Time Series Forecasting using Genetic Algorithm

A Case Study of Maintenance Cost Data For Tunnel Fans

Yamur K. Al-Douri, Hussan Al-Chalabi, Jan Lundberg
Division of Operation and Maintenance Engineering
Luleå University of Technology
Luleå, Sweden
emails: {yamur.aldouri, hussan.hamodi, jan.lundberg}@ltu.se

***Abstract*— Time series forecasting is widely used as a basis for economic planning, production planning, production control and optimizing industrial processes. The aim of this study has been to develop a novel two-level Genetic Algorithm (GA) to optimize time series forecasting in order to forecast cost data for fans used in road tunnels by the Swedish Transport Administration (Trafikverket). The first level of the GA is responsible for the process of forecasting time series cost data, while the second level evaluates the forecasting. The first level implements GA based on the Autoregressive Integrated Moving Average (ARIMA) model. The second level utilizes a GA based on forecasting error rate to identify proper forecasting. The results show that GA based on the ARIMA model produces good forecasting results for the labor cost data objects. It was found that a multi-objective GA based on the ARIMA model showed an improved performance. The forecasted data can be used for Life Cycle Cost (LCC) analysis.**

Keywords*— *ARIMA model; Time series forecasting; Genetic Algorithm (GA); Life Cycle Cost (LCC); Maintenance cost data.

I. INTRODUCTION

Time series forecasting predicts future data points based on observed data over a period known as the lead-time. The purpose of forecasting data points is to provide a basis for economic planning, production planning, production control and optimizing industrial processes. The major objective is to obtain the best forecast function, i.e., to ensure that the mean square of the deviation between the actual and the forecasted values is as small as possible for each lead-time [1]. Much effort has been devoted over the past few decades to the development and improvement of time series forecasting models [2].

The Genetic Algorithm (GA) is often compatible with nonlinear systems and uses a particular optimization from the principle of natural selection of the optimal solution on a wide range of forecasting populations [3]. The proposed multi-objective GA optimizes a particular function based on the ARIMA model. The ARIMA model is a stochastic process modelling framework [4] that is defined by three parameters (p,d,q). The parameter p stands for the order of the autoregressive AR(p) process, d for the order of integration (needed for the transformation into a stationary stochastic process), and q for the order of the moving average MA(q) process [4]. A stationary stochastic process means a process where the data properties have the same variance and autocorrelation [5].

The weakness of the ARIMA model is the difficulty of estimating the parameters. To address this problem, a process for automated model selection needs to be implemented in the automated optimization to achieve an accurate forecasting [6]. The GA is a well-established method which helps in solving complex and nonlinear problems that often lead to cases where the search space shows a curvy landscape with numerous local minima. The GA is designed to find the best forecasting solution through automated optimization of the ARIMA model and to select the best parameters (p,d,q) to compute point forecasts based on time series data. The parameters of the ARIMA model are influenced by the selecting process of the GA. In addition, the GA can evaluate the forecasting accuracy using multiple fitness functions based on statistics models.

Vantuch & Zelinka [7] modified the ARIMA model based on the Genetic Algorithm and particle swarm optimization (PSO) to estimate and predict data of time. They found that the Genetic Algorithm could find a suitable ARIMA model and pointed to improvements through individual binary randomization for every parameter input of the ARIMA model. Their model shows the best set of coefficients obtained with PSO compared with the best set obtained with a classical ARIMA prediction. However, these authors present the ARIMA parameters in a binary setting with limited possibilities and they consider the forecasting based on an ARIMA evaluation only.

Hatzakis & Wallace [3] proposed a method that combines the ARIMA forecasting technique and a multi-objective GA based on the Pareto optimal to predict the next optimum. Their method is based on historical optimums and is used to optimize AR(p) and MA(q) to find a non-dominated Pareto front solution with an infinite number of points. They found that their method improved the prediction accuracy. However, these authors assumed that the data were accurate and used the Pareto front solution to select a proper forecasting. In addition, they did

 ISBN: 978-1-61208-677-4

not use any forecasting error rate to evaluate the forecasting results.

The aim of this study has been to develop a novel two-level multi-objective GA to optimize time series forecasting in order to forecast cost data for fans used in road tunnels. The first level of the GA is responsible for the process of forecasting time series cost data, while the second level evaluates the forecasting. The first level implements GA based on the ARIMA model. This level gives possibilities of finding the optimal forecasting solution. The second level utilizes GA based on forecasting error rate to identify a proper forecasting. We argue that a GA decreases the complexity, increases the flexibility, and is very effective when selecting an approximate solution interval for forecasting.

II. METHODOLOGY

A. Data collection

The cost data concerns tunnel fans installed in Stockholm in Sweden. The data had been collected over ten years from 2005 to 2015 by Swedish Transport Administration (Trafikverket) and were stored in the MAXIMO computerized maintenance management system (CMMS). In this CMMS, the cost data are recorded based on the work orders for the maintenance of the tunnel fans. Every work order contains corrective maintenance data, a component description, the reporting date, a problem description, and a description of the actions performed. Also included are the repair time needed and the labor, material and tool cost of each work order.

In this study, we consider the one cost objects of labor based on the work order input into the CMMS for the ten-year period mentioned above. The selected data were clustered, filtered and inputed for the present study using a Multi-objective GA (MOGA) based on a fuzzy c-means algorithm [14]. It is important to mention that all the cost data used in this study concern real costs without any adjustment for inflation. Due to company regulations, all the cost data have been encoded and are expressed as currency units (cu).

B. Two-level system of Genetic Algorithms

In this study, a novel two-level GA has been developed, as shown in Figure 1. The levels of the GA are as follows: (1) a GA based on the ARIMA model for forecasting the cost data, and (2) GA based on multiple functions for measuring the forecasting accuracy for validation of the forecasted data. Level 1 of the GA is applied to the cost data objects (labor) to forecast data for the next level and for each of 15 different generations. The second level validates the forecasted data for the cost object. Using two levels allows us to reduce the computational cost [8], while reaching an effective and reasonable solution [9].

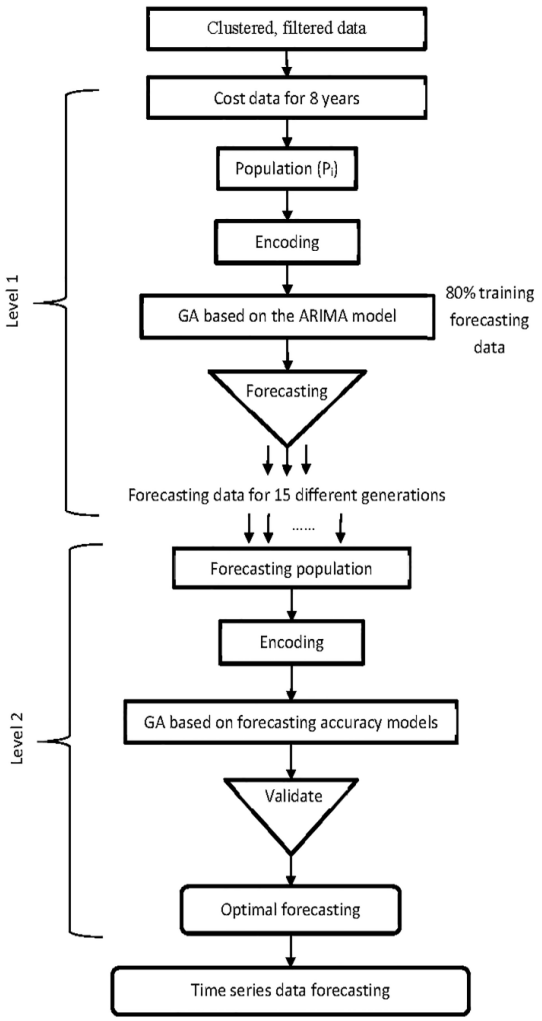

Figure 1. Two-level of Genetic Algorithm (GA)

1) GA based on the ARIMA model

The proposed GA method uses a particular optimization based on the principle of natural selection of the optimal solution and applies this optimization on a wide range of forecasting ***populations***. The GA creates populations of ***chromosomes*** as possible answers to estimate the optimum forecasting [3]. This algorithm is robust, generic and easily adaptable because it can be broken down into the following steps: initialization, evaluation, selection, crossover, mutation, update and completion. The evaluation (***fitness function***) step creates the basis for a new population of chromosomes. The new population is formed using specific Genetic operators, such as ***crossover*** and ***mutation*** [10], [11]. The fitness function is derived from the ARIMA forecasting model. A GA with automated optimization avoids the weakness of the ARIMA model by estimating the parameters for forecasting [6].

The GA is a global optimization technique that can be used to achieve an accurate forecasting based on the ARIMA model. The GA is known to help in solving

complex nonlinear problems that often lead to cases where the search space shows a curvy landscape with numerous local minima. Moreover, the GA is designed to find the optimal forecasting solution through automated optimization of the ARIMA model. In addition, the GA can evaluate the forecasting accuracy using multiple fitness functions based on statistical models.

The first level utilizes a GA which is based on the ARIMA model and is implemented four different times using a cross-validation randomization technique. The technique aims to select the best time series data for forecasting. The process is the following: a random number of cost data are selected based on encoding in each of the four implementations; the modified random cost data are generated 15 times. The modifications are used to find the optimal cost data for forecasting. The following steps are implemented when applying the multi-objective GA in level 1.

Step 1: Initial population

A longitudinal study of each cost object ($Z^{labour}, Z^{material}$) is used to forecast data using the multi-objective GA for the two objects in parallel.

Step 2: First GA generation and selection

The first generation is performed by selecting each cost object and checking whether the data are stationary (i.e., trend-stationary) or non-stationary using a Dickey-Fuller test [12]. To apply the ARIMA model, the data should be stationary, i.e. the null hypothesis of stationarity should not be rejected. When applying the Dickey-Fuller Test (DFT) in equation (1), the hypothesis p = 1 means that the data are non-stationary and p < 1 that the data are stationary [12].

$$DFT(x_t) = \alpha + px_{t-k} + \epsilon_t \quad (1)$$

α : constant estimated value of the time series data;

p : the hypothesis is either p = 1 or p < 1;

t : time {1,…,k};

ε : the white noise of the time series data.

Step 3: Encoding

Random values, either ones or zeros, are generated for each cost data object. Encoding is the process of transforming from the phenotype to the genotype space before proceeding with GA operators and finding the local optima.

Step 4: Fitness function

The fitness function is based on the ARIMA model for the forecasting of time series cost data objects individually, as seen in the equation below. The fitness function consists of an autoregression (AR) part and a moving average (MA) part [1]. The ARIMA model uses AR and MA polynomials to estimate (p) and (q) [7].

The fitness function is formulated as the equation (2) follows:

$$fitness(p,d,q) = \mu + \sum_{i=1}^{p}(\sigma y_{t-1}) + \sum_{i=1}^{q}(\theta\epsilon_{t-1}) + \epsilon_t \quad (2)$$

where the following notation is used:

μ : the mean value of the time series data;

p : the number of autoregressive lags;

d : the number of differences calculated with the equation $\Delta y_t = y_t - y_{t-1}$;

q : the number of lags of the moving average process;

σ : autoregressive coefficients (AR);

θ : moving average coefficients (MA);

ϵ : the white noise of the time series data.

The parameters (p,q) are estimated using an autocorrelation function (ACF) and a partial autocorrelation function (PACF) [1]. The estimated values produced by the previous equation will be used to create a forecast for 20 months (m) using the equation (3). These forecasted values will be evaluated using the second level of GA to find the optimal forecasting with high accuracy.

$$fitness(t+m) = \mu + \sum_{i=1}^{p}(\sigma y_{t-1}) + \sum_{i=1}^{q}(\theta\epsilon_{t-1}) + \epsilon_t \quad (3)$$

where fitness(t+m) is the time series forecasting at time (t+m) and

m : months {1, 2, 3,…, m}.

Step 5: Crossover and mutation

In this study, a one-point crossover with a fixed crossover probability is used. This probability decreases the bias of the results over different generations caused by the huge data values. For chromosomes of length l, a crossover point is generated in the range [1, 1/2 l] and [1/2 l, l]. The values of objects are connected and should be exchanged to produce two new offspring. We select two points to create more value ranges and find the best fit.

Randomly, ten percent of the selected chromosomes undergo mutation with the arrival of new chromosomes. For the cost object values, we swap two opposite data values. The purpose of this small mutation percentage is to keep the forecasting changes steady over different generations.

Step 6: New generation

The new generation step repeats steps 3 to 5 continuously for 15 generations. Fifteen generations are enough for these data because the curves of the fitness functions are repeated after fifteen generations. The selected fifteen generations are used individually for the second level to validate the forecasting accuracy for each object and population. This step yields fully correlated data for the next step.

2) GA for measuring the forecasting accuracy

In this level, the GA is applied longitudinally to the data. The GA operates with a population of chromosomes that contains labor cost and material cost objects. The GA operates on the selected population over different generations to find the appropriate forecasting accuracy. During the GA generations, the chromosomes in the population are rated concerning their adaptation, and their

mechanism of selection for the new population is evaluated. Their adaptability (fitness function) is the basis for a new population of chromosomes. The new population is formed using specific Genetic operators such as crossover and mutation. The GA is used to evaluate the forecasting accuracy for each generation of the first level.

Level 2 utilizes a GA which is based on different forecasting error rates and is implemented for each generation from the first level and for four different populations using a cross-validation randomization technique. This technique aims to select the best evaluation of the time series data forecasting and the process is as follows. A random number of cost data are selected based on the encoding in each generation of the four implementations, and the modified random cost data are generated five times. The modifications are then used to find the optimal cost data forecasting. In this study, due to the size of the training data, five generations are sufficient to obtain valid results. The following steps are implemented when applying the GA in level 2.

Step 1: Initial population

A longitudinal study of each cost object ($Z^{labour}, Z^{material}$) is used to forecast data using the multi-objective GA for the two objects in parallel.

Step 2: First GA generation, encoding and selection

The first generation is performed by selecting each cost object and encoding through generating random values, either ones or zeros, for each cost data object. The selection for each cost data object is based on encodings with the value of 1. This selection is used to evaluate the forecasted data using the multi-objective fitness function.

Step 3: Fitness function

The multi-objective fitness function is based on multiple functions for measuring the forecasting accuracy. The mean absolute percentage error (MAPE) is used to evaluate the selected forecasting data from the previous step [13]. The fitness functions are formulated as equation (4):

$$fitness(MAPE) = mean(|p_i|) \quad (4)$$

where

$p_i = \frac{100e_i}{Y_i}$ and $e_t = Y_i - F_i$

t : time $\{1, \ldots, k\}$;

Y_t: the actual data over time;

F_t: the forecasted data over time.

Step 4: Crossover and mutation

In this study, we use a one-point crossover with a fixed crossover probability. This probability decreases the bias of the results over different generations due to the huge data values. For chromosomes of length l, a crossover point is generated in the range [1, 1/2 l] and [1/2 l, l]. The values of objects are connected and should be exchanged to produce two new offspring. We select two points to create more value ranges and find the best fit.

Randomly ten percent of the selected chromosomes undergo mutation with the arrival of new chromosomes. For the cost object values, we swap two opposite data values. The purpose of this small mutation percentage is to keep the forecasting changes steady over different generations.

Step 5: New generation

The new generation step repeats steps 2 to 4 continuously for five generations. Five generations are enough for these data, because the fitness function is repeated after the fifth generation. The selected generation is used for the second level to validate the forecasting accuracy for each object. This step yields fully correlated data that can be used for forecasts covering several months.

III. RESULTS OF THE TWO-LEVEL GAS

A. Results of Level 1: GA based on the ARIMA model

In this part of the study, we tested GA based on the ARIMA model to generate forecasting data for the labor cost object. The forecasted data for each population obtained with 15 different generations were then evaluated using the second level. The second level evaluation helped in deciding the best generation of the forecasted data. In this section, we present only the best forecasted curves with the historical data because of the huge number of possibilities considered in this study.

Figure 2 shows the forecasted labor data curve for 20 months from 2013 to 2015, for the second population and, specifically, for generation 13. In addition, it shows the historical data with the polynomial trend to illustrate the relationship between the independent variables over a timeline with monthly intervals. The selected labor data show a better forecasting than that obtained with the ARIMA model in that the forecasted data are close to the actual data and the polynomial trend. The ARIMA parameters for the selected labor cost data covering 47 months were p= 0.22, d= 1 and q= 0.23.

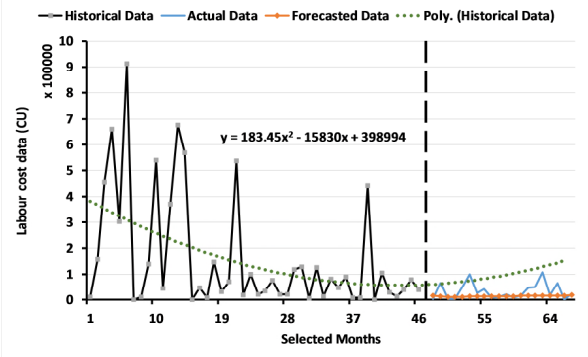

Figure 2. Labor cost data forecasting using the GA based on the ARIMA model

The forecasted data for the labor cost object were evaluated using the second level, applying a GA based on the statistical forecasting error rate. The model for the forecasting accuracy evaluated the forecasted data for 20 months from 2013 to 2015 based on the actual values of this period. Implementing level 2, the accurate forecasted

data were found, i.e. the proper selection of data for each object to be used for forecasting.

Figure 3 show an example of bias in forecasting at the second population and seventh generation. The forecasted data for the 20-month period do not seem to be in sync with the historical data before the vertical line or with the polynomial trend curve after the vertical line. The forecasted data are higher than the historical and polynomial trends.

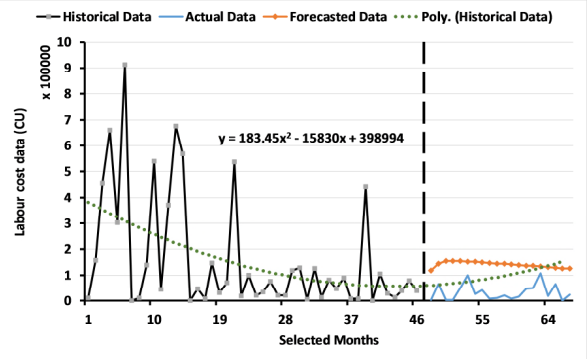

Figure 3. Labor cost data forecasting using the GA based on the ARIMA model

B. *Results of level 2: GA for measuring the forecasting accuracy*

The outcome from the first level, specifically for each generation for each population, indicates the forecasting accuracy for each cost object. For each generation, the GA based on multiple fitness functions was used to find the best fitness value through five different generations. The fitness functions (forecasting error rate models) provide an accurate data forecasting through comparing the behaviour of the different models and revealing which forecasting model is appropriate.

Selecting a proper population for the labor cost data is quite difficult due to the variety of fitness values. In this study, we considered the population that was selected by fitness value that have a low forecasting error rate. The twelve generation for the forth population was selected as having the lowest forecasting error rate with a suitable selection of input data. The labor cost data were selected using fitness (MAPE), with value of 0.4 as seen in Figure 4.

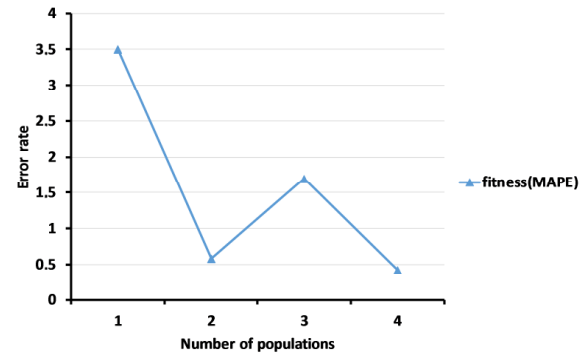

Figure 4. Fitness(MAPE) for four populations

IV. CONCLUSIONS

The GA based on the ARIMA model provides other possibilities for calculating the parameters (p,d,q) and improves the data forecasting. The outcome of the multi-objective GA based on the ARIMA model can be used to forecast data with a high level of accuracy, and the forecasted data can be used for LCC analysis.

ACKNOWLEDGMENT

The authors would like to thank Dr. Ali Ismail Awad, Luleå University of Technology, Luleå, Sweden, for his support concerning the research methodology and for allowing us to use the computing facilities of the Information Security Laboratory to conduct the experiments in this study. In addition, we would like to extend our gratitude to Prof. Peter Soderholm at the Swedish Transport Administration (Trafikverket) for supplying the data for this study.

REFERENCES

[1] G. E. Box, G.M. Jenkins, G.C. Reinsel and G.M. Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 2015.

[2] Y. Chen, B. Yang, J. Dong and A. Abraham, "Time-series forecasting using flexible neural tree model," Inf.Sci., vol. 174, no. 3, pp. 219-235, 2005.

[3] I. Hatzakis and D. Wallace, "Dynamic multi-objective optimization with evolutionary algorithms: a forward-looking approach," in Proceedings of the 8th annual conference on Genetic and evolutionary computation, 2006, pp. 1201-1208.

[4] N. R. Herbst, N. Huber, S. Kounev and E. Amrehn, "Self-adaptive workload classification and forecasting for proactive resource provisioning," Concurrency and computation: practice and experience, vol. 26, no. 12, pp. 2053-2078, 2014.

[5] D. Kwiatkowski, P.C. Phillips, P. Schmidt and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" J.Econ., vol. 54, no. 1-3, pp. 159-178, 1992.

[6] R. J. Hyndman and Y. Khandakar, Automatic time series for forecasting: the forecast package for R, Monash University, Department of Econometrics and Business Statistics, 2007.

[7] T. Vantuch and I. Zelinka, "Evolutionary based ARIMA models for stock price forecasting," in ISCS 2014: Interdisciplinary Symposium on Complex Systems, 2015, pp. 239-247.

[8] S. Thomassey and M. Happiette, "A neural clustering and classification system for sales forecasting of new apparel items," Applied Soft Computing, vol. 7, no. 4, pp. 1177-1187, 2007.

[9] C. Ding, Y. Cheng and M. He, "Two-level Genetic Algorithm for clustered traveling salesman problem with application in large-scale TSPs," Tsinghua Science & Technology, vol. 12, no. 4, pp. 459-465, 2007.

[10] O. Cordón, F. Herrera, F. Gomide, F. Hoffmann and L. Magdalena, "Ten years of Genetic fuzzy systems: current framework and new trends," in IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th, 2001, pp. 1241-1246.

 ISBN: 978-1-61208-677-4

[11] C. Shi, Y. Cai, D. Fu, Y. Dong and B. Wu, "A link clustering based overlapping community detection algorithm," Data Knowl.Eng., vol. 87, pp. 394-404, 2013.

[12] S. J. Leybourne, T.C. Mills and P. Newbold, "Spurious rejections by Dickey–Fuller tests in the presence of a break under the null," J.Econ., vol. 87, no. 1, pp. 191-203, 1998.

[13] R. J. Hyndman and A.B. Koehler, "Another look at measures of forecast accuracy," Int.J.Forecast., vol. 22, no. 4, pp. 679-688, 2006.

[14] Yamur K. Aldouri, Hassan Al-Chalabi, and Zhang Liangwei. "Data clustering and imputing using a two-level multi-objective genetic algorithm (GA): A case study of maintenance cost data for tunnel fans." Cogent Engineering 5.1: 1-16, 2018.

 ISBN: 978-1-61208-677-4