engelhardt (1997), the distribution of contaminants in typically anomalous sites, has a high asymmetry to the right, caused by the effects of outliers. As can be seen in Figures 4, 5, 6, and 7, the data are distributed asymmetrically to the right because of the outlier values, as stated by Singh, Singh, Engelhardt (1997). Although the sample size is small, the Violinplot showed the trend of data distribution (Figures 4, 5, 6, and 7). This is very important from the point of view of environmental statistical planning, as it facilitates the understanding of the researched locality.
Silva (2019) studied two major natural radioactive anomalies existing in the Northeast region of Brazil. The first is located in the State of Pernambuco, in the municipalities of Pedra and Venturosa. The second is in the State of Paraíba, in the municipality of Pocinhos. Both anomalies have high concentrations of uranium and thorium in rocks and soil. Figure 8 shows the graphical representation of the Violinplot for the concentration of radium-228 ($^{228}$Ra) in the water samples of the Ipanema River located on the anomalies of uranium and thorium existing in the municipalities of Pedra and Venturosa, in the Agreste Semi-arid region of Pernambuco. In this case, the concentrations of $^{228}$Ra ranged from 71 to 2667 mBq L$^{-1}$ ($n = 12$). Figures 9 and 10 show the Violinplot of the concentrations of $^{226}$Ra and $^{210}$Pb in soil samples of uranium and thorium anomalies in the municipality of Pocinhos (Paraíba), whose values ranged from 1 to 289 Bq kg$^{-1}$ ($n = 42$) and from 7 to 5098 Bq kg$^{-1}$ ($n = 42$), respectively.

![Figure 8. Violinplot of $^{228}$Ra concentrations in Ipanema River water.](source)

Source: the authors.
Figure 9. Violinplot of $^{226}$Ra concentrations in soil.

Source: the authors.

Figure 10. Violinplot of $^{210}$Pb concentrations in soil.

Source: the authors.
Especially analyzing Figures 8 and 9, we observed how the outlier values “stretched” the distribution to the right, causing severe asymmetry. Statistically, the very discrepant anomalous values “stretch” the “strings” of the Violinplot (Figures 8 and 9). In this case, therefore, a thorough evaluation is necessary to know if there was an experimental error or if such values are characteristic of the region studied. According to Singh, Singh, and Engelhardt (1997), the presence of outliers in a typically anomalous location is almost certain. These authors also stated that the concentrations of chemical elements in typically anomalous locations can vary from the background to extremely high values. According to Downing and Clark (1998), once the probable causes of the existence of outliers are identified, they should be left in the statistical analysis of the data.

In radioecological studies on the determination of radionuclide concentrations in the environment, the presence of very distant values is sufficient reason for particular attention to be paid to the identification of the conditions under which such values were obtained. In fact, either such values are the result of errors in experimental procedures, or the distribution of the measurements is peculiar, deviating from what would usually be expected (MURTEIRA, 1990). Some authors even recommend that the analysis be initially performed by removing such values from the set considered. This type of practice can only be accepted provided that the experimental conditions that led to such values are known and identified and that there is a conviction that these conditions can be effectively considered anomalous. Modernly, we seek to build models that accommodate outliers and do not lead to their exclusion; they are models of contamination, of which there are numerous versions, such as those cited by Barnet and Lewis (1996). Since, in general, in radioecological studies the causes that generate discrepant values are not fully known, it is difficult to discriminate the measurement errors since some types of outliers are inherent to the population (SINGH; SINGH; ENGELHARDT, 1997).

Due to the high asymmetry in the radioecological data set, it is very common to use the lognormal distribution in the statistical analysis of the data (BLACKWOOD, 1992). The lognormal distribution has become a common choice among radioecologists to represent intrinsically positive and often highly discrepant environmental data in statistical analyses. However, the implications of its use are often not carefully considered,
which leads to misinterpretations of the data obtained. However, what must be evident in monitoring applications in radioecology, is what it means to assume lognormality in terms of data analysis and interpretation. According to Blackwood (1992), when assuming that a given set of data has a lognormal probability distribution, one should consider using normal theory methods on log data transformed for multiplicative errors and hypothesis testing on the original scale. As shown in Figures 4 to 10, it is incorrect to state always and categorically that data obtained from radioecological studies have a lognormal probability distribution, especially for small data sets.

Figures 4 to 10 observe trends for different types of data distribution that need to be studied with great attention and care. On the other hand, since Violinplot integrates the estimation of a probability density function per kernel (KDE) and a boxplot, without necessarily defining the type of distribution, but solving the fundamental problem of data smoothing, as shown by Parzen (1962), its statistical measures provided are robust, which makes the median an excellent measure of central tendency to represent the data set. A median is widely used to represent environmental data, as it is a statistical parameter quite resistant to the effects caused by outlier values and when the data set is small (BERTHOUEX; HUNTER, 1981). It is very common for radioecologists to make use of the median in their statistical analyses of data (SHEPPARD et al., 2006).

In addition to the lognormal probability distribution, the Weibull probability distribution function is also widely used in the statistical analysis of environmental data (MIKOLAJ, 1972). In fact, the Weibull distribution function has been applied empirically in radioecology in the analysis of atmospheric radioactivity datasets (APT, 1976). When comparing the use of Violinplot in the statistical analysis of radioecological data, concerning any type of probability distribution, the following advantages are obtained: (1) it does not depend on a specific type of probability distribution to obtain the median; (2) optimizes data smoothing; and (3) replaces with great statistical efficiency the histogram and the isolated boxplot.

The unquestionable advantage of the Violinplot graph over the boxplot is that, in addition to showing the pertinent statistical measures, it also shows the complete distribution of the data (Figures 4 to 10). This is very important, especially when it comes to
multimodal data, that is, a distribution with more than one peak. The kernel density plot (KDE) used to create the Violinplot graph is the same one added at the top of a histogram. The wider sections of the Violinplot graph (Figures 4 to 10) represent a higher probability of observations assuming a certain value, while the thinner sections correspond to a lower probability. Such information provides a good intuition about what a Violinplot chart is and what kind of information it contains. Certainly, Violinplot provides more information than boxplot and histogram in decision-making.

At the top of Figures 8 and 9, there are small, much more elongated distributions (thinner sections) compared to the center of each distribution (wider sections), with no distinct peaks. This indicates that there is a high possibility that the experimental (original) data belong to different populations. According to Singh, Singh, Engelhardt (1997), data obtained from the natural environment usually come from mixtures of populations (mixture of Gaussian distributions), especially if there are quite discrepant values. Certainly, the anomalies of the small distributions (thinner sections) of the Violinplot (Figures 8 and 9) were due to the effects caused by the outliers. This important information would have been lost if the graphical modeling had been performed using only the boxplot or histogram.

A comparative analysis can provide specified information that influences important decisions. Benchmarking brings together multiple data sets to make comparisons between various options. When you want to ensure an effective decision-making process, it can be beneficial to learn about benchmarking. In such circumstances, the Violinplot is not only an excellent statistical tool but also a beautiful graphic design capable of integrating the data with the type of probability distribution specified (Figure 11).
4 CONCLUSION

The Violinplot is a visual representation that describes the distribution of numerical data using density curves. Each curve’s width represents the approximate frequency of data points in the region. By incorporating a smoothed probability density function (KDE) instead of a histogram, the Violinplot combines the strengths of a boxplot and a histogram while avoiding the subjectivity of binning. So, it’s possible to include the minimum, first quartile, median, third quartile, and maximum, that are represented in the same manner as a traditional boxplot. It allows for a quick graphical examination and comparison of one or more datasets.

Also, the Violinplot has demonstrated its effectiveness as a statistical analysis tool for radioecological data because it offers advantages over isolated boxplots and histograms, providing valuable insights and assisting in the identification of anomalous values.

So, it’s important to note that the utility of the Violinplot extends beyond radioecology. It can be applied to various fields of knowledge that require robust numerical analysis in statistical planning for environmental management.

In summary, the Violinplot is a powerful graphical tool that facilitates data analysis, assists in anomaly detection, and contributes to the execution and effectiveness
of statistical planning in environmental management, not limited to radioecology but applicable across disciplines.
REFERENCES


AMARAL, Romilton dos Santos. Determinação de urânio na fosforita por meio de medidas radiométricas e análise por ativação. Tese de Mestrado-PRONUCLEAR, nº 41, Dissertação apresentada ao Departamento de Energia Nuclear da Universidade Federal de Pernambuco, 1987.


