Proceedings of the 6th WSEAS Int. Conf. on NEURAL NETWORKS, Lisbon, Portugal, June 16-18, 2005 (pp250-255)

Analysis of the results of lotteries using statistical methods and artificial neural networks

ARTURO A. HERNANDEZ G.
Escuela de Ingeniería de Sistemas
Facultad de Ingeniería
Universidad de los Andes
Mérida, VENEZUELA

FRANCKLIN RIVAS ECHEVERRÍA
Laboratorio de Sistemas Inteligentes
Facultad de Ingeniería
Universidad de Los Andes
Mérida, VENEZUELA

EVELENIR BARRETO GONZALEZ
Departamento de Algoritmos y Programación
Facultad de Ingeniería
Universidad Católica Andrés Bello
Caracas, VENEZUELA

FELIPE PACHANO
Departamento de Inv. de Operaciones
Facultad de Ingeniería
Universidad de Los Andes
Mérida, VENEZUELA

Abstract: - The popularity of lotteries has increased worldwide in the last few years, and so its revenues, making them an useful study case. This paper shows the research done for testing the randomness of the results of three lotteries and to check if there were any patterns, studying them as time series. This research was done using three methods, statistical randomness tests, autoregressive integrated moving averages (ARIMA), and artificial neural networks (ANN). At the end any sign of patterns or suspicious data was not found.

Key-Words: - **Artificial neural networks, ARIMA, lottery, randomness tests, pattern recognition**

1 Introduction

Everyone would like to know, or at least have a pretty good idea of what will happen in the future. It would be of great help for the human being if the behavior of many variables in natural or economic sciences could be forecasted or at least recognized or classified. That is why many forecasting models and pattern recognizing techniques have developed greatly [11].

As well as scientific variables goes, there is great research but not on everyday-life variables, such as the lottery results. The gambling activity, especially the lottery, has grown a lot in the past few years making this industry a very wealthy one. So, it is interesting to verify if the behavior of the results of the lotteries is the right one, which means random [11].

The main objective of this research is to test the randomness of the data with diverse statistical methods and then apply two different modeling techniques, one with a merely statistical approach and one with a computational approach.

Three lotteries where chosen, two in Venezuela and one in USA. The data from each lottery has a different time span. The three lotteries have a 3-digit result.

A random event is one that can not be foreseen o that can not be predicted [7]. Several methods to verify the randomness of a data set have been developed, among them there is the Chi-square goodness of fit test, Pearson Chi-square test in contingency tables, interval test and run-length test.
Each one of them was applied to the results of three lotteries. First taking the resulting number as a whole and then the numbers where separated and the tests were applied to the last two numbers individually.

The Chi-square goodness of fit test compares the observations of a variable with the expected results for this variable. It will give the answer for the null hypothesis that the observed variable and the expected variable come from the same population.

The hypothesis that the sequence of appearances of a specific number in a data series is random or equivalently to say that the appearances are independent will be checked in the run-length test. [8].

Another test is the Pearson Chi-square in contingency tables. This test was done taking the last two numbers of each result and separating them, having the last number in the columns and the middle number in the columns. It will be checked if the variable in the columns is independent from the variable in the rows [8].

The last test used to verify the randomness of the data was the interval test. The basic idea is to look at the intervals between successive occurrences of the same value. Counting the number of separations of length 1, 2, 3, etc. and computing a chi-square statistic using these counts and the expected counts for each separation length [13].

2 Time Series

A time series is a set of observations made sequentially in time, usually measured in intervals of the same size [3]. Time series are formed by various components such as: Tendencies, cyclic movements, seasonal movements, and random movements. Traditional methods decompose a time series in such movements [11].

There are several objectives for time series study: Description, Explanation, Prediction and Control [6].

Time series can be classified due to the certainty of their behavior. Deterministic series are the ones that follow an exact behavior defined by mathematical rules. Stochastic series can be associated to a probability distribution, so exact forecasting of stochastic series is almost impossible [11].

There are several methods for stochastic time series modeling, two of them will be covered in this paper, ARIMA and Artificial Neural Networks.

3 ARIMA

The ARIMA method is widely used in statistics for time series modeling and forecasting [4]. It is based in Autoregressive Integrated Moving Average processes.

In order to apply the ARIMA method, the time series must be stationary, this is, its mean and variance should remain constant over time. In those cases where the series is not stationary, there are several methods, such a differentiation or logarithmic transformation, which will change the series behavior to a stationary one [11].

This method could be applied to autoregressive processes (AR), moving average processes (MA) and mixed processes (ARMA). The importance of mixed process lies in the fact that they might have less parameters than an AR process or a MA process by itself, following the parsimony principle. Most time series are not stationary, so to apply the method, differentiation is needed, such process are called integrated [5].

The ARIMA method is unable to model seasonal processes, but if a seasonal differentiation is applied they could be modeled. The parameters for the seasonal component would only be taken into account every S periods of time [5].

The model is developed through the multiplication of the polynomial operators that define each model, obtaining model known as ARIMA(p,d,q)(P,D,Q). This model may be expressed as:

$$\phi_P(B^S)\phi_p(B)(1-B)^d\,(1-B^{S})^D\, y_t = \delta + \theta_Q(B^S)\,\theta_q(B)a_t \qquad (1)$$

To find the orders of the model, there are tools known as the autocorrelation function (ACF) and the partial autocorrelation function (PACF)[4].

3.1 ARIMA Time Series Analysis

Time series can be studied in frequency domain or in time domain. This paper focuses time series analysis in time domain [12].

ARIMA method works under the philosophy of letting the information talk by itself [9].

Since this is a very general method, which is one of its strengths, experience is very helpful when applying its phases [5].

ARIMA method is applied following the general steps showed in figure 1 and explained below:
1) Exploratory analysis: it is a first approach for the series. Graphs and Histograms are made and main statistics are found in order to have a brief idea of the behavior of the series. Validation data is selected. At last it is checked if the series is stationary, if not stationary, differentiation and/or transformations are in order [11].

2) Model identification: Using ACF and the PACF, the amount of AR and MA parameters is identified for the seasonal and non-seasonal part.

3) Parameter estimation: the values for the parameters ϕ and θ are estimated through the minimization of the squared sum error, through the conditional or non conditional minimum square method or the maximum likelihood method [5].

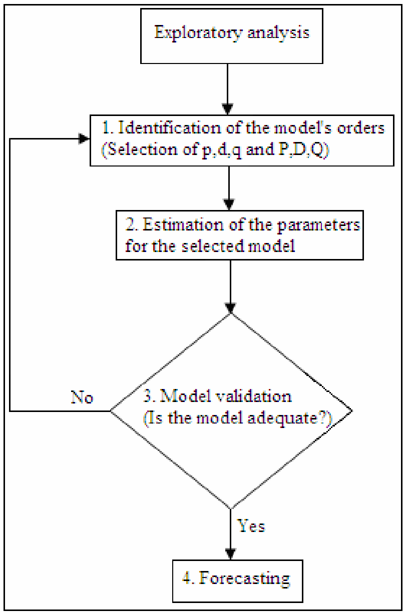
Exploratory analysis
1. Identification of the model's orders (Selection of p,d,q and P,D,Q)
2. Estimation of the parameters for the selected model
No
3. Model validation (Is the model adequate?)
Yes
4. Forecasting

Figure 1. ARIMA Method. [4][11]

4) Checking the adequacy of the model: a model is adequate if it possesses the following conditions: a) The residuals follow a white noise behavior. b) The parameter must be significatively different from zero and it must be incorrelated. c) The goodness of fit can be evaluated with Akaike information criterion and Schwartz Bayesian criterion, among others [5].

5) Forecasting: In this last step, a forecast of the series is done using the selected model. These forecasted values are compared with observed values for the validation of the model. Then the residuals are calculated so the goodness of the model can be evaluated.

If the model is considered final and adequate for forecasting, prediction for n periods with its confidence intervals would be the final step. The length of the prediction would be of a predefined size [11].
For evaluation purposes, the mean error, the mean absolute error, the mean square error are calculated.

4 Artificial Neural Networks

Artificial neural networks (ANN) are inspired in biological neurons; they try to follow certain human abilities done by the brain and by millions of interconnected elements, called neurons. Neurons are the functional cells of the nervous tissue, fundamental for the nervous system. They are capable of receiving signals, processing them and transmitting them to other neurons [2][5].

The biological neural networks were the base for a mathematical model for the transmission of information, that tries to emulate their behavior, at the beginning as an artificial neuron and then growing up to more complex models called Artificial neural networks.

A neuron model has the following components: an input, an output, synaptic weights, a sum point for weighted inputs, an activation function and a threshold;

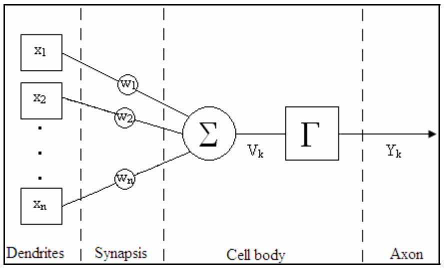

Figure 2. Neuron model. [2]

In figure 2 it can be observed that neuron k can be described as follows:

$$v_k = \sum_{i=0}^{p} x_i w_{ik} \tag{2}$$

$$y_k = \Gamma(v_k) \qquad (3)$$

where x_i are the inputs for the k neuron; w_{ik} are the weights for the k neuron; v_k is the sum of the inputs multiplied by its corresponding weights; Γ is the activation function and y_k is the output for neuron k. The weights will readjust according to the training rule and the activation function selected depending on the nature of the problem and the output required. There are many activation functions to choose from, step function, sigmoid function, ramp function and identity function are among them [1][2][5].

A major issue when using ANN is its topology design. The number of inputs, outputs, hidden layers, activation function, the interconnection direction between neurons of the same layer and between layers [11]. And last but not least, another important issue is the training algorithm to be used. The ANN learning is given following the training algorithm which is a procedure to modify the weights. The most widely used algorithm for training multi-layered feed-forward networks is backpropagation [5].
The stopping criterion for the algorithm is the minimum function error. This minimum could be found using the gradient descent method [14].

Once the network has been trained, it has the ability to answer close to the desired ones when new examples are provided.

4.1 ANN Time Series Analysis

There are several models for time series modeling and forecasting with ANN. The most popular are:

1) Multi-layer perceptron**:** in this type of network every neuron does a biased weighted sum and goes through an activation function to finally give an output. It has a feed-forward topology. This network can model almost any kind of complex function [11].

The number of input and output neurons is defined by the problem. The number of neurons in the hidden layer is no easy to define but a good start would be one hidden layer with the average between the inputs and the outputs as the number of neurons in it [11].

Then, the weights and the threshold should adjust so the prediction error is minimized. This is achieved through the training algorithm, backpropagation in commonly used [11].

2) Radial basis function (RBF) network: it is a three layer feed-forward network. The hidden layer has radial units that give Gaussians outputs. The connection weights between the hidden and the output layer will be adjusted with the pseudo-invert method or the delta rule. The activation function of the neurons is a radial basis function [11].

The output of the hidden layer neurons represents a basis function. The activation functions of the output layer are lineal functions and produce a weighted sum of the output of the previous layer. [10]
RBF networks may have a faster training than the multi-layer perceptron because the lineal transformation of the output can be optimized [3].

5 Results

The methods explained before where applied to three sets of data: 1) Winning numbers of Zulia Lottery on its daily Triple A - 12 pm draw from 11/02/2003 to 01/09/2004 (Monday through Friday). 2) Winning numbers of Caracas Lottery on its daily Triple draw from 03/06/1995 to 07/04/1998 (Monday trough Saturday). 3) Winning numbers of the Lottery of Florida's Lotto on its daily draw from 01/01/1995 to 04/09/2004 (Monday trough Sunday). The size of the sample for each lottery was 465, 828 and 2840 respectively.

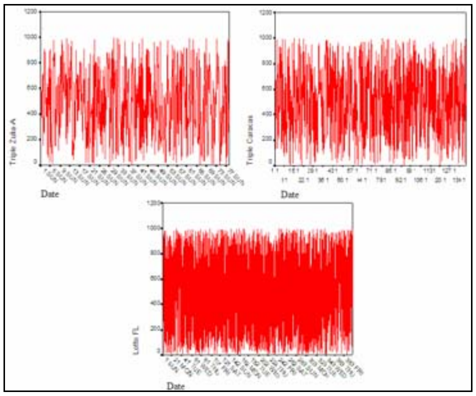

Figure 3. Raw data. Zulia lottery (top left), Caracas lottery (top right), Florida lottery (bottom)

Figure 3 shows the winning numbers time series for each lottery. All the main statistics were found, and the histograms were graphed.

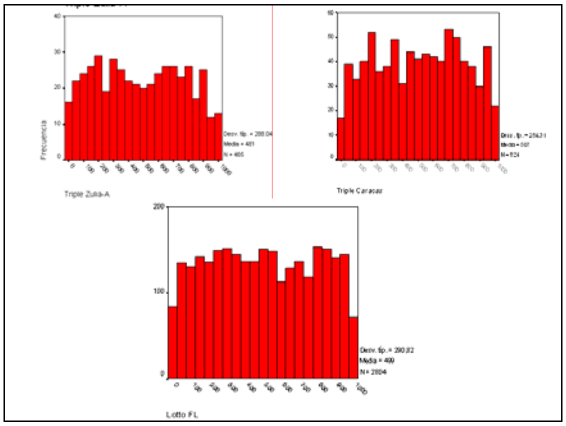

Figure 4. Histograms. Zulia lottery (top left), Caracas lottery (top right), Florida Lottery (bottom).

The randomness tests were applied to the data: The Chi-square goodness of fit test, the Pearson Chi-square test in contingency tables, the interval test and the run-length test. Neither of them but one showed proof of non-random events. In Zulia lottery the numbers 7 and 9 appeared in the second digit of the result 4% more than the expected, rejecting the null hypothesis in the Chi-squared goodness of fit test. The value of the test statistic for theses tests was closer to the rejection zone in Venezuelan lotteries than the USA lottery, but not enough to reject them with a 95% confidence.

When the ACF and PACF were calculated the same result was constant to the three lotteries. The series were stationary with no seasonal component, and no significant values showed in the ACF and PACF graphs, so the data could be considered as white noise or random. As showed in the next figure:

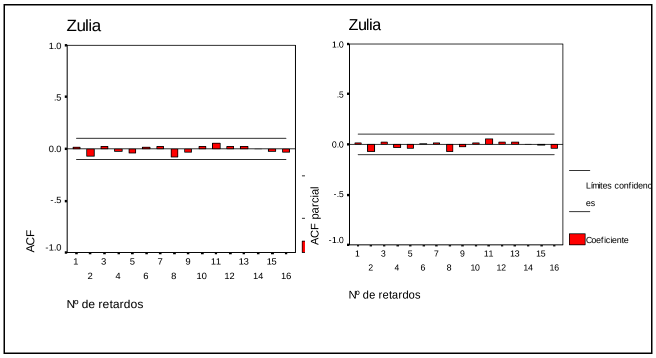

Figure 5. ACF and PACF for Zulia lottery

An ARIMA(1,0,1) model was assumed for the three lotteries. After verifying the adequacy of the model, and no better model was found, the ARIMA(1,0,1) model was taken as best fit for the three series.

In order to apply the ANN, several variables are chosen by trial and error. Assuming one input and that one output is wanted, the amount of neurons and hidden layers is set to one. The learning algorithm will modify the weights, which had a random value at the starting point of the algorithm.

Three different ANN were found to be the better fit for the data with the starting the randomly chosen weights. A two hidden layers perceptron with 20 neurons each for Zulia lottery, a two hidden layers perceptron with 13 neurons on each layer for Caracas lottery and for the Florida Lottery it turned to be a radial basis function network with one layer and 13 neurons as the most fitted.

As figure 6 shows, neither one of the techniques fitted the data well enough to consider the model useful; also confirming the randomness of the data.

Conclusions

Three lotteries were studied, two in Venezuela and one in USA. After applying the randomness tests (Chi-square goodness of fit test, Pearson Chi-square test in contingency tables, interval test and run-length test), there were not statistical significant values to reject the hypothesis of randomness. Even though, after watching the histograms, the lottery of USA showed more uniformity on its frequencies and this crosschecks the results of the randomness tests where Venezuelan lotteries are closer to the rejection zone.

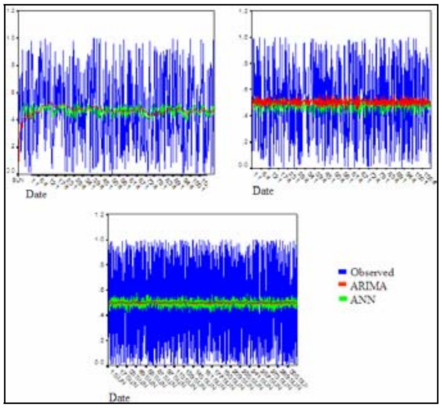

Figure 6. Results. Zulia lottery (Top left), Caracas lottery (Top right), Florida lottery (bottom)

The ARIMA method and the ANN are very useful due to their great adaptability to non-random time series.

The fit and training errors given by both methods were similar and very high, likewise, when validation was done, it gave high errors, so the prediction done would be of no use.

It is recommended to make new tests with other study cases in other countries or other data to monitor and evaluate the randomness of the lotteries worldwide.

Also, a greater diffusion of this modeling/forecasting techniques and the up-to-date software that supports them. Especially for undergraduate students because these methods are considered of great help for research and for scientific problem solving.

References:

[1] Acosta, M. Zuluaga, C. "Redes Neuronales". http://ohm.utp.edu.co/neuronales/main.htm. 2000.

[2] Aguilar, J. y Rivas, F. "Introducción a las técnicas de computación inteligente". Universidad de Los Andes. 2001.

[3] Barreto, E. y Ávila, F "Reconstrucción de data caótica mediante métodos basados en redes neuronales artificiales y dinámicas no lineales". Universidad Metropolitana. 2001.

[4] Box, G., Jenkins, G. y Reinsel, G. "Time series analysis, forecasting and control". Prentince Hall. 1994.

[5] Collantes, J. "Predicción con redes neuronales: comparación con las metodologías de Box y Jenkins". Universidad de Los Andes. 2001.

[6] Chatfield, C. "The analysis of time series: an introduction". Chapman and Hall. 1980.

[7] Diccionario de la lengua española. 22ª Edición. Real Academia Española. 2003.

[8] Ferrán, M. "SPSS para Windows: Programación y análisis estadístico". McGraw-Hill/Interamericana de España. 1996.

[9] Gujarati, D. "Econometría". McGraw-Hill. 1997

[10] Hagan, M., Demuth, H. y Beale, M. "Neural networks design". PWS Publishing Company. 1996.

[11] Hernandez, A. "Análisis de resultados de loterías venezolanas mediante métodos estadísticos y redes neuronales artificiales". Universidad de los Andes. 2005

[12]Mora C., "Modelos Arima: Poblaciones de pequeños mamíferos en la selva nublada de Mérida". Universidad de los Andes. 1996.

[13] Sinclair, B. (1998). "How random is random()?". http://www.owlnet.rice.edu/~elec428/rng/test.html. 1998

[14] Viloria, M. "Evaluación de tecnologías para la predicción de trayectorias de comportamiento para pozos petroleros Universidad de Los Andes. 2002.