CHAOS
Theory and Applications
in Applied Sciences and Engineering

e-ISSN: 2687-4539
RESEARCH ARTICLE
Vol.6 / No.4 / 2024 / pp.264-272
https://doi.org/10.51537/chaos.1486869

# Tax Audit in Turkiye: Simulation and Estimations Based on Kernel and Weight Functions

**Mehmet Niyazi Çankaya** *,1 **and Murat Aydın** α,2

*Faculty of Applied Sciences, Department of International Trading and Finance, Uşak University, Uşak, Turkiye, αFaculty of Applied Sciences, Department of Accounting Finance and Management, Uşak University, Uşak, Turkiye.

## ABSTRACT

This research examines the use of kernel estimation and `FindDistribution` methods in `Mathematica` software to analyze the ratio of taxpayer audits to total taxpayers, focusing on two large populations: one with approximately 80,000 audits per 100,000 taxpayers and the other with 4.5 million audits per 6 million taxpayers. Comparing the maximum statistics, the study shows that a larger number of taxpayers leads to more audits. The dataset also includes a weighted average for audits and taxpayers with a maximum of around 75,000 and 4 million respectively. These numerical values have been determined using the simulation carried out after modeling the real data sets of the total number of taxpayers and their audits from the years 2012 to 2023. These results show that different taxpayer populations require the targeted audit strategies and highlight the importance of the statistical models with corresponding estimation method to better understand complex distributions and improve tax audit processes.

## KEYWORDS

Inference
Non-parametric models
Robust statistics
Simulation
Taxpayers

## INTRODUCTION

Kernel density estimation (KDE) has been widely used in various fields, including income distribution analysis (Papatheodorou *et al.* 2004), poverty assessment (Minoiu and Reddy 2008), and population variance estimation (Hanif and Shahzad 2019). However, its application to grouped data has been found to introduce biases in poverty estimates (Minoiu and Reddy 2008). To address this, a method that combines auxiliary information with a kernel estimate has been proposed (Kuk 1993). Furthermore, a bipartite recursive algorithm based on KDE has been developed for measuring the scale of a given income population (Chen and Wang 2011).

KDE stands as a versatile tool widely deployed across diverse fields, ranging from income distribution analysis to poverty assessment and population variance estimation. Papatheodorou *et al.* (2004) underscores its efficacy in unveiling nuanced disparities within income distributions across different European countries, shedding light on the ramifications of income polarization and concentration. However, Minoiu and Reddy (2008) brings attention to the inherent biases introduced when KDE is applied to grouped data, particularly in poverty estimation, urging for caution in parameter selection. Addressing this concern, Kuk (1993) proposes a method amalgamating auxiliary information with kernel estimation to mitigate such biases, while Chen and Wang (2011) devises a bipartite recursive algorithm grounded in KDE for gauging the scale of specific income populations. This confluence of research highlights the promise KDE holds in estimating taxpayer numbers and scrutinizing taxpayer audit. Nevertheless, the discourse underscores the imperative of meticulous consideration of data characteristics and parameter choices to ensure robust and reliable estimations. A merge between a parametric model used for distribution of error term in the polynomial regression model and the polynomial movement on the data set as time series is studied by (Çankaya and Aydın 2024).

In Turkiye, the tax audit process is managed by the Presidency of the Tax Audit Board under the Ministry of Treasury and Finance. An important element of tax audit is expressed by the term "tax inspection" as it is understood in the activity reports of the Presidency. Article 134 of the Tax Procedure Act states that the main objective of tax inspection is to investigate and ensure the correctness of tax payments. Accordingly, tax inspectors check whether taxpayers have fulfilled their tax obligations in accordance with the legislation and whether they have correctly determined the

**Manuscript received:** 20 May 2024,
**Revised:** 8 October 2024,
**Accepted:** 8 November 2024.

[1]mehmet.cankaya@usak.edu.tr
[2]murat.aydin@usak.edu.tr (**Corresponding author**).

actual tax base. Tax inspection is not limited to the detection of tax evasion, but also includes the purpose of informing taxpayers of their tax obligations and verifying the elements of their tax returns.

A range of studies have explored the distribution of taxpayers and taxpayer audit. Chamberlain and Prante (2007); Piketty *et al.* (2018); Serikova *et al.* (2020) both highlight the progressive nature of the U.S. tax system, with the former emphasizing the impact of government spending on this progressive. Johns and Slemrod (2010); Davidson and Duclos (1997) delve into the distributional consequences of income tax noncompliance and the statistical inference for measuring the incidence of taxes and transfers, respectively. Ruggles and O'Higgins (1981); Piketty *et al.* (2017) both examine the distributive impact of government expenditures, with the latter focusing on the distribution of national income. Chotikapanich (2008); Perese (2015) provide methodological approaches for estimating income distributions and analyze the distribution of household income and federal taxes, respectively. Tax audit outcomes can lead to considerable adjustments in how companies recognize and value tax benefits, ultimately affecting their financial statements and tax strategies (Brushwood *et al.* 2018; Cowx and Vernon 2023).

The organization of the paper is given in the following order: The first section is for the introductory knowledges from literature. The second section gives real data. The method and objective are represented by third section. The forthcoming sections provide the statistical evaluations and their numerical results. The last section is divided into section for the conclusion.

## DATA ON TAXPAYERS IN TURKIYE

Within the scope of the study, the data were obtained from the annual reports published on the official website of the Presidency of the Tax Audit Board. The reports covering the period between 2012 and 2023 contain informations which are total number of taxpayers and taxpayer audit.

**Table 1** Taxpayers and their Audit by Years (VDK 2023)

| Year | Total number of taxpayers | Taxpayer audit |
|---|---|---|
| 2012 | 2,422,975 | 46,845 |
| 2013 | 2,460,281 | 71,352 |
| 2014 | 2,472,658 | 55,284 |
| 2015 | 2,527,084 | 58,676 |
| 2016 | 2,541,016 | 49,817 |
| 2017 | 2,636,370 | 44,182 |
| 2018 | 2,727,208 | 44,376 |
| 2019 | 2,813,452 | 40,763 |
| 2020 | 3,004,329 | 47,597 |
| 2021 | 3,221,084 | 54,065 |
| 2022 | 3,443,964 | 77,610 |
| 2023 | 3,621,478 | 60,242 |

Table 1 presents the total number of taxpayers and the number of taxpayers audited for certain years. In general, the table shows that the number of taxpayers increases each year and that the number of taxpayer audit generally shows an increasing trend. This may imply that tax controls cannot be applied to all taxpayers due to the limited resources of the tax administration or other priorities. In particular, there can be a significant decrease in the numbers of audit in years 2013-2014, 2015-2016, 2018-2019 and 2022-2023. There is an increase in trending of taxpayer audit from years 2012-2013 and 2021-2022. These numbers may indicate that the tax administration's strategies or resources have changed or that it has turned to other ways of administrative process of tax management system.

The continuous increase in the total number of taxpayers may reflect the expansion of the tax system or the fact that more people are becoming taxpayers as the economy grows. However, low numbers of audit may indicate that tax compliance is not at the desired level or that the tax administration is not using its audit resources effectively. The next section provides the modeling of data sets in Table 1 and the artificial data sets generated from the estimated functions determined by modeling.

## METHOD AND OBJECTIVE

### Kernel Estimation Method

One of the key advantages of the kernel mixture distribution is its ability to fit complex and multimodal data distributions. Unlike traditional parametric models, which make assumptions about the underlying data distribution, the kernel mixture distribution is non-parametric, meaning that it can adapt to the shape and structure of the data without imposing strict constraints, that is, it is data-adaptive and so the smoothness property will guarantee to fit data set well. This flexibility makes it particularly suitable for analyzing data sets with different patterns and characteristics.

It also provides a versatile framework for a variety of statistical tasks, including kernel mixture distribution, density estimation, clustering, and anomaly detection. By adjusting parameters such as the bandwidth of the kernels and the number of components in the mixture, analysts can fine-tune the distribution to capture different aspects of the data and achieve the desired level of granularity (Wand and Jones 1994).

In cases where the sample size is small, it is considered prudent to use applied techniques and artificial datasets to avoid bad effects. It should be noted, however, that for the smoothing technique, the alternative smoothing function can also be tried to obtain a possibly more accurate modeling; the figures will be close to the results already obtained, since the number of replications is increased to generate the artificial data. In addition, the parametric model proposed provides a comparison between the parametric function and the smoothing function. The results of the proposed distribution show that the smooth function is able to perform an accurate fit compared to the trimodal normal distribution as a parametric model (Vila *et al.* 2024b).

Various techniques and measures are used in data analysis to overcome the difficulties of working with small sample sizes. The use of artificial datasets can be very useful in such cases, allowing the creation of additional data points to supplement the original sample. This can help to correct for irregularities or gaps in the data and increase the robustness of the analysis. In addition, the use of smoothing techniques can improve the modelling process by reducing noise and highlighting important patterns in the data. Exploring alternative smoothing functions can further improve the modelling process and potentially lead to more accurate results. The comparison between parametric models and smooth functions, as proposed by Vila *et al.* (2024b), sheds light on the effectiveness

of different modelling approaches. The smooth function appears to outperform the parametric model, especially when compared to the tri-modal normal distribution.

Kernel estimation as a non-parametric method is a statistical method used to estimate the probability density function, $f(x)$, of a random variable based on a sample of data. It uses kernel functions to set a smoothness in order to fit the data and create estimates for the parameters such as location, scale, etc. of the underlying distribution.

Given a data sample, $x_1, x_2, \ldots, x_n$, the kernel density estimate of the function $f(x)$ at a point $x$ is calculated as:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

- $f(x)$ is the estimated density at point $x$ for the function $f(x)$.
- $n$ is the number of data points in the sample.
- $h$ is the bandwidth parameter, which controls the width of the kernel.
- $K(\cdot)$ is the kernel function, a smooth, symmetric function centered around zero. Common choices for the kernel function include the Gaussian, Epanechnikov, and uniform kernels. The choice of kernel affects the shape of the estimated density (Wand and Jones 1994; Wolfram 2003).

In `Mathematica`, the `KernelMixtureDistribution` function is used to create a kernel mixture model for density estimation. By default, `KernelMixtureDistribution` uses a Gaussian (normal) kernel for the estimation (Wand and Jones 1994; Wolfram 2003).

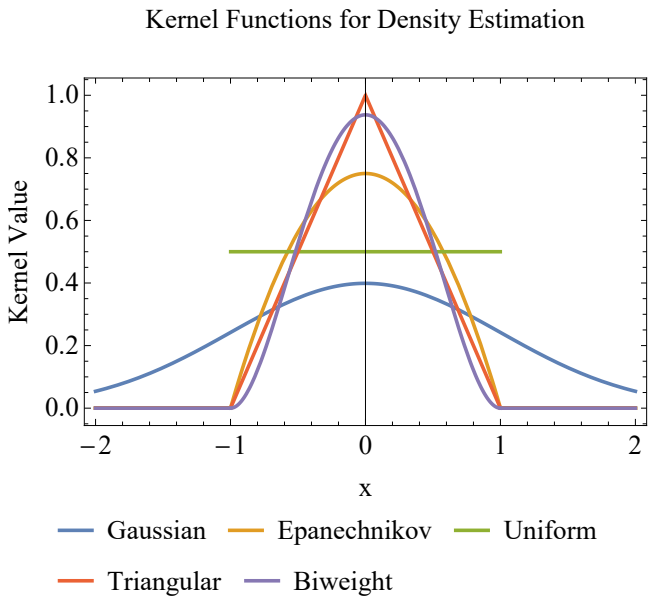



**Figure 1** Kernel functions

Since the kernel functions used in `Mathematica` are very close to each other, the mixture form of the normal distribution was preferred due to the number of pages and the complexity of the results in the Figures at this paper. On the other hand, Gaussian (normal) kernel have an ability to fit the data where take the values at the interval $[-2, 2]$ when compared with other kernel functions such as Epanechnikov, Triangular, Biweight, etc. In addition, the function `FindDistribution` is also used to be able to perform a precise evaluation while getting the weights from differences of cumulative distribution function (see codes in Appendix).

### Robust Estimations for Location and Scale Parameters

Robust statistics and kernel estimation share the goal of dealing with non-standard data distributions and mitigating the effects of outliers. Robust statistics focuses on developing methods that are resistant to outliers and deviations from standard assumptions. Robust estimators, such as the median or trimmed mean, are less affected by extreme values than traditional estimators such as the mean (Maronna *et al.* 2019).

Kernel estimation, often used in non-parametric density estimation, involves smoothing the data using a kernel function to estimate the underlying probability density function. This approach provides flexibility in modelling complex data distributions without assuming a specific parametric form. However, kernel estimation can be sensitive to outliers, leading to biased estimates, especially in regions of sparse data (Wand and Jones 1994).

The connection between robust statistics and kernel estimation lies in their complementary roles in dealing with challenging data scenarios. While kernel estimation provides flexibility and adaptability in modelling diverse data distributions, robust statistical techniques provide stability and resistance to outliers. By combining the principles of robust statistics with kernel estimation, researchers can develop methods that are both flexible and robust, enabling more reliable inference and analysis in the presence of non-standard data distributions and outliers.

The log-likelihood form of location and scale family is used to obtain the weighted mean and the weighted variance. The mean is given by

$$\text{weightedMean}(w) = \frac{\sum_{i=1}^{n} (\text{Sort}(\text{ND}(w))_i \cdot \text{weights}(w)_i)}{\sum_{i=1}^{n} \text{weights}(w)_i} \tag{2}$$

- $\text{ND}(w)$: Function that returns the numerical data derived from $w$.
- $\text{weights}(w)$: Function that returns the weights corresponding to the elements of $\text{ND}(w)$.
- $\text{Sort}(\text{ND}(w))$: Sorted version of $\text{ND}(w)$ in ascending order.

The square root of the weighted variance is defined as weighted standard deviation given by the following form:

$$\text{weightedStD}(w) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \text{weights}(w)_i \, (\text{Sort}(\text{ND}(w))_i - \text{weightedMean}(w))^2} \tag{3}$$

- $\text{ND}(w)$: Function that returns the numerical data derived from $w$.
- $\text{weights}(w)$: Function that returns the weights corresponding to the elements of $\text{ND}(w)$.
- $\text{Sort}(\text{ND}(w))$: Sorted version of $\text{ND}(w)$ in ascending order.
- $\text{weightedMean}(w)$: Weighted mean of $w$, as defined previously.
- $n$: Length of $\text{ND}(w)$.

The theory of robust statistics is based on weights from the assumed or the chosen function (Maronna *et al.* 2019). In our case, the weights come from two separate functions. One is the kernel estimator and the other is the `FindDistribution` function included with `Mathematica` software `12.0.0.0`. The `FindDistribution` is a powerful tool for fitting a probabilistic model to a given dataset. Implemented in version 12.0.0.0, this function automatically identifies the most appropriate distribution from a set of candidate distributions using the statistical goodness-of-fit tests. It allows users to quickly determine the underlying statistical properties of their data, simplifying the statistical modelling (Wolfram 2003).

**Kernel Mixture and Find Distributions in `Mathematica`**

In `Mathematica`, the functions `KernelMixtureDistribution` and `FindDistribution` can be used to estimate a distribution based on a data sample and a chosen kernel function.

The syntaxes for the function are:

```
KernelMixtureDistribution[data,Automatic,"SemiCircle"]
FindDistribution[data]
```

- 'data' is the input data sample.
- 'Automatic' allows `Mathematica` to automatically select an appropriate bandwidth.
- '"SemiCircle"' specifies the semi-circle kernel function, which may be useful for specific types of data.
- The function `FindDistribution` returns the name or symbolic representation of the distribution that best fits the data. It can handle a wide range of distribution families, including but not limited to normal, uniform, exponential, gamma, beta, and many others.

The 'KernelMixtureDistribution' function returns a nonparametric distribution that can be used for further analysis. Consideration of factors such as numerical optimization and manufacturing process is crucial to ensure the validity and reliability of the results. The numerical values obtained from Figures 2-13 are likely to represent the results of these optimization processes and can give an idea of the performance of the modeling techniques automatically performed by the functions in `Mathematica 12.0.0.0`. By following these steps, you will have a comprehensive understanding of the characteristics of your synthetic datasets and be able to analyze their statistical properties effectively (Wolfram 2003).

***Algorithmic Schema in Order***

1 Transfer data set into the case where the unit interval is set:
If your original data is not in the unit interval (i.e., the range [0, 1]), you'll need to scale it to fit within this range. The number of taxpayer audit is proportioned to the total number of taxpayers, thus obtaining data will be in the unit interval. If the randomly generated ratio values from estimated density, $f(x)$, are multiplied by the total number of taxpayers, then the number of taxpayer audit is obtained. If the number of taxpayers is divided by the randomly generated ratio values from estimated density, $f(x)$, the total number of taxpayers is obtained.

2 Model the unit interval data set:
Once your data is in the unit interval, you can model it using a kernel estimation method. Since using `Mathematica`, the function, 'SK=KernelMixtureDistribution[x,Automatic,"SemiCircle"]', could be used to create a smooth kernel density estimate of your data. Using the model generated in the previous step, generate a synthetic dataset with the same characteristics as your original data.

3 Generate artificial data set with sample size $n = 12$:
'RandomVariate[SK,n=12]'; Use the 'RandomVariate' function in `Mathematica` to generate random samples from your estimated density, $f(x)$.

4 Once you have generated one synthetic dataset, replicate this process 10,000 times by using 'SK'. For each iteration, generate a new synthetic dataset. Multiplication by total taxpayers with generated ratio values gives the taxpayer audit and division of taxpayer audit with ratio values gives the total number of taxpayers (see also step 1).

5 Provide statistics: Once you have your 10,000 synthetic datasets, calculate statistics such as the first moment (mean), scale estimate (standard deviation), minimum, maximum, 1th, 25th, 75th and 99th percentiles for each dataset. You can also use functions like first moment and scale estimated from the estimated density, 'Min', 'Max', 'Quantiles (1%, 25%, 75% and 99%)', etc., in `Mathematica` to evaluate these statistics summarizing the general representation of the generated data set from estimated density.

When the smooth function from kernel method is used, the corresponding statistics such as first moment, scale estimate, etc. are calculated. Therefore, these statistics are more accurate due to the precise fitting performed with the smooth function. Since the artificial data set is replicated with a sample size of $n = 12$, the minimum and maximum values are selected for each set. The same process is done for the data sets at the 1th, 25th, 75th and 99th percentiles so that we can observe the behavior of the data set at these percentiles as probability values indicating what the values generated for these percentiles are. In other words, we can see the overall picture of the data generated for these values. Note that the computational and methodological processes are also used by references (Vila *et al.* 2024b; Özen and Çankaya 2023; Aydın and Çankaya 2024).

## STATISTICAL EVALUATIONS

The kernel smoothing method in `Mathematica` is capable of performing a fitting on the data set. Further, since the assumed nonparametric density in this software is used to generate artificial data sets, we have a well-defined computational schema for evaluating various statistics and properties of your synthetic datasets. To summarize:

**1** Empirical First Moment and Scale Estimate from Data Generated SK:
These statistics are computed using built-in functions in `Mathematica` ($'Moment[data,1]'$ and $'Sqrt[Moment[data,2] - Moment[data,1]^2]'$) and are considered more accurate due to the precise fitting performed by the smooth function.

**2** Minimum and Maximum Values:
For each replicated artificial dataset, the minimum and maximum values are chosen. This provides insight into the range of values generated by the model.

**3** Quantiles at 1th, 25th, 75th and 99th Percentiles:
Similarly, for each replicated artificial dataset, the values at 1th, 25th, 75th and 99th percentiles are determined. This provides a picture of the overall distribution of the generated data and allows the behavior of the dataset around these quantities to be observed. It should also be noted that this calculation scheme is used with references (Vila *et al.* 2024b; Özen and Çankaya 2023; Aydın and Çankaya 2024), which shows its validity and suitability in practice. By following this scheme, you can effectively analyze the characteristics and behavior of your synthetic datasets and help your optimization process.

**4** Weighted Statistics for Location Scale Parameters:
The weighted mean and the weighted standard deviation based on the differences of cumulative distribution function from kernel smooth (KS) and FindDistribution (FD) are calculated.

The statistical evaluations are detailed quantitatively in the following section 'Numerical Results'. This section presents the empirical findings, providing a comprehensive analysis of the data and their implications for the study.

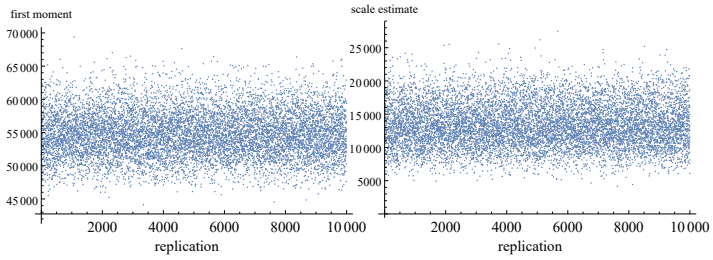
## NUMERICAL RESULTS OF STATISTICAL EVALUATIONS

The generation of accurate simulated data can have a significant impact on various stakeholders, including policymakers, researchers, and industry professionals. These stakeholders rely on the quality and precision of the data to make fully informed decisions. While the simulated data are confidential, it is important to thoroughly and fully document the methods and approaches used to generate them. This documentation allows for reproducibility and validation of the outputs of simulation by other researchers or analysts in the field.

Each of Figures 2-13 shows summary statistics from the artificially generated data set for the sample size $n = 12$, replicated 10,000 times. There are two types of Figures. One of them represents the statistical values for taxpayer audit given by Figs. 2 - 5. The second one represents the statistical values for taxpayer given by Figs. 6 - 9. Further, note that we focus on the maximum values of the generated data set for the sake of the fact that the future probable prediction can also be evaluated and suggested as well. In such case, the maximum, 1%, 25%, 75% and 99% as order statistics for taxpayer audit and taxpayer are an open issue which should be studied intensively. We prefer to omit the topic about order statistics and Figures 2-13 give the general appearance, as mentioned above. Consequently, the numerical results can be the values shown in Figures 2-13, taking into account the numerical optimization and generation procedure from the estimated density, $f(x)$ estimated by using two estimation method which are non-parametric being kernel smooth and parametric being FindDistribution. The simulated data are therefore considered confidential as they are the best possible match to the observed data.

Figures 2a-2b and 6a-6b show the empirical first moment on average from the measure of central tendency and the scale estimate from the measure of dispersion. Figures 3 and 7 represent the minimum and maximum values of the data at a sample size of $n = 12$. Figures 4-5 and 8-9 represent the simulated data for $n = 12$ at 1%, 25%, 75% and 99% cut-offs, respectively.

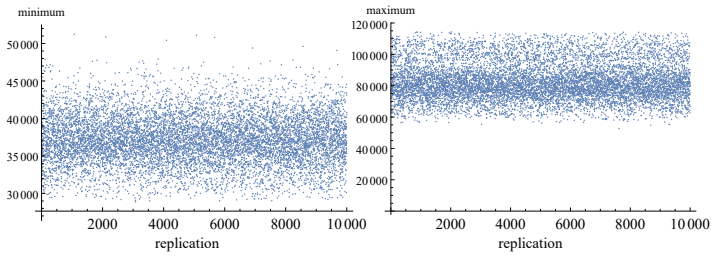
When comparing Figures 3b and 7b for maximum of the artificial numbers from taxpayers audit and taxpayers, respectively, the more taxpayers lead to have detection of the more taxpayer audit, as expected.




**(a)** Empirical first moment **(b)** Scale estimate

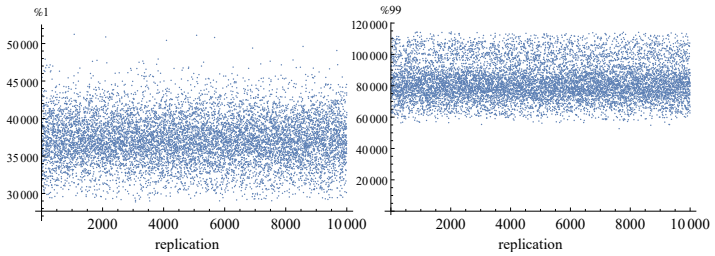
**Figure 2** The simulated data for statistics of the taxpayer audit within years 2012-2023

Figures 10-13 show the robust estimates replicated at 10,000 times for location and scale parameters. When Figure 10a is compared with Figure 11a, the results show that the values of weighted mean from FD tends to take lower values, which shows that the chosen function for fitting data set affects the results we will get, because the function chosen plays role in determining the weights for the robust estimation. In addition, the scale estimate given



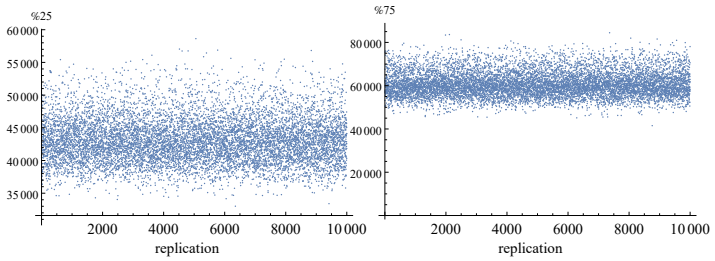

**(a)** Minimum values of artificial data **(b)** Maximum values of artificial data

**Figure 3** The simulated data for minimum and maximum of the taxpayer audit within years 2012-2023



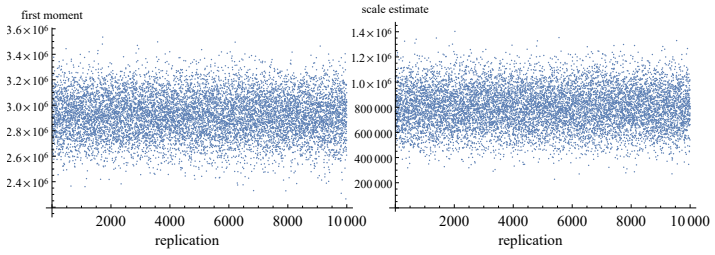

**(a)** Quantile at 1% of artificial data **(b)** Quantile at 99% of artificial data

**Figure 4** The simulated data for quartiles at 1% & 99% of the taxpayer audit within years 2012-2023




**(a)** Quantile at 25% of artificial data **(b)** Quantile at 75% of artificial data

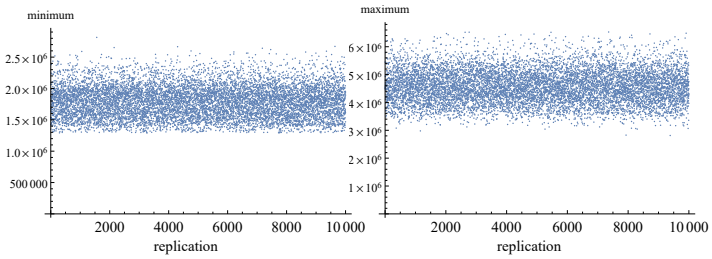
**Figure 5** The simulated data for quartiles at 25% & 75% of the taxpayer audit within years 2012-2023




**(a)** Empirical first moment **(b)** Scale estimate

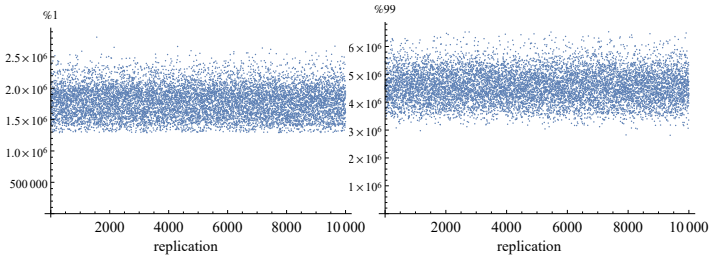
**Figure 6** The simulated data for statistics of the taxpayers within years 2012-2023

by Figure 2b have values which are bigger than that of values in Figure 10b. The same situation for taxpayers at Figures 6b and 12b is observed. Figures 10-11 represent the case where the numbers of taxpayer audit are around. Figures 12-13 represent the case



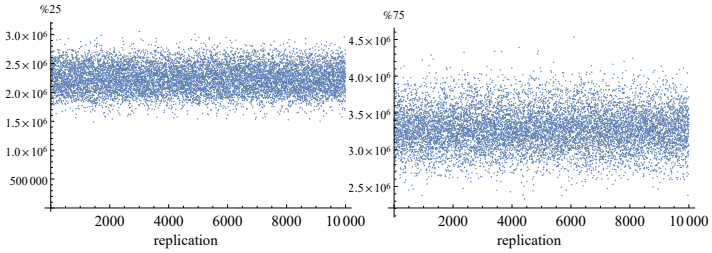

**(a)** Minimum values of artificial data **(b)** Maximum values of artificial data

**Figure 7** The simulated data for minimum and maximum of the taxpayers within years 2012-2023



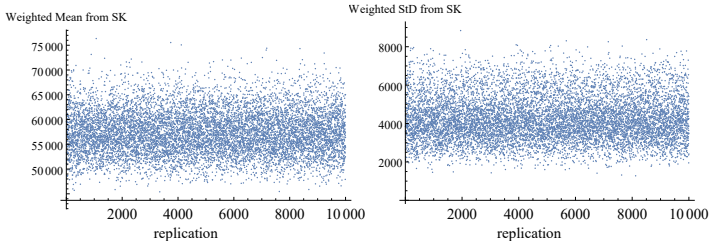

**(a)** Quantile at 1% of artificial data **(b)** Quantile at 99% of artificial data

**Figure 8** The simulated data for quartiles at 1% & 99% of the taxpayers within years 2012-2023




**(a)** Quantile at 25% of artificial data **(b)** Quantile at 75% of artificial data

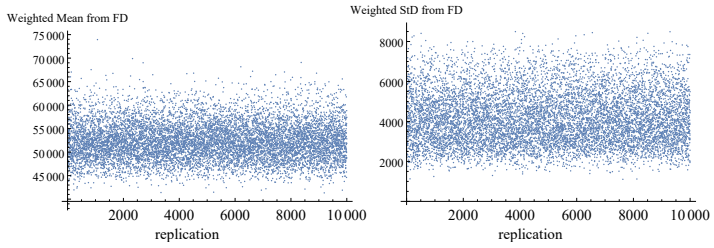
**Figure 9** The simulated data for quartiles at 25% & 75% of the taxpayers within years 2012-2023




**(a)** Weighted mean from Smooth Kernel(SK) **(b)** Weighted standard deviation from Smooth Kernel

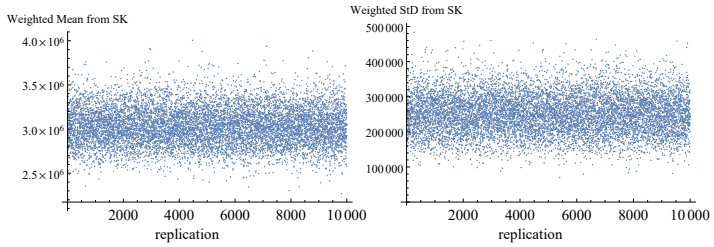
**Figure 10** The weighted forms of location and scale estimates from Smooth Kernel, $n = 12$ for the taxpayer audit

where the numbers of taxpayer are around.



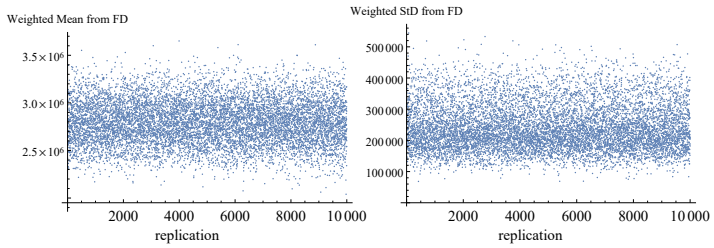

**(a)** Weighted mean from Find Distribution (FD) **(b)** Weighted standard deviation from Find Distribution

**Figure 11** The weighted forms of location and scale estimates from Find Distribution, $n = 12$ for the taxpayer audit




**(a)** Weighted mean from Smooth Kernel **(b)** Weighted standard deviation from Smooth Kernel

**Figure 12** The weighted forms of location and scale estimates from Smooth Kernel, $n = 12$ for the taxpayers




**(a)** Weighted mean from Find Distribution **(b)** Weighted standard deviation from Find Distribution

**Figure 13** The weighted forms of location and scale estimates from Find Distribution, $n = 12$ for the taxpayers

### Number of Taxpayers at per square Kilometer of Turkiye

To find the number of taxpayers per square kilometre, you must first find the total number of taxpayers in a given area, and then divide that number by the total area of that area, measured in square kilometre. This calculation will give you the density of taxpayers, which indicates how many taxpayers live or are based in each square kilometre of the area in question.

Considering the maximum value in the simulated data from kernel estimation method, there are approximately 6 million taxpayers from Figure 7b. Turkiye, with a population of 85.8 million, is in the taxpayer's role at a rate of $6/85.8 = 0.06993$.

Since there are 110 people per square kilometre in Turkey, $110 \cdot 0.06993 = 7.69$, approximately 8 out of 110 people per square kilometre will be identified as taxpayers if they are evenly distributed across the regions of Turkey (Wolfram 2003). This measure is essential for understanding the distribution of taxpayers across the country, which can help in effective policy making, resource allocation and economic planning.

## CONCLUSION

The study has highlighted the importance of using sophisticated statistical methods to accurately model these complex, multimodal data distributions which can be modeled by using the kernel estimation methods which provide robust and versatile framework for statistical analysis. They can handle complex/multimodal data distributions such as the ratio values between the taxpayer audit and its total numbers. Firstly, the kernel estimation method has been used. After that, the `FindDistribution` function included with `Mathematica` software `12.0.0.0` is used to model the generated data artificially from `SK=KernelMixtureDistribution` with `RandomVariate[SK, n=12]`.

The values of ratio give an advantage for us to fit the data set well. In addition, the total taxpayers and the taxpayer audit can be calculated from the values of ratio. When the results for the maximum number of taxable persons and taxpayers are compared, the more taxable persons lead to the detection of the more taxpayer audit, as expected. According to the maximum statistic of the simulated data, the total numbers of taxpayers can go up to 6 million, which can occur in the near future. The taxpayer audit will be around 120,000 from maximum statistics. The simulation results for the taxpayer audit have shown that there are two blocks for the numbers which are 80,000 and 100,000. The values around 100,000 are few when compared with that of 80,000, which shows that the audit on the taxpayers is commented as two populations and some precautions for tax audit can be necessary when the maximum statistics are taken into account.

In the same way, when observing the number of taxpayers, there can be two populations which are 4.5 million and 6 million. Additional statistics, including the weighted means for taxpayer audits and overall taxpayer population, which are approximately 75,000 and 4 million respectively, are also provided to summarize general characteristics of the data set. As a result, more taxpayers should be surveyed. Further, by improving our understanding of these populations; policymakers and tax authorities can implement more effective policies to optimize tax audit processes and ensure fair tax compliance among different taxpayer groups. Our ability for modeling, estimating and understanding the number of taxpayers and its audit form with precision and confidence intervals will continue to improve with the continued research and improvements in this area.

## APPENDIX

### The `Mathematica` codes for computation and statistical evaluations

`Mathematica`, developed by Wolfram Research, is a comprehensive computational software system widely used in various fields of science, engineering, mathematics and computing. It features a high-level programming language, powerful computational capabilities, and a wide range of built-in functions, making it an indispensable tool for research, education, and industrial applications.

The codes were used to model the proportional data using the kernel mixture distribution in `Mathematica` 12.0.0.0 software.

```
For[w=1, w <= rep, w++,
  (* Fitting via Smooth Kernel Method *)
SK[w]=KernelMixtureDistribution[x, Automatic, "SemiCircle"];

(* Vector for scaling the generated random numbers *)
  vec={data};

(* Generate random numbers based on the kernel
distribution and scale them by vec *)
  ND[w]=RandomVariate[SK[w], 12] * vec;

(* Calculate the CDF of the kernel distribution for
sorted ND[w] *)
 CDFSK[w]=CDF[KernelMixtureDistribution[ND[w]],Sort[ND[w]]];

(* Calculate the CDF of a fitted distribution for
  sorted ND[w] *)
 CDFFD[w]=CDF[FindDistribution[ND[w]],Sort[ND[w]]];

(* Calculate weights based on the differences in the CDF
  for the kernel distribution *)
  weights1SK[w]=Differences[CDFSK[w]];

(* Calculate the remaining weight to ensure the weights
  sum to 1 *)
weights2SK[w]=1 - Total[weights1SK[w]];

(* Combine the weights and ensure they sum up to 1 *)
  weightsSK[w]=Join[weights1SK[w], {weights2SK[w]}];

(* Calculate weights based on the differences in the CDF
  for the fitted distribution *)
  weights1FD[w]=Differences[CDFFD[w]];

(* Calculate the remaining weight to ensure the weights
  sum to 1 *)
  weights2FD[w]=1 - Total[weights1FD[w]];

(* Combine the weights and ensure they
  sum up to 1 *)
  weightsFD[w]=Join[weights1FD[w], {weights2FD[w]}];

(* Calculate various statistics for ND[w] *)
  sta1[w]:=Moment[ND[w], 1];  (* First moment (mean) *)
  ta1=Table[sta1[w], {w, rep}];

  sta2[w]:=Mean[ND[w]];  (* Mean *)
  ta2=Table[sta2[w], {w, rep}];

  sta3[w]:=Median[ND[w]];  (* Median *)
  ta3=Table[sta3[w], {w, rep}];

 (* Standard deviation based on moments *)
  sta4[w]:=Sqrt[Moment[ND[w], 2] - Moment[ND[w], 1]^2];
  ta4=Table[sta4[w], {w, rep}];

  (* Standard deviation *)
  sta5[w]:=StandardDeviation[ND[w]];
  ta5=Table[sta5[w], {w, rep}];

  (* Median absolute deviation from the median *)
  sta6[w]:=Median[Abs[ND[w] - Median[ND[w]]]];
  ta6=Table[sta6[w], {w, rep}];

   (* Median absolute deviation from the mean *)
  sta61[w]:=Median[Abs[ND[w] - sta1[w]]];
  ta61=Table[sta61[w], {w, rep}];

  (* Mean absolute deviation from the mean *)
```

```
  sta62[w]:=Mean[Abs[ND[w] - sta1[w]]];
  ta62=Table[sta62[w], {w, rep}];

  (* Minimum value *)
  sta7[w]:=Min[ND[w]];
  ta7=Table[sta7[w], {w, rep}];

  (* Maximum value *)
  sta8[w]:=Max[ND[w]];
  ta8=Table[sta8[w], {w, rep}];

  (* Calculate various quantiles *)

  (* 1st percentile *)
  staquan1[w]:=Quantile[ND[w], 0.01];
  taga1=Table[staquan1[w], {w, rep}];

  (* 25th percentile *)
  staquan2[w]:=Quantile[ND[w], 0.25];
  taga2=Table[staquan2[w], {w, rep}];

  (* 50th percentile / median *)
  staquan3[w]:=Quantile[ND[w], 0.5];
  taga3=Table[staquan3[w], {w, rep}];

  (* 75th percentile *)
  staquan4[w]:=Quantile[ND[w], 0.75];
  taga4=Table[staquan4[w], {w, rep}];

  (* 99th percentile *)
  staquan5[w]:=Quantile[ND[w], 0.99];
  taga5=Table[staquan5[w], {w, rep}];

  (* Calculate the weighted mean based on
  the smooth kernel (SK) distribution weights *)
  weightedMeanSK[w]:=
    Total[Sort[ND[w]] * weightsSK[w]]
    /
    Total[weightsSK[w]];
  taWeMeSK=Table[weightedMeanSK[w], {w, rep}];

  (* Calculate the weighted standard deviation
  based on the smooth kernel distribution weights *)
  weightedStDSK[w]:=
Sqrt[Total[weightsSK[w]*(Sort[ND[w]]-weightedMeanSK[w])^2]
    /
    Length[ND[w]]];
  taWeSDSK=Table[weightedStDSK[w], {w, rep}];

  (* Calculate the weighted mean based
  on the fitted distribution (FD) weights *)
  weightedMeanFD[w]:=
    Total[Sort[ND[w]] * weightsFD[w]]
    /
    Total[weightsFD[w]];
  taWeMeFD=Table[weightedMeanFD[w], {w, rep}];

  (* Calculate the weighted standard deviation
  based on the fitted distribution weights *)
  weightedStDFD[w]:=
Sqrt[Total[weightsFD[w]*(Sort[ND[w]]-weightedMeanFD[w])^2]
    /
    Length[ND[w]]];
  taWeSDFD=Table[weightedStDFD[w], {w, rep}];
]
```


#### Acknowledgments

We appreciate the editorial board's and reviewers' valuable comments on the paper.


#### Availability of data and material

The real data set is included by the paper.

#### Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

#### Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

## LITERATURE CITED


Alleva, G. and A. E. Giommi, 2016 *Topics in Theoretical and Applied Statistics*. Springer.

Aydın, M. and M. N. Çankaya, 2024 Assessing the regulatory impact of the turkish competition authority on market dynamics: A statistical approach using kernel estimation and its simulation. Journal of Mehmet Akif Ersoy University Economics and Administrative Sciences Faculty **11**: 837–853.

Brushwood, J. D., D. M. Johnston, and S. J. Lusch, 2018 The effect of tax audit outcomes on the reporting and valuation of unrecognized tax benefits. Advances in Accounting **42**: 1–11.

Çankaya, M. N., 2020a M-estimations of shape and scale parameters by order statistics in least informative distributions on q-deformed logarithm. Journal of the Institute of Science and Technology **10**: 1984–1996.

Çankaya, M. N., 2020b On the robust estimations of location and scale parameters for least informative distributions. Turkish Journal of Science and Technology **15**: 71–78.

Çankaya, M. N. and O. Arslan, 2020 On the robustness properties for maximum likelihood estimators of parameters in exponential power and generalized t distributions. Communications in Statistics-Theory and Methods **49**: 607–630.

Çankaya, M. N. and M. Aydın, 2024 Future prediction for tax complaints to turkish ombudsman by models from polynomial regression and parametric distribution. Chaos Theory and Applications **6**: 63–72.

Çankaya, M. N. and J. Korbel, 2018 Least informative distributions in maximum q-log-likelihood estimation. Physica A: Statistical Mechanics and its Applications **509**: 140–150.

Çankaya, M. N. and R. Vila, 2023 Maximum log q likelihood estimation for parameters of weibull distribution and properties: Monte carlo simulation. Soft Computing **27**: 6903–6926.

Çankaya, M. N., A. Yalçınkaya, Ö. Altındağ, and O. Arslan, 2019 On the robustness of an epsilon skew extension for burr iii distribution on the real line. Computational Statistics **34**: 1247–1273.

Çankaya, M. N., 2021 Derivatives by ratio principle for q-sets on the time scale calculus. Fractals **29**: 2140040.

Chamberlain, A. and G. Prante, 2007 Who Pays Taxes and Who Receives Government Spending? An Analysis of Federal, State and Local Tax and Spending Distributions, 1991-2004. SSRN Electronic Journal .

Chen, Y. and H. Wang, 2011 Construction and application of bipartite recursive algorithm based on kernel density estimation: A new non-parametric method to measure the given income population scale. In *Statistics & Information Forum*, pp. 3–8.

Chotikapanich, D., 2008 *Modeling income distributions and Lorenz curves*, volume 5. Springer Science & Business Media.

Cowx, M. and M. Vernon, 2023 Accounting for tax uncertainty over time. Available at SSRN 4678373 .

Davidson, R. and J.-Y. Duclos, 1997 Statistical Inference for the Measurement of the Incidence of Taxes and Transfers. Econometrica **65**: 1453.

Hanif, M. and U. Shahzad, 2019 Estimation of population variance using kernel matrix. Journal of Statistics and Management Systems **22**: 563–586.

Johns, A. and J. Slemrod, 2010 The distribution of income tax noncompliance. National Tax Journal **63**: 397–418.

Kuk, A. Y., 1993 A kernel method for estimating finite population distribution functions using auxiliary information. Biometrika **80**: 385–392.

Maronna, R. A., R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, 2019 *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Minoiu, C. and S. Reddy, 2008 Kernel density estimation based on grouped data: The case of poverty assessment. IMF Working Papers **08**: 1.

Özen, E. and M. N. Çankaya, 2023 Estimation of the turkish stock investor numbers based on kernel method. In *Competitivitatea şi inovarea în economia cunoaşterii*, pp. 445–454.

Papatheodorou, C., P. Peristera, and A. Kostaki, 2004 Kernel density techniques as a tool for estimating and comparing income distributions: a cross european–country study. Journal of Income Distribution **13**: 2–2.

Perese, K., 2015 The distribution of household income and federal taxes, 2011. Current Politics and Economics of the United States, Canada and Mexico **17**: 695.

Piketty, T., E. Saez, and G. Zucman, 2017 Distributional National Accounts: Methods and Estimates for the United States*. The Quarterly Journal of Economics **133**: 553–609.

Piketty, T., E. Saez, and G. Zucman, 2018 Distributional national accounts: methods and estimates for the united states. The Quarterly Journal of Economics **133**: 553–609.

Ruggles, P. and M. O'Higgins, 1981 The distribution of public expenditure among households in the united states. Review of Income and Wealth **27**: 137–164.

Serikova, M., L. Sembiyeva, K. Balginova, G. Alina, A. Shakharova, *et al.*, 2020 Tax revenues estimation and forecast for state tax audit. Entrepreneurship and Sustainability Issues **7**: 2419–2435.

VDK, T., 2023 Vdk annual reports. https://en-vdk.hmb.gov.tr/annual-reports, [Online; accessed 10-May-2024].

Vila, R., L. Alfaia, A. F. Menezes, M. N. Çankaya, and M. Bourguignon, 2024a A model for bimodal rates and proportions. Journal of Applied Statistics **51**: 664–681.

Vila, R. and M. N. Çankaya, 2022 A bimodal weibull distribution: properties and inference. Journal of Applied Statistics **49**: 3044–3062.

Vila, R., V. Serra, M. N. Çankaya, and F. Quintino, 2024b A general class of trimodal distributions: properties and inference. Journal of Applied Statistics **51**: 1446–1469.

Wand, M. P. and M. C. Jones, 1994 *Kernel smoothing*. CRC press.

Wolfram, S., 2003 *The mathematica book*. Wolfram Research, Inc.

***How to cite this article:*** Cankaya, M.N., and Aydin M. Tax Audit in Turkiye: Simulation and Estimations Based on Kernel and Weight Functions *Chaos Theory and Applications*, 6(4), 264-272, 2024.