A Neutrosophic Statistic Method to Predict Tax Time Series in Ecuador

Lilia Esther Valencia Cruzaty1, Mariela Reyes Tomalá2, Carlos Manuel Castillo Gallo3

1 Universidad Estatal Península de Santa Elena, Ecuador. E-mail: lvalencia@upse.edu.ec;
2 Universidad Estatal Península de Santa Elena, Ecuador. E-mail: mreyes@upse.edu.ec;
3 Universidad Estatal Península de Santa Elena, Ecuador. E-mail: ccastillo@upse.edu.ec

Abstract. Prediction of tax collection behavior is an essential tool for social planning by the State of any country. The tax is the State’s mechanism for budget collection, which is necessary to accomplish public services that benefit the whole society. This paper firstly aims to propose a method of predicting time series where values can be given in form of intervals rather than numbers. This form permits to obtain more truthful results, but with a greater indeterminacy. Because statistical prediction methods are used, where data in form of intervals are included, we can classify this approach as a kind of Neutrosophic Statistics technique. Basically, the method converts a set of predicted numerical values into intervals. The second objective is to apply the method to predict the monthly income from taxes in Ecuador for the year 2019.

Keywords: time series, tax, prediction, neutrosophic statistic, autocorrelation, median average

1 Introduction

The tax is a kind of tribute regulated by the public law, see [1]. Taxes generally have the State as a creditor and aims to finance state expenditures. These are based on the principle that those who have more are those who must contribute more, in order to guarantee equity and social freedom.

People and companies must pay taxes on a mandatory basis, because this is the way to finance the operation of the State, otherwise it would collapse. This is the mechanism to guarantee the building of infrastructure such as roads, ports, airports, hydraulic and electrical installations, to provide public health, education, defense, social security services, as support for the unemployed, disability benefits or occupational accidents, the payment of pensions, among many other essential aspects of society.

Some kind of taxes are income taxes, capital gains and profits, social security contributions by workers, employers and freelancers, payroll taxes, taxes for the property, and goods and services taxes.

In this paper, the prediction of taxes in Ecuador is made for the year 2019. For this, we start from the monthly measurement of the sum of income taxes, consumption taxes and value-added taxes. Thus, we propose and apply a neutrosophic method of time series prediction, which differs from those appeared in [2-4]. Such a method uses a set of statistical techniques to obtain classical numeric values to form intervals that have as a limit the lowest and highest values of all the individual considered methods. So, predictions are based on intervals rather than numeric values.

The problem of predicting has a certain complexity; therefore there exist many methods to predict future values, see [5-12]. This quantity of tools is due to the fact that each of them is effective for a set of cases, while for others it is not. Moreover, the most elementary methods can guarantee better results than the more sophisticated ones. That is why the integration of the results of the methods can give more acceptable results than if each one is used separately.

Thus, we offer to increase the probability of obtaining a more truthful result, in exchange for increasing indeterminacy. This corresponds to the Theory of Neutrosophic Statistics, where models and parameters are applied over intervals instead of numerical values, see [13][1].

This method is applied to solve the tax prediction problem in Ecuador. This prediction is essential, because it allows social planning by the Ecuadorian State. It will be possible to estimate in advance the income that the State will have for taxes. This approach is based on the principle that the given information in form of intervals is useful for public administration in this way.

This paper is divided as follows; it begins with a section where the classic statistical concepts and methods of time series are explained, as well as the essential ideas of the theory of Neutrosophic Statistics. The proposed prediction method is next presented and applied to the tax prediction problem in Ecuador. We finish with the
section of conclusions.

2 Preliminaries

This section summarizes the theory of statistical methods of time series and neutrosophic statistic theory divided in two subsections.

2.1 Times Series

A time series is a gathering of observations made consecutively in time, see [14-17][2].

When observations are made continuously in time, time series is called continuous. Whereas, when the observations are taken only at specific moment of times it is called discrete, usually time variable are equally spaced. If time series can be predicted exactly it is called deterministic, however future values of most of time series can only be partially predicted by past values, and they are called stochastic.

Among the objectives to study time series, in this paper we select the prediction or forecasting, which is the calculus in advance of future values of the series since the past values.

Time series can have the following components:

a. **Seasonal effect** is manifested when time series fluctuates in short-term periods. In annual time series they correspond to changes in periods shorter than one year.

b. **Cyclic changes** occur when the time series oscillates with not fixed long-term periods. In annual time series these cycles correspond to oscillations in many years.

c. **Trend** is a long-run development. This component can be interpreted like the overall tendency of the series when oscillations are not considered.

d. A component that does not correspond to any of the previous ones, which is called residual.

Many fields of knowledge are used in time series forecasting, we select particularly a statistical approach.

A time series is called strictly stationary if the joint distribution of \(X_{t_1}, X_{t_2}, \ldots, X_{t_n}\) is the same as the joint distribution of \(X_{t_1+q}, X_{t_2+q}, \ldots, X_{t_n+q}\) for every \(t_1, t_2, \ldots, t_n, q\).

In the following we consider normal stationary processes, and we use the estimated coefficients and statistics of this kind of process. A normal process is such that the joint distribution of \(X_{t_1}, X_{t_2}, \ldots, X_{t_n}\) is a multivariate normal distribution.

**Definition 1:** Given \(\{X_t\}\) a discrete time series and \(q \in \mathbb{Z}_+\), elements of another time series \(\{Y_t\}\) is defined by Equation 1.

\[
Y_t = \frac{\sum_{r=0}^{q-1} X_{t+r}}{q} \tag{1}
\]

For \(t = 1, 2, \ldots, n-q+1, n\) is the number of elements in \(\{X_t\}\).

\(\{Y_t\}\) is called the time series of Moving Averages (MA).

The time series of MA is used to attenuate seasonal variations.

**Definition 2:** Given \(N\) observations \(X_1, X_2, \ldots, X_N\), on a discrete time series and \(q \in \mathbb{Z}_+\), Equation 2 is the Autocorrelation coefficient (AR coefficient) of \(X_1, X_2, \ldots, X_N\), with lag \(q\).

\[
r_q = \frac{\sum_{t=1}^{N-q}(X_t-\bar{X})(X_{t+q}-\bar{X})}{\sum_{t=1}^{N}(X_t-\bar{X})^2} \tag{2}
\]

Where \(\bar{X}\) is the mean of the time series.

In practice \(r_q\) is usually calculated since the autocovariance coefficients, \(\{c_q\}\) with Equation 3 with the autocovariance coefficient with lag \(q\).

\[
c_q = \frac{1}{N} \sum_{t=1}^{N-q}(X_t-\bar{X})(X_{t+q}-\bar{X}) \tag{3}
\]

Then, \(r_q = \frac{c_q}{c_0}\)

To estimate the autocovariance it is recommendable to take \(N > q+1\), \(N \geq 50\) and \(q \leq N/4\), see [17].

A correlogram is a graph that aids for interpreting the set of autocorrelation coefficients. It consists on a plot of lags \(q\) versus \(r_q\).

A completely random time series, for a large \(N\) satisfies \(r_q=0\) for every \(q\). This can be seen in the correlogram. Also, this kind of graph exhibits the presence of seasonal fluctuations in the original series when it has fluctuations as well.

The correlogram can contain the upper and lower bounds for autocorrelation with the significance level \(\alpha\), which is given by Equation 4:

\[
B = \pm z_{1-\alpha/2} s_r(r_q) \tag{4}
\]
Where $r_q$ is the estimated autocorrelation of lag $q$, $z_{1-\alpha/2}$ is the quantile of the normal distribution and $S(\cdot)$ is calculated as follows:

$$S(r_q) = \frac{1}{\sqrt{N}}$$  \hspace{1cm} (5)

$$S(r_q) = \frac{1 + 2 \frac{q}{N} \frac{1 - q^2}{1}}{N}$$  \hspace{1cm} (6), for $q>1$.

The trend is calculated by curve fitting of the time series observations, usually by means of a linear function and least square estimation.

MA and AR processes are used as models of time series. When they are combined or integrated they form the ARMA and ARIMA models, see [14, 17-18] for more details.

### 2.2 Neutrosophic Statistic

Neutrosophic Statistics is an extension of classical statistics where crisp numerical values are replaced by values in form of intervals, see [13][3]. This substitution can be applied to parameters, not only to random variables. The sample size can also be considered as indeterminate or inaccurate. In this theory the data can be ambiguous, vague, inaccurate, incomplete or indeterminate.

It is necessary to emphasize that there is a difference between the concepts of indeterminacy and randomness. Classical statistics deal with random variables, while in a neutrosophic framework they can also be indeterminate. From this starting point, new concepts are defined from the classical ones, they are the following:

- **Neutrosophic Descriptive Statistics** “is comprised of all techniques to summarize and describe the neutrosophic numerical data characteristics.”
- **Neutrosophic Inferential Statistics** “consists of methods that permit the generalization from a neutrosophic sampling to a population from which it was selected the sample.”
- **Neutrosophic Data** “is the data that contains some indeterminacy.”
- **Neutrosophic Frequency Distribution** “is a table displaying the categories, frequencies, and relative frequencies with some indeterminacy.”
- **Neutrosophic Statistical Graphs** “are graphs that have indeterminate (unclear, vague, ambiguous, unknown) data or curves.”

Other essential definitions are the following:

**Neutrosophic Survey Results** “are survey results that contain some indeterminacy. A Neutrosophic Population is a population not well determined at the level of membership (i.e. not sure if some individuals belong or do not belong to the population).”

A simple random neutrosophic sample of size $n$ from a classical or neutrosophic population is a sample of $n$ individuals such that at least one of them has some indeterminacy.

**Neutrosophic Random Numbers** can also be generated using, instead of only crisp numbers, a pool of sets.

A Neutrosophic Normal Distribution of a continuous variable $X$ is a classical normal distribution of $x$, but such that its mean $\mu$ or its standard deviation $\sigma$ (or variance $\sigma^2$), or both, are imprecise.

Other neutrosophic distributions are: neutrosophic standard normal distribution, neutrosophic bivariate normal distribution, neutrosophic uniform distribution, neutrosophic sampling distribution, neutrosophic geometric distribution, neutrosophic hypergeometric distribution, neutrosophic Poisson distribution, neutrosophic chi-squared distribution, neutrosophic exponential distribution, neutrosophic frequency distribution, neutrosophic Pareto distribution and neutrosophic t-distribution.

The Neutrosophic Least-Squares Lines that approximates the neutrosophic bivariate data $(x_1, \hat{y}_1)$, $(x_2, \hat{y}_2)$, …, $(x_n, \hat{y}_n)$ has the same formula as in classical statistics, i.e.,

$$\hat{y} = a + b\bar{x},$$

where the slope $b = \frac{\sum_{i=1}^{n} x_i (\bar{y} - \hat{y})}{\sum_{i=1}^{n} x_i^2}$ and the y-intercept $a = \bar{y} - b\bar{x}$, were $\bar{x}$ is the neutrosophic average of $x$, whereas $\bar{y}$ is the neutrosophic average of $y$.

The Neutrosophic Residuals are computed in the same way as in classical statistics:

$$y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n,$$

where $y_i$ are the real values of variable $y$, and $\hat{y}_i$ are respectively their predicted values.

It is worthy to explain the definition and operations with neutrosophic numbers:

The neutrosophic numbers has the form $a + bI$, where $a$ and $b$ are real numbers, and $I$ = indeterminate.

Given $N_1 = a_1 + b_1I$ and $N_2 = a_2 + b_2I$ two neutrosophic numbers, some operations between them are defined as follows:

- $N_1 + N_2 = a_1 + a_2 + (b_1 + b_2)I$ (Addition);
\[ N_1 - N_2 = a_1 - a_2 + (b_1 - b_2)I \] (Subtraction),
\[ N_1 \times N_2 = a_1a_2 + (a_1b_2 + b_1a_2 + b_1b_2)I \] (Product),
\[ \frac{N_1}{N_2} = \frac{a_1+b_1I}{a_2+b_2I} = \frac{a_2 + 2b_1-a_1b_2I}{a_2(2a_2+b_2)} \] (Division).

Additionally, given \( I_1 = [a_1, b_1] \) and \( I_2 = [a_2, b_2] \) we have the following operations between them (see [19]):
\[ I_1+I_2 = [a_1+a_2, b_1+b_2] \] (Addition);
\[ I_1-I_2 = [a_1-b_2, b_1-a_2] \] (Subtraction);
\[ I_1\cdot I_2 = \min\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}, \max\{a_1-b_1, a_1-b_2, a_2-b_1, a_2-b_2\}\] (Product),
\[ I_1/I_2 = I_1(1/I_2) = \{a/b: a \in I_1, b \in I_2\}, \text{ always that } 0 \notin I_2 \] (Division).

3 Ecuadorian Tax Forecasting

Usually forecasting is referred to a unique value for every future time obtained from a unique model of prediction. However, the forecasted value is not necessarily accurate, in view that there not exists an ideal method of prediction. A very sophisticated method can yield to not sufficiently accurate results, while a simple one could be more accurate or vice versa.

In this paper we propose a method where the forecasted value is statistically obtained like an interval rather than a unique value. This predicted interval value is used to predict other value of the same kind and so on. Such a method is applied to forecast the Ecuadorian tax during 2019. It is based on the single-valued statistical forecasting methods which define an interval-valued one. We describe it below:

1. To graphically represent the time series and visually establish its properties of seasonal effect, cyclic changes and trend.
2. If \( \{X_t\} \) for \( t = 1, 2, \ldots, n \) are the values of the time series. Apply the correlogram to statistically establish the seasonal effect.
3. Let \( M_1(\{X_t\}), M_2(\{X_t\}), \ldots, M_k(\{X_t\}) \) be \( k \) forecasting methods conveniently selected according to some criteria of accuracy for this type of time series.
   Observe \( X_{n+1} = M_i(\{X_t\}) \), for \( i = 1, 2, \ldots, k \); which are the forecasted values for every method at time \( t_{n+1} \).
4. Form the intervals \( I_{n+1} = [\min\{X_{n+1}\}, \max\{X_{n+1}\}] \). Now, we have new time series \( \{X_t\} \) for \( t = 1, 2, \ldots, n, n+1 \), where \( X_{n+1} = I_{n+1} \).
5. Apply the Step 3 to the new \( \{X_t\} \). In case that \( \{X_t\} \) contains interval values, apply the prediction methods to both, the minimums and the maximums of the intervals and obtain a new value according to Step 3.

In case that we have more than one interval instead of numeric values, we form the new interval that ranges from the minimum of every lower value to the maximum of every upper value of the input intervals.

This Step is applied until we predict the future values we needs in advance.

Let us remark that the precedent method, which we shall denote as \( IM(S, n) \) for the time series \( S \) with \( n \) elements, allows predicting future values in form of intervals, i.e., we obtain \( S_{n+1} = IM(S, n) \) the predicted \( n+1 \) value which is an interval-valued solution. Moreover, some elements of \( S \) can be intervals. Therefore, this is a kind of Neutrosophic Statistical method.

If we want to obtain a representative single value since \( IM(S, n) \), we can calculate the following formula to interval \( I_m = [a_1, a_2] \); see [20].

\[ \lambda(I_m) = \frac{a_1+a_2}{2} \tag{7} \]

The rationale of \( IM \) is the following, because every single-valued predicting method has its advantages and disadvantages over the other ones respect to accuracy, we cannot \textit{a priori} establish which is the most appropriate method to solve the problem. But, if \( I_m \) is the interval that contains all these individual values obtained from the methods \( M_1(\{X_t\}), M_2(\{X_t\}), \ldots, M_k(\{X_t\}) \); denoting by \( x_{n+1} \) the predicted single valued number corresponding to the \( i^{th} \)-method and \( v(n+1) \) the actual unknown value to predict, then we can assume that \( \text{probability}(v(n+1) \in I_m) \geq \max\{\text{probability}(v(n+1) = x_{n+1})\} \). Nevertheless, the cost for assuming this principle is that we increase the indeterminacy when we accept intervals rather than numbers as the predicted value.

Let us calculate the prediction of the Ecuadorian tax applying the precedent method. We support our calculus on the R language software Version 2.11.1, see [21][4]. The time series is the monthly observation of the taxes paid in Ecuador in thousand of USD collected as the sum of the income tax, consumption tax and value-added tax, which is depicted in Figure 1.
Figure 1: Depiction of the monthly time series of tax in Ecuador from 2014 to 2018.

For plotting we use the `plot` function of R. See that the series seems to have seasonal fluctuations, to determine this property we obtained the correlogram aided by the function `acf` of R, see Figure 2.

Figure 2: Depiction of the correlogram of the tax time series.

The dotted lines in Figure 2 bound the interval of randomness. We can appreciate that the seasonal effect is more evident yearly, although there also exists a quarterly one.

The linear trend can be calculated from least squares fit obtaining the function $T(t) = 922349.3 + 1561.91 \times t$, which is graphically represented with the solid line in Figure 3. We used `lsfit` function of R. Let us appreciate that the trend is slightly increasing.

Figure 3: Tax time series with its linear trend.

For prediction we use the R functions `predict.ar`, `predict.Arima`, and `predict.StructTS`, which are based on
fitting methods using AR, ARIMA, and Maximum Likelihood respectively. We selected those methods because they are designed for time series prediction. The results are summarized in Table 1 and depicted in Figure 4.

<table>
<thead>
<tr>
<th>Month</th>
<th>Lower prediction</th>
<th>Upper prediction</th>
<th>Middle prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>887369.3</td>
<td>983527</td>
<td>935448.15</td>
</tr>
<tr>
<td>February</td>
<td>724684.8</td>
<td>970209.5</td>
<td>847447.15</td>
</tr>
<tr>
<td>March</td>
<td>964698.5</td>
<td>1311692</td>
<td>1138195.25</td>
</tr>
<tr>
<td>April</td>
<td>948233.3</td>
<td>1107685</td>
<td>1027959.15</td>
</tr>
<tr>
<td>May</td>
<td>813647</td>
<td>972354.5</td>
<td>893000.75</td>
</tr>
<tr>
<td>June</td>
<td>855445.6</td>
<td>972354.5</td>
<td>913900.05</td>
</tr>
<tr>
<td>July</td>
<td>965762.5</td>
<td>1140292</td>
<td>1053027.25</td>
</tr>
<tr>
<td>August</td>
<td>961802.7</td>
<td>1027133</td>
<td>994467.85</td>
</tr>
<tr>
<td>September</td>
<td>872498.3</td>
<td>1036670</td>
<td>954584.15</td>
</tr>
<tr>
<td>October</td>
<td>858309.4</td>
<td>992776.6</td>
<td>925543</td>
</tr>
<tr>
<td>November</td>
<td>940590.4</td>
<td>1086582</td>
<td>1013586.2</td>
</tr>
<tr>
<td>December</td>
<td>962525.2</td>
<td>1315634</td>
<td>1139079.6</td>
</tr>
</tbody>
</table>

Table 1: Predicted interval values for the year 2019 and the intermediate values x 1000 USD.

Let us note that in Figure 4 the upper line corresponds to the upper bound, the lower line corresponds to the lower bound and the intermediate line corresponds to the mean values, according to Equation 7.

Figure 5 contains the plotting of the degree of indeterminacy calculated according to Equation 8 given below.

Let \( I_m = [a_1, a_2] \) be an interval, the indeterminacy of \( I_m \) is calculated as follows:

\[
\gamma(I_m) = a_2 - a_1 \tag{8}
\]

Let us note that in Figure 5 the indeterminacy fluctuates over the time.
Conclusion

This paper proposed a prediction method, where the future value is estimated in the form of an interval rather than in the form of a numeric value. Essentially, the method compiles the results of statistical prediction methods and converts them into a single value in form of an interval. This allows prediction of values with greater veracity, although with less precision. However, this is sufficiently useful to predict the behavior of taxes in Ecuador, in such a way that the monthly predictions of tax collection in Ecuador for the year 2019 are calculated.

References


Received: October 1st, 2019
Accepted: January 15th, 2020