*Available online www.jsaer.com*

***Journal of Scientific and Engineering Research*, 2021, 8(1):55-74**


**Research Article**

**ISSN: 2394-2630**
**CODEN(USA): JSERBR**

# Predicting the On-Time Percentages of Local Trains of TRA in Taiwan

**Chih-Ming Hsu***

*Department of Business Administration, Minghsin University of Science and Technology, Hsinchu, Taiwan

**Abstract** It is a critical issue to accurately predict the on-time percentages of trains since it would significantly affect the management of trains' operation in various fields. However, the problem for predicting the on-time percentages of trains is complex and difficult since there are several factors that can influence the on-time percentages of trains. This study develops a prediction procedure to resolve such a problem based on the clustering technique, feature selection and genetic programming (GP). The clustering approach is utilized to partition the data into clusters, whose clustering performance is evaluated by Davies-Bouldin index (DBI). Several groups of feature variables with different significance to the dependent variable, i.e. on-time percentage, are then determined. Finally, the GP technique is applied to construct prediction models whose independent variables coming from the feature variables determined previously. A case study on predicting the on-time percentages of local trains operated by the Taiwan Railway Administration (TRA) in Taiwan is demonstrated to illustrate the usefulness, effectiveness, and efficiency of the proposed approach. According to the experimental results, the clustering technique along with DBI can effectively identify the importance of each independent variable, as well as the GP can construct an adequate model for predicting the on-time percentages of local trains. In addition, a comparison shows that the feature selection technique can balance the accuracy and complexity of a GP prediction model. Hence, our proposed prediction approach can be considered as a useful, effect, and efficient procedure for dealing with a predicting problem in the real world.

**Keywords** On-time percentage, Clustering technique, Feature selection, Genetic programming, Taiwan Railway Administration (TRA)

## Introduction

There are various kinds of prediction problems existing everywhere in our living world, such as predicting the stock prices, temperature, house prices, election, etc. In the domain of rail transportation, predicting the on-time percentages of trains accurately is an important task because it would greatly affect the management of trains' operation in many fields, e.g. setting up an appropriate timetable, arranging the waiting of trains, allotting tracks to trains, determining the sufficient manpower requirement etc. However, it is a very complicated and difficult prediction problem since there are numerous factors that might influence the on-time percentages of trains with different levels. There are lots of approaches, that apply techniques originating from different areas, had been proposed to deal with the prediction problems in rail transportation, as well as to demonstrate their effectiveness and usefulness. For example, [1] explores the prediction and analysis regarding the train-vehicle crash at passive highway-rail grade crossings by using a nonparametric statistical method, along with the hierarchical tree-based regression (HTBR) where the Federal Railroad Administration (FRA) database focusing on 27 years (from 1980 through 2006) of train-vehicle accident history in the United States is used. The authors make a cross-sectional statistical analysis through using the HTBR to investigate the public highway-rail grade crossings which had been upgraded from the crossbuck-only to the stop signs, without involving of additional traffic-control instruments or other automatic counterplots. The HTBR models utilized to predict the frequencies of train-vehicle crash both for the passive grade crossings controlled by crossbucks only, as well as forth crossbucks that

are coupled with the stop signs. In addition, the change of crash frequencies after the stop-signs have been used at the crossbuck-only-controlled crossings are also assessed. According to the results from the research, the installation of stop-signs can be an effective engineering countermeasure for improving the safety of passive grade crossings. Furthermore, the HTBR models can be utilized to decision makers and traffic engineers can use to analyze the train-vehicle crash frequencies at passive crossings, as well as to synthesize the specific attributes of given crossings for evaluating the potential effectiveness of installing stop-signs. [2] applied the semi-analytical/FEM model topropose an approach for predicting the ground vibrations induced by the high-speed trains while passing through continuous girder bridges. The proposed method consists of two steps where the first one deals with the reaction on top of a pier through using the semi-analytical dynamic interaction model for a train-track-continuous girder while the dynamic characteristics of train, track, and bridge are considered, as ell as the other one resolves the ground vibrations by constructing a 3-dimensional pile foundation-soil finite element model along with the application of the negative reaction force yielded from the pier top to the pile foundation-soil model. They demonstrate the effectiveness of the proposed method through comparing to the other existing approach, as well as the experimental results indicate that their proposed method is reliable and can be practically used to resolve the prediction problem regarding the ground vibration induced by the high-speed railway train while moving along the continuous girderbridges. [3] analyzed the main factors that can influence the safe operations of the high-speed railway to develop an approach for resolving the forecasting problems regarding the high-speed train's safe operation by constructing a fuzzy logic model based on building object set, factor set and judgment set. In their study, the fuzzy assessment function is utilized to make the judgement about the safety of a high-speed railway in real time and assessment indicators. Through introducing the fuzzy assessment function into the coastal high-speed railway construction, the Coastal Economy and Development could be furtherly accelerated. According to the experimental results, their proposed methodology can effectively forecast the safety of a high-speed train on moving thus providing the appropriate guarantee for it. [4] presented a model to predict the aerodynamic noise arisen from the train pantograph based on the semi-empirical component. In their study, an assembly of cylinders and bars with particular cross-sections is used to approximate a pantograph. In addition, the coefficients of the model that accounts for the factors, including the incident flow speed, diameter, cross-sectional shape, yaw angle, rounded edges, length-to-width ratio, incoming turbulence and directivity, are obtained by considering an empirical database. The incoherent sum of predicted noise resulting from different pantograph struts is then used to obtain the overall noise. The validation of their model is made by using available wind tunnel noise measurements with two full-size pantographs. According to the demonstration results, the semi-empirical model has the potential to be a quick tool for predicting the aerodynamic noise resulting from the train pantographs. [5] considered the Holt-Winters model, taking advantage of time series characteristics of passenger flow, as well as the changes of TSF for the OD in different time during a day to propose a hybrid model for exploring the impact of the train service frequency (TSF) regarding the HSR on the passenger flow. In their study, the final hybrid model is generated by integrating the two models based on the minimum absolute value method. Their developed model is verified through making a case study in analyzing the operational data of Beijing-Shanghai high-speed railway from 2012 to 2016. Their model can also be applied to forecast the effects of the TSF except for providing the forecasting ability with a definite formation. [6] proposed an iterative computing approach for the train-load-induced uneven settlement of the transition zone to predict the subgrade uneven settlement, regarding the rail deflections, that mainly occurs in the bridge-embankment transition zones in a high-speed railway. They explore the vehicle track interactions, as well as the deviator stress field about the transition zone by utilizing a vehicle-track-subgrade model. The deterioration process regarding the uneven settlement of the transition zone can be therefore obtained through combining with the soil cumulative plastic strain model. Their experimental results show that the uneven settlement of the transition zone arisen from the train loads tend can be steady while the number for the repeated load applications exceeds 40,000. In addition, both of the settlement of the subgrade at the first five meters measured from the abutment, and the 25 to 30 meters from the abutment can vary instantly. In other words, the more strengthening should be appropriately made, as well as the further attentions should be provided for the two regions in the process of track maintenance. [7] developed an integrated model for simultaneously forecasting the demand and planning the train stop for the high-speed rail (HSR). In their study, the modal

choice or modal split that forecast the demand for forecasting the demand of the HSR is first established. Then, the train stop planning problem (TSPP) that determines the stations at which the train trip should stop thus forming the train's stop-schedule to satisfy the demand while the travel demand and the number of train trips are given. A nonlinear model is applied to integrate and formulate the modal choice problem (MCP) and train stop planning problem (TSPP) with a goal that intends to maximize the total demand arisen from a high-speed rail system. The authors develop a heuristic iterative algorithm to solve such an integrated problem. In addition, a case study on investigating the relationship between the demand and service for the Beijing-Shanghai HSR corridor in China is conducted. Based on the empirical analyzed results, the modal choice and train stop planning should be combined and considered simultaneously toprovide a sustainable design forthe HSR train services. Furthermore, their proposed method can also provide a theoretical basis for evaluating the adaptability of the service network to the travel demand by simulating the impact regarding the number of stops onits mode share through reflecting the changes of travelers' behaviors based on the HSR train stop planning.[8] proposed a method for calculating the failure rate of an Automatic Train Protection (ATP) system where the C-C algorithm is utilized to determine the delay time and embedded dimension. The phase space then can be reconstructed from one-dimensional time series to a high-dimensional space. Therefore, an intelligent forecasting model for the short-term failure rate in an ATP system is established according to the chaotic characteristics of a failure rate. A case study aiming to forecast the failure rates from 2010 to 2018 is made to confirm the validity of their proposed model. The proposed chaos prediction model can provide an accuracy of 99.71%, that is superior than the pure support vector machine (SVR) method according to the prediction results. The maintenance inflexibility and imbalance of supply and demand of spare parts configuration thus can be resolved by predicting the failure rate intelligently. [9] developed a model to resolve the prediction problem regarding the train delays in the complex train operations with a dependency nature by using a Bayesian network-(BN).In their study, there are three BN schemes including (1) heuristic hill-climbing, (2) primitive linear, and (3) hybrid structure, are applied to explore the real-world train operation data gathered from a high-speed railway line. The dependency graph of the developed structures is first rationalized through utilizing the historical data. Then, the over-fitting problem is avoided, as well as the prediction performance is evaluated against the other models by training each BN structure with a verifying approach of the gold standard k-fold cross validation. According to the validation results, the superposition and interaction effects of train delays can be efficiently captured through using a BN-based model. In addition, However, a superior model can be obtained by developing a well-designed hybrid BN structure based on the domain knowledge and judgments from the expertise and local authorities. The performance comparison results from the hybrid BN structure against the real-world benchmark data reveal that their proposed method can achieve over 80%prediction accuracy within a 60-min period, thus yielding the lower prediction errors while evaluating through the mean absolute error (MAE), mean error (ME) and root mean square error (RMSE) measures on average.[10] applied the auto regression (AR) as well as the support vector regression (SVR) models, whose weights are optimized through using the chaotic particle swarm optimization (CPSO) algorithm, to develop a hybrid model, named the AR-SVR-CPSO model, for the improvement about the prediction accuracy. They first utilize the AR model along with various methods to construct the prediction model for the vibration time series. The SVR methodology combined with the phase space reconstruction is then employed to construct a model for predicting the vibration time series again. Finally, the CPSO method is used to optimize the weights with which the prediction values obtained from the AR and SVR models are weighted and summed together. Their proposed approach is verified by conducting a case study with the data collected from the reliability test platform for the high-speed train transmission systems, as well as from the "NASA prognostics data repository". According to the experimental results, their proposed hybrid model can provide the superior prediction performance than the traditional AR and SVR models. [11] proposed a hybridized approach for predicting the disruptions and disturbances during train operations, including the primary delay, the number of affected trains, and the total delay times by using the Bayesian network (BN) paradigm. First, they explore the dependencies of the concerned factors on each station and among the adjacent stations while the domain knowledge and expertise about the operational characteristics are well-known for obtaining an effective BN structure. Through integrating the expert knowledge, interdependencies learned from real-world data, as well as real-time prediction and operational requirements,

four candidate BN structures are then presented. Finally, a 5-fold cross-validation methodology is applied to train these candidate structures with the operational data gathered from the Wuhan-Guangzhou (W-G) and Xiamen-Shenzhen (X-S) high-speed railway (HSR) lines in China. The structure with the best prediction performance is designated as the model for predicting the consequences of disruptions and disturbances in the two HSR lines. The comparison results reveal that their proposed approach can outperform the three other predictive models commonly used by providing the superior average prediction accuracy. [12] applied the state-space equations to develop an approach for alleviating the errors regarding the predicted delays of a train in the neighbouring stops. In their study, the errors are reduced by using a linear quadratic regulator based on the modern control theory, as well as the proposed method is also supported through making the required derivation and proving. In addition, the necessary parameters about the regulator are optimized by utilizing an artificial fish swarm algorithm, and the corresponding simulation is carried out via the SIMULINK architecture. The execution results reveal that their proposed method can be an effective reference for both theoretical innovation and practical application in operating a railway system.

The above literature review shows that the prediction problems regarding the on-time percentages of trains had not been investigated in the previous researches. However, many issues, such as making a suitable timetable, arranging the stations for waiting other trains, allocating the running tracks for trains, determining the appropriate manpower etc., are considerably affected by the on-time percentages of trains, thus further influencing the operation of a railway corporation. Therefore, the predicting regarding the on-time percentages of trains has become a crucial task for the sustainable operation of railways. However, various factors, e.g. the total number of passengers, passengers' types, operating days of trains, the initiation and terminal stations or time, the running distances for trains etc., that can affect the on-time percentages of trains. These factors are not identified easily, as well as simultaneously considered completely. Furthermore, gathering the operation data of trains is also a very difficult work. Next, these considered factors may have their own importance while predicting the on-time percentages of trains. Besides, the importance of each factor is not simple to be determined either. Hence, the predicting for the on-time percentages of trains has been considered a very complicated and thorny problem, thus obtaining far fewer studies. Therefore, the clustering technique, feature selection and genetic programming (GP) are sequentially applied to develop a systematic prediction approach to tackle the problems regarding the on-time percentages of trains. The remaining sections are organized as follows. Section 2 briefly introduces the analyzing and modelling methods. Our proposed prediction approach is then presented in Section 3. In Section 4, a real case study, aiming to predict the on-time percentages of local trains operated by the Taiwan Railway Administration in Taiwan, is provided to demonstrate the usefulness, effectiveness, as well as efficiency of our proposed prediction approach. Conclusions are finally provided in Section 5.

## Methodologies

### *TwoStep Cluster Analysis*

In various situations, the original data of a problem must be grouped, for some purpose, into a certain number of clusters such that the data allocated in the same cluster are more similar to each other than to the data assigned to the other clusters. Such a technique is called cluster analysis that can be achieved throughusing various approaches, and TwoStep cluster analysis is one of these famous methods. A likelihood distance measure is applied, as well as the variables in the cluster model are assumed to be independent in the TwoStep cluster analysis. In addition, the distributions for each continuous variable and categorical variable are assumed to be normal multinomial, respectively. There are several desirable features, including: (1) It can cluster the data for both categorical and continuous variables; (2) It can automatically determine the optimal number of clusters; (3) It can efficiently analyze the data with even a large scale, while comparing the TwoStep cluster analysis to the traditional clustering techniques. The procedure of TwoStep cluster analysis is summarized as follows:

**Step 1. Constructing a cluster features (CF) tree**

Firstly, the initial case is arranged at the root of the tree in a leaf node that involves the variable information regarding that case by the cluster features (CF) tree. Then, each succeeding case either joins to an existing node or forms a new node based on its similarity to the existing nodes along with the similarity criterion that

measuring the distance between nodes. Each node therefore can summarize the information of variables for all cases clustered in that node. In other words, the summary information for the data file then can be capsuled by the CF tree.

**Step 2. Grouping leaf nodes**

A range of solutions can be produced through grouping the leaf nodes of a CF tree by using an agglomerative clustering algorithm. The optimal number of clusters, i.e. groups, then can be determined through using the Schwarz's Bayesian Criterion (BIC) [13] or the Akaike Information Criterion (AIC) [14] clustering criterion.

***Genetic Programming***

The famous theory of natural selection and evolution, put forward by Darwin, describes the evolution progress regarding organisms in the natural world. According to the inspiration from Darwin's evolution theory, [15] therefore presented a well-known optimization method for resolving an optimization problem by imitating the evolutionary procedure of living beings, called genetic algorithms (GAs). The Gas use a chromosome, consisting a series of genes, that mimics the chromosome of a living thing to stand for a feasible solution, i.e. an individual, to an optimization problem. A population for an optimization problem is then formed by all individuals. A fitness function is well-designed according to the objective function of an optimization problem to assess the quality of a feasible solution in the population for tackling the optimization problem. The obtained evaluation value is called fitness. Besides, an appropriate mechanism of natural selection and matching is also designed to simulate the marriage process about individuals, thus forming a matching pool. Hence, each pair of individuals, named parents, in the matching pool can hopefully breed new individuals, called offspring, with the better quality by using the well-defined crossover functions that closely correlate to the fitness of a feasible solution in the optimization problem. Furthermore, the unusual situation of crossover, i.e. extraordinary genetic changes, is achieved by devising a mutation function. Finally, the individuals in the offspring are also assessed by the fitness function, and a new population in the next generation then can be developed through replacing the worse (weak) individuals in the previous generation, i.e. the parents' generation, by the better individuals among the offspring. In order to extend the GAs into a field of computer programs, [16] presented the genetic programming (GP) where a feasible solution, i.e. program, is expressed by using a tree-based structure as shown in Figure 1. The tree in Figure 1 represents a computer program that can be obtained by decoding the tree from left to right, as well as from bottom to top, as follows

$$3 - \frac{x}{15} + 8 \times \sqrt{y} \quad (1)$$

The elements coming from the two parts, including the (1) terminal set and (2) function set, made up the GP tree. The available elements for each terminal branch of a GP tree are defined by the terminal set. These elements can be an independent variable, zero-argument function, or random constant, etc. For example, the 3, *x*, 15, 8, and *y*in Figure 1 are the elements from the terminal set. On the other hand, the function set define a set of primitive functions available for each branch in a GP tree, e.g. addition, square root, multiplication, sine and others. In Figure 1, the +, -, ×, ÷, and √ are the elements in the function set. The fitness corresponding to the solution expressed by Equation (1) then can be evaluated by feeding variables *x* and *y* into the equation, as well as referring the objective function that it is intended to optimize. In addition, the crossover and mutation operators in GAs must be transformed into the styles that can agree with the solution of a tree-based structure in GP. As illustrated in Figures 2, the original paired solutions are

$$3 - \frac{x}{15} + 8 \times \sqrt{y} \quad (2)$$

And

$$4 + \cos(x) - \frac{\log(y)}{6z} \quad (3)$$

By using the crossover operator, the new paired solutions can be obtained as follows

$$3 - \cos(x) + 8 \times \sqrt{y} \quad (4)$$

and

$$4 + \frac{x}{15}\cos(x) - \frac{\log(y)}{6z} \quad (5)$$

As for the mutation operator, the original tree $3 - \frac{x}{15} + 8 \times \sqrt{y}$ mutates into a new solution $3 - \frac{x}{15} + 8 \times (5 + y)$ as shown in Figure 3.

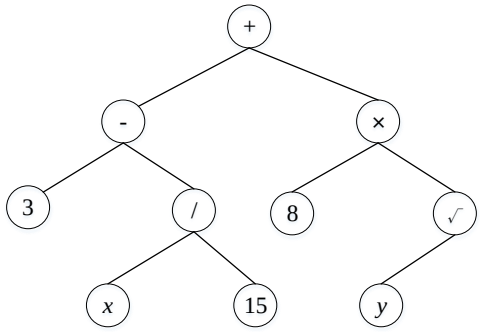



*Figure 1: Tree-based structure in GP*

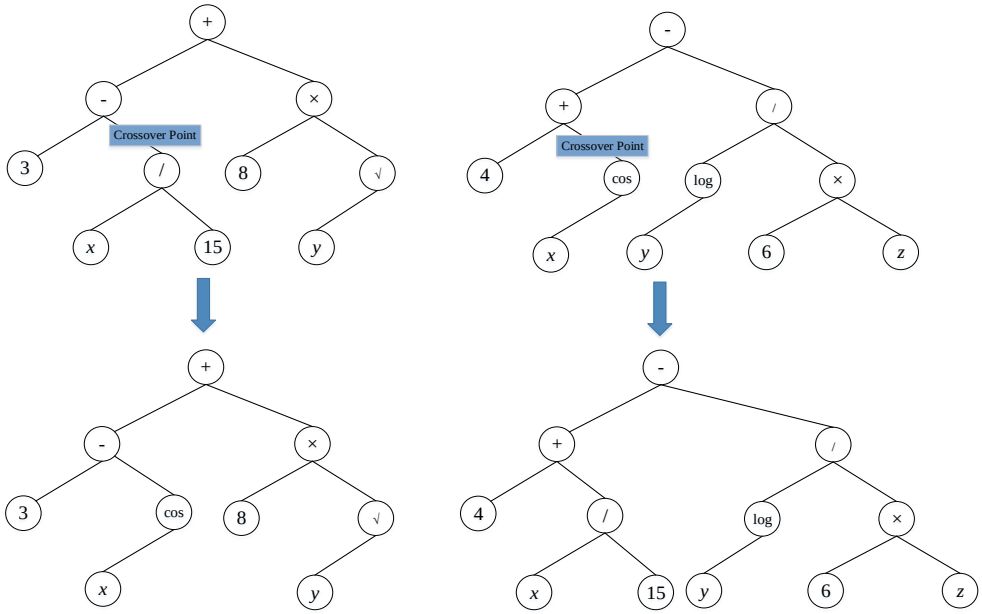



*Figure 2: Crossover in GP*

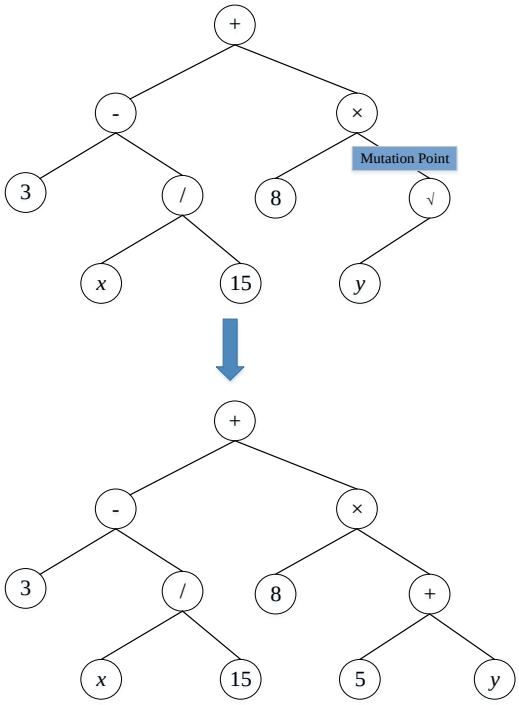



*Figure 3: Mutation in GP*

Based on the above definitions, the procedure of a simple GP algorithm can be illustrated in Figure 4 and briefly summarized as follows [17-19]:

1. Initialize a problem
   The necessary parameters of GP, including the population size, maximum size of programs, crossover rate, mutation rate etc., are first determined. According to a well-designed or random mechanism, the initial solutions (programs or individuals) of a population with the pre-specified population size are then generated. Generally speaking, the produced programs must be generated according to the limitation of a pre-specified maximum size for a feasible program. In addition, different sizes and shapes can be applied to these individuals.
2. Evaluate fitness
   Firstly, each program in the population is executed. The fitness (adaptability) corresponding to the program in the population is explicitly or implicitly assessed by means of measuring how well it can resolve the optimization problem through a pre-defined fitness function. There are various evaluation methods, e.g. the amount of error between its output and target, the total cost/time for bringing the system to a desired state, or the classification accuracy. The obtained evaluation result is defined as the fitness.
3. Create the next generation
   Based on the probability determined by the fitness corresponding to each program, some individuals can be selected to form a matching pool. The well-designed genetic operators are then applied to these selected individuals (programs), including:
   (1) Reproducing: This operator can duplicate the selected program to create a new individual.
   (2) Crossover: Two paired programs are first chosen from the matching pool randomly. Two chosen programs, called parents, are then recombined with random crossover points to form two new programs, called children, in the offspring generation.
   (3) Mutating: A part of a selected program is randomly chosen and mutate to produce a new offspring individual.
   (4) Altering architectures: Generate a new offspring program by altering the architecture of the selected program.
   After applying the above genetic operators, the programs in the offspring population replace the original individuals in the current population (the now-old generation) according to a certain strategy, such as the elitist strategy, thus creating a new population (the next generation).
4. Check the termination criteria
   The best-so-far individual, i.e. the best program ever encountered during the running process of GP, is designated as the final solution for the optimization problem when the stopping criteria can be satisfied. Otherwise, Steps 2 to 4 should be run iteratively until the termination criteria can be fulfilled.

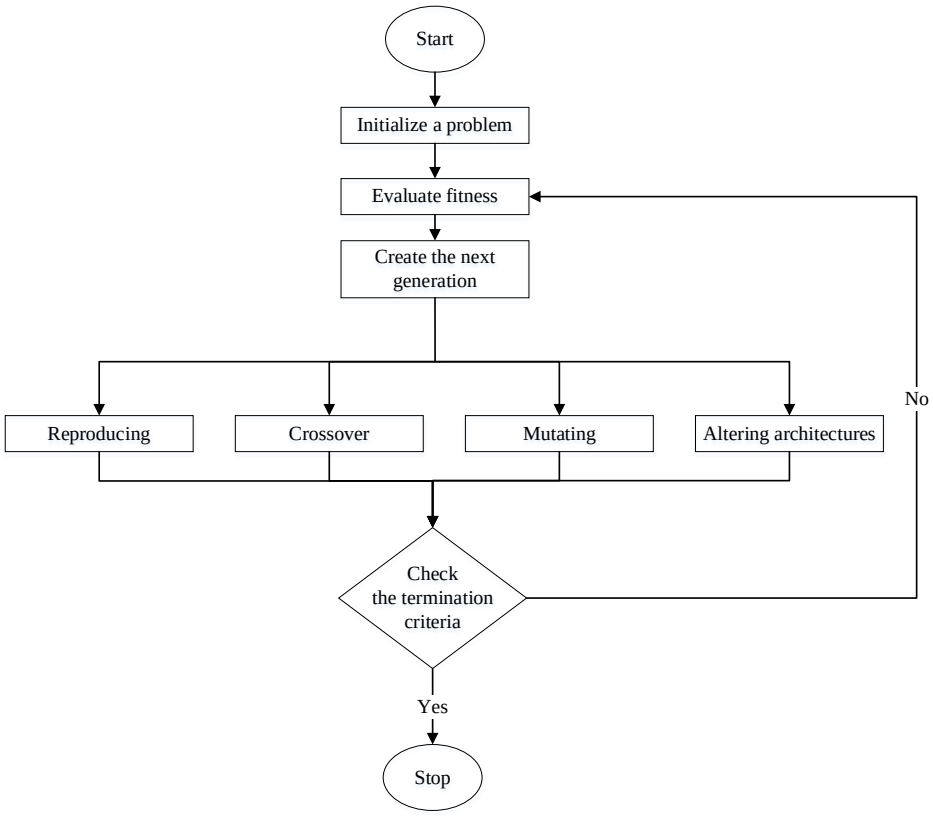



*Figure 4: The procedure of a simple GP*

***Feature Selection***

In some cases, the irrelevant and redundant variables should be removed to lower the influence of the noises existing in the original variables thus determining a minimal subset of variables, called feature selection. Through this way, the original problem can be resolved equally well, even better, than by using a full set consisting all of the original variables (features) [20]. By using the feature selection mechanism, researchers can avoid to select too many or too few features, i.e. more or less than necessary, to reduce the data size and complexity thus lowering the training time, as well as increasing the computational efficiency [21]. The valuable information contained in the data may be duplicated or shadowed while too many (irrelevant) features are selected. On the other hands, the information provided by the selected subset of features might be low if choosing too few features. There are two classifications for the approaches of feature selection: (1) the wrapper approach and (2) the filter approach. The wrapper approach for feature selection acts as a wrapper around the induction algorithm [22] and tries to find a good subset of features by evaluating features' subsets through the induction algorithm. That is, the searching process for the best subset of features utilizing and interacting the selection and induction algorithms simultaneously. However, a preprocessing methodology is utilized to screen out critical features, thus finding a good set of features in the filter approach. Therefore, the filter approach entirely ignores the performance of the selected subset's features on the induction algorithm. In other words, the feature selection algorithm neglects the induction algorithm and operates independently [22].

***Davies-Bouldin Index***

The Davies-Bouldin index (DBI) [23] is a metric for evaluating the performance regarding a clustering algorithm. The DBI is an internal evaluation scheme that validates how well the obtained clustering results can made by using the quantities and features inherent to the data. A good DBI value cannot guarantee to retrieve the best information from the original data. Given data points with $n$ dimensions, let $C_i$, whose feature vector is represented by $X_j$, be a cluster for these data points. The scatter within the cluster can be measured by $S_i$ defined as follows

$$S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i}\left|X_j - A_i\right|^p\right)^{1/p} \quad (6)$$

where $A_i$ and $T_i$ are the centroid of $C_i$ and the size of cluster $i$, respectively, and $p$ is a parameter that determines the distance metrics.

Usually, $p$ is set for 2 to measure the Euclidean distance between the data point and centroid of the cluster. There are various distance metrics can be used in the situations where Euclidean distance may not be the optimal measurement to determine the cluster for the data with a higher dimension. In addition, the distance metric applied in a problem must be in accord with the metric that uses in the clustering scheme. Then, a measure of separation between clusters $i$ and $j$ can be defined as

$$M_{i,j} = \left\|A_i - A_j\right\|_p = \left(\sum_{k=1}^{n}\left|a_{k,i} - a_{k,j}\right|^p\right)^{1/p} \quad (7)$$

where $a_{k,i}$ is the $k$th element of $A_i$. Notably, $M_{i,j}$ is the Euclidean distance measuring the centers of clusters $i$ and $j$ when $p$ is set for 2 essentially.

Davies and Bouldin defines an index to evaluate for the quality of a clustering scheme as follows [23]

$$DBI \equiv \frac{1}{N}\sum_{i=1}^{N} D_i \quad (8)$$

where $N$ is the total number of clusters, and $D_i$ is calculated by

$$D_i \equiv \max_{j\neq i} R_{i,j} \quad (9)$$

and

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad (10)$$

Notably, the DBI is dependent both on the data and algorithm. Next, a smaller DBI implies a better clustering result. In addition, a smaller $R_{i,j}$ indicates that the distances for the data points within clusters $i$ and $j$ are relatively smaller than the distance between the centroids of clusters $i$ and $j$, i.e. a better clustering result. Hence, the $D_i$ considers the worst case according to Equation (9).

**Proposed Approach**

This study utilizes the clustering method, feature selection technique and genetic programming (GP) to propose an approach for resolving the prediction problems as briefly depicted in Figure 5 and illustrated as follows:

**Step 1: Collect Data**

The required data are first collected according to the dependent output variable for a prediction problem, as well as all potential independent input variables corresponding to the output variable. Generally, the input and output variables have their own physical meaning with diverse scales. Therefore, each variable in the collected data must be identically normalized into the range between -1 and 1 based on the corresponding maximum and minimum values of that variable, thus avoiding that the variables with the wider measurement ranges will dominate the effects of variables which have relatively the narrower evaluation ambits. The normalized data are then divided into two parts including the training and test data groups according to ratio, e.g. 3:1, that is pre-determined by the users.

**Step 2: Cluster Data**

For the training data, each input variable along with the output variable are fed into the TwoStep cluster analysis, thus grouping the data into several clusters. The optimal number of clusters is determined based on the Schwarz's Bayesian Criterion (BIC) [13] or the Akaike Information Criterion (AIC) criterion [14].The clustering performance for the optimal clustering results is then evaluated to obtain its Davies-Bouldin index (DBI) [23] value. Therefore, a smaller DBI value implies a better clustering result, i.e. the data can be separated into clusters more appropriately. In other words, the causal relationship between the input and output variables is stronger, thus can simultaneously reduce the distances for the data in the same cluster as well as increase the distances among different clusters.

**Step 3: Group Input Variables**

The DBI values acquired in Step 2 are first ranked increasingly. Then, the incremental ratios, defined as a ratio of the DBI increment to its previous DBI value, are calculated except for the first, i.e. smallest, DBI value. Next, an analysis diagram is drawn where the cross axis is made up by the input variables that are arranged according to the increasing DBI increment ratios gradually, as well as the DBI incremental ratios corresponding to these ranked input variables form the vertical axis. The analysis diagram then can be divided into some sub-diagrams by observing the changes of DBI increment ratios. The input variables corresponding to the DBI increment ratios allocated in each sub-diagram create a group. Therefore, the input variables classified in the same group have the causal relationships with a similar degree to the output variable. That is, the input variables in the same group provide the similar importance in predicting the output variable. Furthermore, the DBI increment ratios in each group gradually increase. Hence, the input variables in the first and last groups have the highest and weakest influence on predicting the output variable, respectively.

**Step 4: Construct Final GP Models**

For each group determined in Step 3, the corresponding input variables are first made up by merging the input variables from that cluster as well as from all of the groups with smaller DBI increment ratios. The GP tool is then utilized to established the GP prediction model for each group determined in Step 3 with its corresponding input variables made up previously, along with the output variable. Therefore, the total number of the final GP prediction models is identical with the total number of groups decided in Step 3. Furthermore, the GP procedure is implemented several times since each execution result of GP, i.e. the candidate GP model, may vary. The criteria including the root-mean-square error (RMSE), R squared ($R^2$), and mean absolute percentage error (MAPE) are then used to evaluate the prediction performance of these candidate GP modes, thus selecting the model with the best prediction achievement as the final GP prediction model for each group selected in Step 3.

**Step 5: Select Features**

The prediction performance of the final GP models built in Step 4for the training data is compared to each other. In addition, the test data prepared in Step 1 are also fed into these final GP models established in Step 4 to measure their generalizability for dealing with the data that are never encountered before. Through balancing the prediction capability for simultaneously handling the known training data, as well as the unknown test data, the input variables that are critical to the output variable can be identified. In other words, the important features, i.e. input variables, to the response, i.e. output variable, can be screened out.

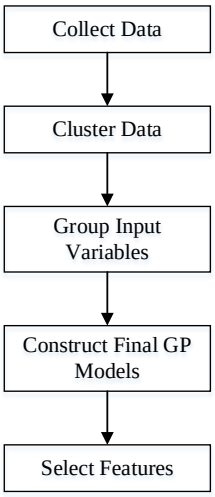



*Figure 5: Propose approach*

**Case Study of TRA**

In this section, the usefulness, effectiveness as well as efficiency of the proposed approach are demonstrated by presenting a case study on predicting the on-time percentages of local trains operated by the Taiwan Railway Administration (TRA) in Taiwan.

***TRA***

Taiwan Railway Administration (TRA), owned by the government of the Republic of China (ROC), is a company that operates the traditional railway transportation. In 1891, the first railway was constructed with 28.6km of track in Taiwan. Nowadays, the TRA operates twelve railway lines reaching a total length of 1065kmrailways as conceptually shown in Figure 6. Notably, the traffic of railways between Zhunan and Keelung stations is considered the busiest owing to the most part of population primarily congregates in the western areas of Taiwan, especially in the northwest region. Furthermore, the TRA must simultaneously operate multiple types of passenger trains, e.g. Tzu-Chiang Limited Express, Puyuma Express, Taroko Express, Chu-Kuang Express, Fu-Hsing Semi Express, Fast Local, and Local Trains etc., as well as freight trains on the same tracks. Next, Local Trains must often wait for Express or Fast Local Trains, that have the higher priority, to pass. Therefore, it becomes a difficult task for predicting the on-time percentages of Local Trains since the factors that can influence the on-time percentages of Local Trains are various and not easy to identified. So, this study focuses on the prediction problem regarding the on-time percentages of Local Trains operating in the region covering from Keelung station to Zhunan station.

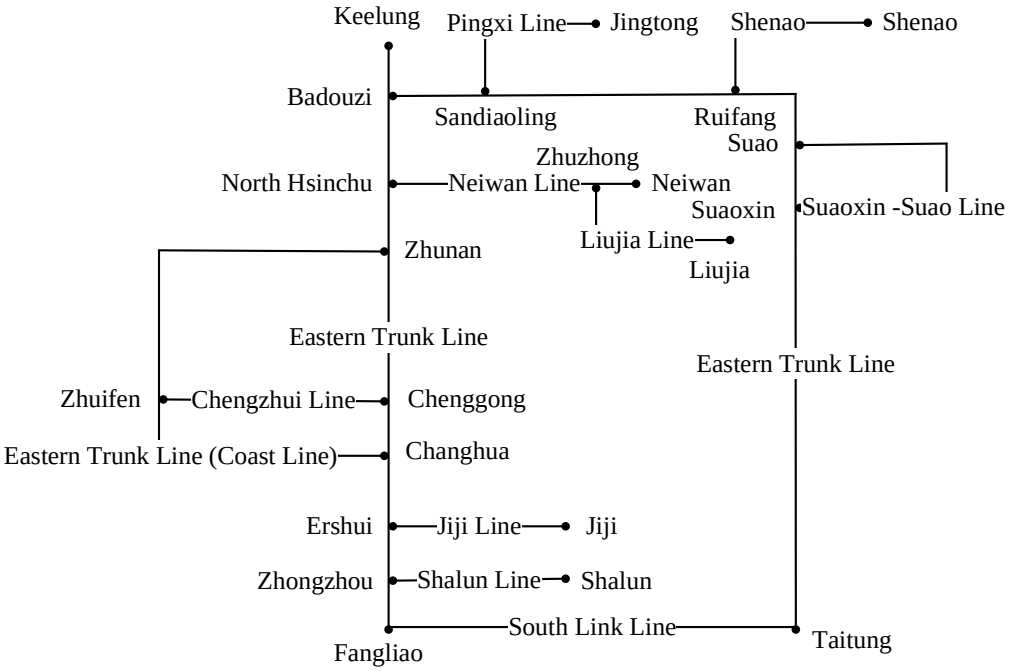



*Figure 6: Railways operated by TRA*

***Data Collection***

The response, i.e. dependent variable, in this study is the on-time percentage of a Local Train. Hence, the possible factors that might affect the response are first identified and determined as summarized in Table 1interpreted by referring to Figure 6. Notably, the TRA defines the on-time percentage of a Local Train as the total number of trains that can arrive at the terminal station on time divided by the total number of trains that arrive at the terminal station during an operating period, e.g. one month. Furthermore, a train that can arrive at the terminal station within five minutes of its original scheduled time is deemed to be on time. Furthermore, the categorical factor with three or more statuses must be encoded through several binary variables. For example, the second factorin Table 1 represents the operation day of a train and has four statuses, including (1) daily, (2) on Sundays, (3) daily except Sundays, and (4) daily except Saturdays, which are encoded by utilizing four

variables as (1,0,0,0), (0,1,0,0), (0,0,1,0), and (0,0,0,1). Therefore, forty independent variables are finally used to n code the thirty-three factors shown Table 1. For each Local Train, the forty independent variables, along with its corresponding on-time percentage, i.e. dependent variable, are then put in a row. There are 278 rows gathered from the operation data collected by the TRA on March 2019. In addition, each independent variable has its own physical meaning and measurement scale. For example, the independent variables encoding for the fifth and sixth factors are time and distance, as well as have measurement ranges of (100, 1400) and (2.1,130.9), respectively. Therefore, these variables are linearly and identically normalized into the range of (-1,1)according to its corresponding maximum and minimum values in order to prevent a variable with a large measurement range might dominates the effect of a variable, that has a relatively small range, on the dependent variable. Here, a ratio of 3:1 is used to randomly divide these normalized data of 278 rows into the two sets including training and test data sets with 209 and 69 rows, respectively.

**Table 1:**Possible factors for predicting the on-time percentages of Local Trains

| No. | Factors(Independent variables) | Explanation |
|---|---|---|
| 1 | Northbound/Southbound | The train is northbound/southbound. |
| 2 | Operation days | Train runs daily/on Sundays/daily except Sundays/daily except Saturdays. |
| 3 | Operation lines | Train runs via Eastern Trunk line/ Eastern Trunk Line (Coast Line)/Cross-Line (Mixed Lines). |
| 4 | Initial kilometer marker | Kilometer marker ("kmarker") of station from which train initiates. Zhunan station kmarker is set at 0. Kmarker of each station is calculated based on distance from Zhunan station. |
| 5 | Initial time | Time when train sets out. |
| 6 | Train-kilometers | Distance of train's trip. |
| 7 | Traveling time | Duration of train's trip. |
| 8 | Train-kilometers before train enters region | Northern point at which Western trunk and Western trunk line (coast line) merge is Zhunang station. Eastern trunk line starts at Badouzi station. Trains not setting out from points between Zhunang and Keelung stations have run for a distance before entering region between Zhunang and Badouzi stations. This distance is defined as "Train-kilometers before train enters region". |
| 9 | Travel time before train enters region | Similar to explanation in previous factor. Trains not initiating between Zhunang and Keelung stations have run for a duration before entering region between Zhunang and Badouzi stations. This duration is defined as "Travel time before train enters region". |
| 10 | Number of waiting times before train enters the region | Similar to situation explained in factor 8. Trains not initiating between Zhunang and Keelung stations must wait for Express/Fast Local Train several times before it enters region between Zhunang and Badouzi stations. This *number of* waiting times is defined as "Number of waiting times before train enters region". |
| 11 | Waiting time before train initiates | Similar to situation explained in factor 8. A train not initiating between Zhunang and Keelung stations must wait for Express/Fast Local Train for a duration before it enters region between Zhunang and Badouzi stations. This waiting *duration* is defined as "Waiting time before train initiates". |
| 12 | Train-kilometers after train leaves | Similar to situation explained in factor 8. A train not |

| No. | Factors(Independent variables) | Explanation |
|---|---|---|
| | region | terminating between Zhunang and Keelung stations must run for a distance after leaving region between Zhunang and Badouzi stations. This distance is defined as "Train-kilometers after train leaves region". |
| 13 | Traveling time after train leaves region | Similar to situation explained in factor 8. A train not terminating between Zhunang and Keelung stations must run for a duration after it leaves region between Zhunang and Badouzi stations. This duration is defined as "Traveling time after train leaves region". |
| 14 | Number of waiting times after train leaves region | Similar to situation explained in factor 8. A train that does not terminate between Zhunang and Keelung stations must wait for Express/Fast Local Train several times after it leaves region between Zhunang and Badouzi stations. This number of waiting times is defined as "Number of waiting times after train leaves region". |
| 15 | Waiting time after train leaves region | Similar to situation explained in factor 8. A train that does not initiate between Zhunang and Keelung stations must wait for Express/Fast Local Train for some duration after it leaves region between Zhunang and Badouzi stations. This waiting duration is defined as "Waiting time after train leaves region". |
| 16 | Number of waiting times during operating (Tzu-Chiang Limited Express) | The total number of delays due to waiting for Tzu-Chiang Limited Express to pass while running between Zhunang and Keelung stations. |
| 17 | Number of passing trains during operating (Tzu-Chiang Limited Express) | Total number of passing Tzu-Chiang Limited Express trains while a train runs between Zhunang and Keelung stations. |
| 18 | Operating distance of awaited Express (Tzu-Chiang Limited Express) | Total distance which passing Tzu-Chiang Limited Express trains have run beforeTzu-Chiang Limited Express passes a train that runs between Zhunang and Keelung stations. |
| 19 | Waiting time of a train (Tzu-Chiang Limited Express) | The total time a train must wait for Tzu-Chiang Limited Express to pass while running between Zhunang and Keelung stations. |
| 20 | Number of waiting times during operating (Puyuma/Taroko Express) | The total number of delays due to waiting for Puyuma/Taroko Express to pass while running between Zhunang and Keelung stations. |
| 21 | Number of passing trains during operating (Puyuma/Taroko Express) | Total number of passing Puyuma/Taroko Express trains while a train runs between Zhunang and Keelung stations. |
| 22 | Operating distance of awaited Express (Puyuma/Taroko Express) | Total distance passing Puyuma/Taroko Express trains have run before Puyuma/Taroko Express passes a train that runs between Zhunang and Keelung stations. |
| 23 | Waiting time of a train (Puyuma/Taroko Express) | Total time a train must wait for Puyuma/Taroko Express to pass while running between Zhunang and Keelung stations. |
| 24 | Number of waiting times during operating (Chu-Kuang Express) | Total number of delays while a train must wait for Chu-Kuang Express to pass while running between Zhunang and Keelung stations. |
| 25 | Number of passing trains during operating (Chu-Kuang Express) | Total number of passing Chu-Kuang Express trains while a train runs between Zhunang and Keelung stations. |

| No. | Factors(Independent variables) | Explanation |
|---|---|---|
| 26 | Operating distance of awaited Express (Chu-Kuang Express) | Total distance the passing Chu-Kuang Express trains have run before Chu-Kuang Express pass a train that runs between Zhunang and Keelung stations. |
| 27 | Waiting time of a train (Chu-Kuang Express) | Total waiting time for a train while waiting for Chu-Kuang Express to pass while running between Zhunang and Keelung stations. |
| 28 | Number of waiting times during operating (Fast Local Train) | Total number of delays while a train must wait for Fast Local Train to pass while running between Zhunang and Keelung stations. |
| 29 | Number of passing trains during operating (Fast Local Train) | Total number of passing Fast Local Trains while a train runs between Zhunang and Keelung stations. |
| 30 | Operating distance of awaited Express (Fast Local Train) | Total distance which passing Fast Local Trains have run before Fast Local Trains pass a train that runs between Zhunang and Keelung stations. |
| 31 | Waiting time of a train (Fast Local Train) | Total waiting time for a train waiting for Fast Local Train to pass while running between Zhunang and Keelung stations. |
| 32 | Number of waiting times for awaiting one train | Total number of delays while a train must wait for any type of train while running between Zhunang and Keelung stations. |
| 33 | Number of waiting times for successively awaiting two trains | The total number of delays when a train must successively wait for two of any type of train while running between Zhunang and Keelung stations. |

***Data Clustering***

For the training data set, each independent variable along with the output variable are fed into the TwoStep cluster analysis for further clustering the original data to yield data groups with enough diversity. In this study, the SPSS software is used to implement the TwoStep cluster analysis where the distance between two items are evaluated by the likelihood measure, as well as the BIC [13] clustering criterion is applied. The optimal number of clusters is determined automatically by the SPSS software, and the DBI index [23] is utilized to evaluate the clustering performance of these optimal clustering results as summarized in Table 2.

**Table 2:** Clustering results of TwoStep cluster analysis

| Independent variables | Total number of clusters | DBI |
|---|---|---|
| *x*0-1 | 3 | 0.716205 |
| *x*0-2 | 3 | 0.716205 |
| *x*1-1 | 2 | 0.278191 |
| *x*1-2 | 2 | 0.320441 |
| *x*1-3 | 3 | 0.569602 |
| *x*1-4 | 2 | 0.564948 |
| *x*2-1 | 3 | 0.773055 |
| *x*2-2 | 2 | 0.566315 |
| *x*2-3 | 3 | 0.479828 |
| *x*2-4 | 4 | 0.532615 |
| *x*3 | 3 | 1.067848 |
| *x*4 | 2 | 0.829609 |
| *x*5 | 3 | 0.716585 |
| *x*6 | 3 | 0.727437 |
| *x*7 | 3 | 0.658422 |
| *x*8 | 3 | 0.675174 |
| *x*9 | 3 | 0.643617 |

| | | |
|---|---|---|
| *x*10 | 3 | 0.694348 |
| *x*11 | 3 | 0.760276 |
| *x*12 | 2 | 0.974690 |
| *x*13 | 3 | 0.677692 |
| *x*14 | 3 | 0.757923 |
| *x*15 | 3 | 0.510254 |
| *x*16 | 3 | 0.510254 |
| *x*17 | 3 | 0.615470 |
| *x*18 | 3 | 0.606033 |
| *x*19 | 2 | 0.763000 |
| *x*20 | 2 | 0.949367 |
| *x*21 | 3 | 0.734735 |
| *x*22 | 2 | 0.931355 |
| *x*23 | 2 | 0.562308 |
| *x*24 | 3 | 0.562308 |
| *x*25 | 2 | 0.561943 |
| *x*26 | 2 | 0.647475 |
| *x*27 | 2 | 0.563917 |
| *x*28 | 2 | 0.563917 |
| *x*29 | 2 | 0.563917 |
| *x*30 | 2 | 0.563917 |
| *x*31 | 4 | 0.806230 |
| *x*32 | 2 | 0.560262 |

***Input Variables Grouping***

The DBI values in Table 2 are first ranked increasingly as shown in Table 3(A). The DBI incremental ratios are then calculated as summarized in Table 3(B) and depicted in Figure 7(A) except for the first independent variable *x*1-1.Notably, the DBI incremental ratio is set as 0 for *x*1-1. By observing Table 3(B), the DBI incremental ratio of independent variable *x*2-3 is relatively much larger than the DBI incremental ratio regarding independent variable *x*1-2. For exploring more easily, the DBI incremental ratios are diagramed again by moving the variable *x*2-3 as exhibited in Figure 7(B). Through observing the changes of DBI increment ratios, Figure 7(B) is analyzed and divided into six sub-diagrams. Therefore, the input variables are classified into six groups as shown in Table 3(B).

**Table 3:** Ranking DBI values and incremental ratios

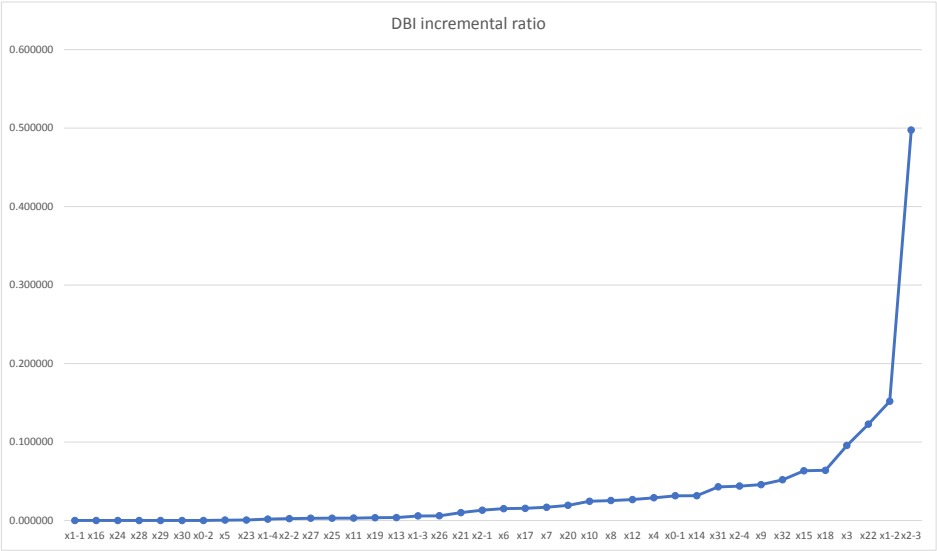
(A)

| Independent variables | DBI values |
|---|---|
| *x*1-1 | 0.278191 |
| *x*1-2 | 0.320441 |
| *x*2-3 | 0.479828 |
| *x*15 | 0.510254 |
| *x*16 | 0.510254 |
| *x*2-4 | 0.532615 |
| *x*32 | 0.560262 |
| *x*25 | 0.561943 |
| *x*23 | 0.562308 |
| *x*24 | 0.562308 |
| *x*27 | 0.563917 |
| *x*28 | 0.563917 |
| *x*29 | 0.563917 |

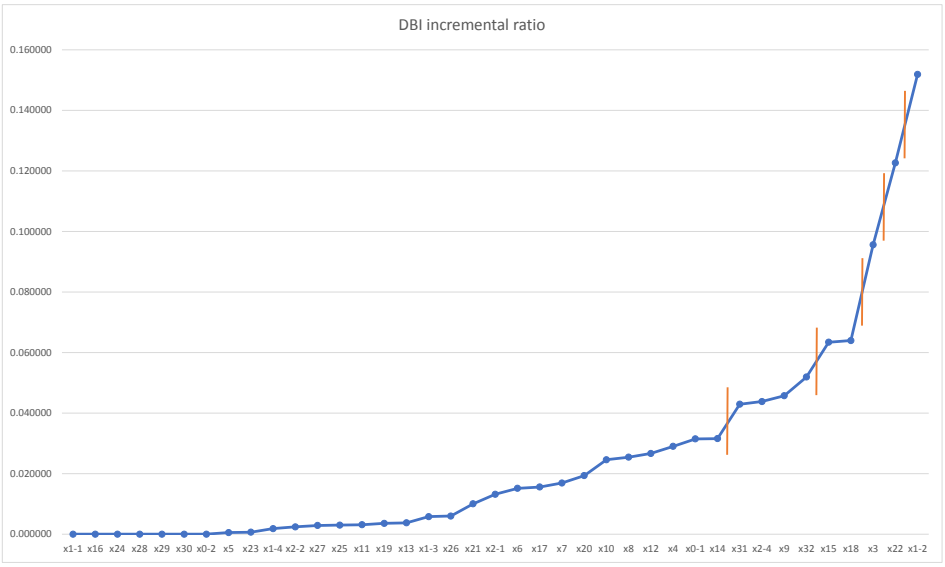
(B)

| Groups | Independent variables | DBI incremental ratios |
|---|---|---|
| 1 | *x*1-1 | 0.000000 |
| | *x*16 | 0.000000 |
| | *x*24 | 0.000000 |
| | *x*28 | 0.000000 |
| | *x*29 | 0.000000 |
| | *x*30 | 0.000000 |
| | *x*0-2 | 0.000000 |
| | *x*5 | 0.000531 |
| | *x*23 | 0.000648 |
| | *x*1-4 | 0.001828 |
| | *x*2-2 | 0.002421 |
| | *x*27 | 0.002862 |

| | | |
|---|---|---|
| | *x*30 | 0.563917 |
| | *x*1-4 | 0.564948 |
| | *x*2-2 | 0.566315 |
| | *x*1-3 | 0.569602 |
| | *x*18 | 0.606033 |
| | *x*17 | 0.615470 |
| | *x*9 | 0.643617 |
| | *x*26 | 0.647475 |
| | *x*7 | 0.658422 |
| | *x*8 | 0.675174 |
| | *x*13 | 0.677692 |
| | *x*10 | 0.694348 |
| | *x*0-1 | 0.716205 |
| | *x*0-2 | 0.716205 |
| | *x*5 | 0.716585 |
| | *x*6 | 0.727437 |
| | *x*21 | 0.734735 |
| | *x*14 | 0.757923 |
| | *x*11 | 0.760276 |
| | *x*19 | 0.763000 |
| | *x*2-1 | 0.773055 |
| | *x*31 | 0.806230 |
| | *x*4 | 0.829609 |
| | *x*22 | 0.931355 |
| | *x*20 | 0.949367 |
| | *x*12 | 0.974690 |
| | *x*3 | 1.067848 |
| | *x*25 | 0.003002 |
| | *x*11 | 0.003105 |
| | *x*19 | 0.003583 |
| | *x*13 | 0.003731 |
| | *x*1-3 | 0.005804 |
| | *x*26 | 0.005994 |
| | *x*21 | 0.010033 |
| | *x*2-1 | 0.013178 |
| | *x*6 | 0.015143 |
| | *x*17 | 0.015571 |
| | *x*7 | 0.016907 |
| | *x*20 | 0.019339 |
| | *x*10 | 0.024577 |
| | *x*8 | 0.025442 |
| | *x*12 | 0.026673 |
| | *x*4 | 0.028997 |
| | *x*0-1 | 0.031479 |
| | *x*14 | 0.031559 |
| | *x*31 | 0.042915 |
| 2 | *x*2-4 | 0.043823 |
| | *x*9 | 0.045733 |
| | *x*32 | 0.051908 |
| 3 | *x*15 | 0.063411 |
| | *x*18 | 0.063959 |
| 4 | *x*3 | 0.095577 |
| 5 | *x*22 | 0.122644 |
| 6 | *x*1-2 | 0.151874 |
| 7 | *x*2-3 | 0.497396 |




(A)




*(B)*

*Figure 7: Diagram of DBI incremental ratios*

***Final GP Models Construction***

First, the GP technique is applied to the training data to build a GP prediction model where the independent variables allocated in the first group act as the input variables, as well as the output variable is the on-time percentages of Local Trains. The Discipulus GP software with its default parameter settings of the population size, crossover rate, and mutation rate of 500, 0.5, and 0.95, respectively, is used in this study. Furthermore, the GP procedure is implemented for ten times, and the criteria including the root-mean-square error (RMSE), R squared ($R^2$), and mean absolute percentage error (MAPE) are used to evaluate the prediction performance of the obtained GP models. Therefore, the model with the best prediction achievement can be selected as the final GP prediction model. Similarly, the same procedure is implemented for the independent variables in other groups. Notably, the input variables should include all the independent variables clustered in the $C_{th}$, $(C\text{-}1)_{th}$,…, second, and first groups for the case corresponding to the $C_{th}$ group. Hence, all the independent variables serve as the input variables for the case corresponding to the $7_{th}$ group. Table 4 summarizes the GP implementation results. Through simultaneously considering RMSE, $R^2$, and MAPE, the optimal execution result among ten runs for each group is then determined and marked by a gray background. Based on Table 4, GP can construct prediction models with prediction performance of sufficient low RMSE and MAPE, as well as high $R^2$.

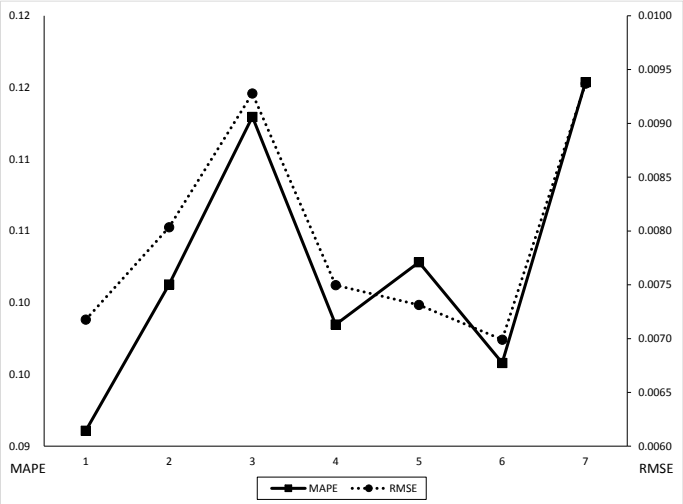
**Table 4:** GP implementation results

| Execution No. | Group#1 | | | Group#2 | | | Group#3 | | | Group#4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE |
| 1 | 0.013836 | 0.62988 | 0.145809 | 0.011505 | 0.65356 | 0.129060 | 0.011833 | 0.68814 | 0.138268 | 0.010814 | 0.71868 | 0.129896 |
| 2 | 0.010347 | 0.70844 | 0.122597 | 0.011835 | 0.70211 | 0.135812 | 0.013305 | 0.65056 | 0.142939 | 0.011216 | 0.71059 | 0.133016 |
| 3 | 0.012933 | 0.67024 | 0.146812 | 0.008190 | 0.78588 | 0.110302 | 0.012836 | 0.66465 | 0.143376 | 0.007497 | 0.78909 | 0.098480 |
| 4 | 0.011266 | 0.67024 | 0.130585 | 0.008035 | 0.76787 | 0.101271 | 0.009673 | 0.72001 | 0.116656 | 0.008507 | 0.75935 | 0.103065 |
| 5 | 0.011153 | 0.67718 | 0.128805 | 0.011286 | 0.69021 | 0.130339 | 0.011002 | 0.70967 | 0.129201 | 0.010694 | 0.75111 | 0.131200 |
| 6 | 0.013900 | 0.59062 | 0.144334 | 0.011749 | 0.69682 | 0.137460 | 0.012252 | 0.68710 | 0.137670 | 0.012111 | 0.71358 | 0.139576 |
| 7 | 0.010250 | 0.72293 | 0.121659 | 0.014006 | 0.63484 | 0.152986 | 0.014114 | 0.61574 | 0.148565 | 0.009437 | 0.73245 | 0.110783 |
| 8 | 0.013967 | 0.62069 | 0.148135 | 0.010969 | 0.73665 | 0.130765 | 0.011789 | 0.72190 | 0.138100 | 0.011931 | 0.71680 | 0.140039 |
| 9 | 0.013022 | 0.66806 | 0.143739 | 0.009488 | 0.75965 | 0.121329 | 0.009278 | 0.74932 | 0.112958 | 0.011341 | 0.68344 | 0.127621 |
| 10 | 0.007177 | 0.81168 | 0.091088 | 0.013372 | 0.63413 | 0.146231 | 0.009984 | 0.72291 | 0.119170 | 0.008874 | 0.73225 | 0.107559 |
| Mean | 0.011785 | 0.67700 | 0.132356 | 0.011043 | 0.70617 | 0.129556 | 0.011607 | 0.69300 | 0.132690 | 0.010242 | 0.73073 | 0.122123 |
| Standard Deviation | 0.002176 | 0.06170 | 0.017765 | 0.001978 | 0.05513 | 0.015526 | 0.001611 | 0.04020 | 0.012463 | 0.001567 | 0.02967 | 0.015564 |
| CV | 0.184642 | 0.09114 | 0.134223 | 0.179111 | 0.07807 | 0.119837 | 0.138758 | 0.05800 | 0.093922 | 0.152992 | 0.04060 | 0.127446 |
| Execution No. | Group#5 | | | Group#6 | | | Group#7 | | | | | |
| | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | | | |

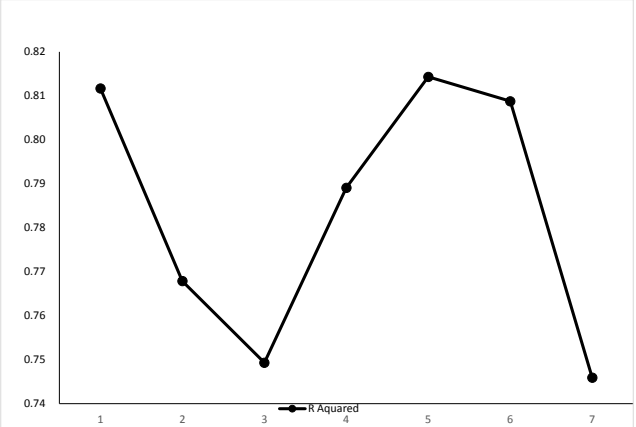
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.012696 | 0.63399 | 0.134627 | 0.009885 | 0.72090 | 0.112719 | 0.009371 | 0.74592 | 0.115382 |
| 2 | 0.013123 | 0.65491 | 0.143890 | 0.007944 | 0.77425 | 0.101446 | 0.010948 | 0.68067 | 0.120900 |
| 3 | 0.012224 | 0.69663 | 0.145909 | 0.013263 | 0.61839 | 0.139994 | 0.013083 | 0.64761 | 0.140240 |
| 4 | 0.009551 | 0.75100 | 0.116888 | 0.012139 | 0.68374 | 0.136995 | 0.011590 | 0.68738 | 0.134680 |
| 5 | 0.011556 | 0.68739 | 0.134262 | 0.011134 | 0.70799 | 0.129542 | 0.012501 | 0.66607 | 0.134812 |
| 6 | 0.011915 | 0.68534 | 0.133429 | 0.011005 | 0.69424 | 0.124697 | 0.013922 | 0.57905 | 0.143096 |
| 7 | 0.008783 | 0.74595 | 0.106330 | 0.010169 | 0.70485 | 0.119830 | 0.010303 | 0.69420 | 0.120768 |
| 8 | 0.008899 | 0.74561 | 0.111414 | 0.006991 | 0.80876 | 0.095810 | 0.010202 | 0.70375 | 0.114451 |
| 9 | 0.007313 | 0.81431 | 0.102828 | 0.011821 | 0.67886 | 0.128507 | 0.011488 | 0.69973 | 0.131878 |
| 10 | 0.012441 | 0.63857 | 0.134211 | 0.012136 | 0.67224 | 0.133679 | 0.012448 | 0.64978 | 0.135004 |
| Mean | 0.010850 | 0.70537 | 0.126379 | 0.010649 | 0.70642 | 0.122322 | 0.011586 | 0.67542 | 0.129121 |
| Standard Deviation | 0.002025 | 0.05785 | 0.015634 | 0.001958 | 0.05336 | 0.014869 | 0.001421 | 0.04434 | 0.010359 |
| CV | 0.186627 | 0.08201 | 0.123711 | 0.183894 | 0.07554 | 0.121554 | 0.122617 | 0.06565 | 0.080226 |

***Feature Selection***

Figure 8 graphs the prediction performance of selected optimal execution results for all groups based on Table 4. Compared to other groups, the thirty (30) input variables in the first group can provide the least RMSE and MAPE according to Figure 8. In addition, the $R^2$ofthe model built by the input variables within the first to seventh groups, i.e. all forty (40) input variables, is the highest. However, the difference between the prediction performance $R^2$ regarding the models, constructed by all input variables and input variables in the first group, is only 0.26% that is extremely small. The entrance of input variables cannot necessarily enhance the prediction performance. In other words, these entranced variables are surplus thus making noise and disturbing the building process of prediction models. Therefore, the aim of feature selection, i.e. selecting the input variables that are critical to predict the output variable, has been reached. Furthermore, gathering the information regarding fewer input variables, instead of all input variables, to establish the prediction model is more economic and faster.




(A) MAPE and RMSE




(B) $R^2$

*Figure 8: The prediction performance of selected optimal execution results*

**Conclusions**

The on-time percentages of trains are valuable information for making appropriate timetables, arranging trains' waiting, determining the running tracks for trains, as well as setting up sufficient manpower etc. Therefore, the prediction of on-time percentages of trains is critical to the operation of trains. However, this is a very complex and difficult task since the on-time percentages of trains are influenced by various factors that are not easily identified and considered completely. Furthermore, the unnecessary input variables might not be helpful, thus making noise and lowing the performance of prediction models. Hence, the clustering technique, feature selection and genetic programming (GP) are utilized in this study to propose a prediction procedure for tackling such a complicated problem. The clustering method along with evaluating clustering performance by Davies-Bouldin index (DBI) is applied to group the data and select featured, i.e. critical, input variables. The GP technique is used to construct prediction models with the independent variables determined based on the grouping results. A case study aiming to predict the on-time percentages of Local Trains of TRA (Taiwan Railway Administration) demonstrates usefulness, effectiveness, and efficiency of the proposed approach. The implementation results show that the clustering method collocated with DBI can group the input variables into several clusters that have different importance to the output variable. Next, the GP can construct adequate models for predicting the on-time percentages of local trains. Furthermore, the comparison also shows that the feature selection process can balance the accuracy and complexity of a GP prediction model. Therefore, the researchers can pay more attention and time to these critical input variables, as well as the operation of trains. The time for gathering and analyzing the data can also be saved. Hence, the prediction procedure proposed in our study can be considered as useful, effective, and efficient for solving the prediction problems in the real world.

**Acknowledgments**

The author would like to thank Jackson Liu in the Secretariat of Taiwan Railway Administration (TRA) in Taiwan for fully supporting this study, as well as thank the partial support of Minghsin University of Science and Technology, Taiwan, R.O.C. under Contract No. MUST-110BA-01.

**References**

[1]. Yan, X. D., Richards, S., & Su, X. G. (2010). Using hierarchical tree-based regression model to predict train-vehicle crashes at passive highway-rail grade crossings. *Accident Analysis and Prevention*, *42*(1), 64-74.

[2]. Cao, Y, Xia, H., & Li, Z. (2012). A semi-analytical/FEM model for predicting ground vibrations induced by high-speed train through continuous girder bridge. *Journal of Mechanical Science and Technology*, *26*(8), 2485-2496.

[3]. Cai, G., Yao, D., Sun, J., Jia, D., & Chen, J. (2015). Predict of high-speed train's safe operation based on fuzzy inference. *Journal of Coastal Research*, *73*, 792-796.
[4]. Iglesias, E. L., Thompson, D. J., & Smith, M. G. (2017). Component-based model to predict aerodynamic noise from high-speed train pantographs. *Journal of Sound and Vibration*, *394*, 280-305.
[5]. Lai, Q., Liu, J., Luo, Y., & Ma, M. (2017). A hybrid short-term forecasting model of passenger flow on high-speed rail considering the impact of train service frequency. *Mathematical Problems in Engineering*, *2017*, Article ID 1828102.
[6]. Shan, Y., Zhou, S. H., Zhou, H. C., Wang, B. L., Zhao, Z. C., Shu, Y, & Yu, Z. (2017). Iterative method for predicting uneven settlement caused by high-speed train loads in transition-zone subgrade. *Transportation Research Record*, *2607*, 7-14.
[7]. Jin, G. W., He, S. W., Li, J. B., Li, Y. B., Guo, X. L., & Xu, H. F. (2019). An integrated model for demand forecasting and train stop planning for high-speed rail. *Symmetry-Basel*, *11*(5), Article ID 720.
[8]. Kang, R. W., Wang, J. F., Cheng, J. F., Chen, J. Q., & Pang, Y. Z. (2019). Intelligent forecasting of automatic train protection system failure rate in China high-speed railway. *Eksploatacja I Niezawodnosc-Maintenance and Reliability*, *21*(4), 567-576.
[9]. Lessan, J., Fu, L. P., Wen, C, (2019). A hybrid Bayesian network model for predicting delays in train operations. *Computers & Industrial Engineering*, *127*, 1214-1222.
[10]. Liu, Y. M., Qiao, N. G., Zhao, C. C., Zhuang, J. J., & Tian, G. D. (2019). Using the AR-SVR-CPSO hybrid model to forecast vibration signals in a high-speed train transmission system. *Proceedings of the Institution of Mechanical Engineers Part F-Journal of Rail and Rapid Transit*, *233*(7), 701-714.
[11]. Huang, P., Lessan, J, Wen, C., Peng, Q., Fu, L., Li, L., & Xu, X. (2020). A Bayesian network model to predict the effects of interruptions on train operations. *Transportation Research Part C: Emerging Technologies*, *114*, 338-358.
[12]. Zhang, L. K., Feng, X. S., Ding, C. C., & Liu, Y. (2020). Mitigating errors of predicted delays of a train at neighbouring stops. *IET Intelligent Transport Systems*, *14*(8), 873-879.
[13]. Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, New York: Springer-Verlag.
[14]. Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, *21*, 243-247.
[15]. Holland, J. H. (1975).*Adaptation in Nature and Artificial Systems*, Ann Arbor, MI: The University of Michigan Press.
[16]. Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Cambridge, Mass: MIT Press.
[17]. Koza J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., & Lanza, G. (2005).*Genetic Programming IV: Routine Human-Competitive Machine Intelligence*, New York: Springer.
[18]. Ciglarič, I., & Kidrič, A. (2006). Computer-aided derivation of the optimal mathematical models to study gear-pair dynamic by using genetic programming. *Structural and Multidisciplinary Optimization*, *32*(2), 153-160.
[19]. Koza, J. R., Streeter, M. J., & Keane, M. A. (2008). Routine high-return human-competitive automated problem-solving by means of genetic programming. *Information Sciences*, *178*(23), 4434-4452.
[20]. Liu H., & Motoda H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*, Boston: Kluwer.
[21]. Piramuthu, S. (2004). Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Researc*h, *156*(2), 483-494.
[22]. Kohavi R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273-324.
[23]. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224-227.