Forecasting of the Broadband Penetration with Genetic Programming method and diffusion models

A Short Research in OECD countries

K.Salpasaranis (*Author*)
Electrical and Computer Engineering Department
University of Patras
Rio-Patras, Greece
salpk@upatras.gr

V.Stylianakis (*Author*)
Electrical and Computer Engineering Department
University of Patras
Rio-Patras, Greece
stylian@upatras.gr

Abstract— This paper presents the implementation of a modified Genetic Programming (GP) method in forecasting fixed broadband telecommunications penetration percentage in Organisation for Economic Co-operation and Development (OECD) countries. The specific GP method combines the use of known diffusion models for technological forecasting purposes, such as Logistic, Gompertz and Bass and the GP. The combination method produces both time dependant and causal models with high performance statistical indicators. Also, multiple approaches to forecasting can be implemented, mainly with no big datasets.

Keywords— Genetic Programming; diffusion models; broadband; penetration;

I. Introduction

Many methods have been focused in forecasting the penetration of new technology in a market [1]. A subcategory of these methods is the diffusion models category [2]. The diffusion models are dynamic models that follow a well defined curve in time. In this paper the models which have been used are the Gompertz, Logistic and Bass for the OECD countries [3][4][5]. The concept implemented here was initially introduced in [2] while in [3] a hybrid GP (hGP) method made possible the combination of diffusion models and GP to produce several forecasting models.

The Genetic Programming method (GP) is an evolutionary programming method derived by Genetic Algorithm (GA). It is a heuristic method that simulates the biological natural selection of an appropriate problem solution [6], [7]. In this paper, the independent variables besides time are macroeconomic statistics indices such as Gross Domestic Product per Capita (GDPpC) and Consumer Price Index (CPI).

II. Diffusion Models

A. General

In [1], [2] and [3] the diffusion process has been extensively described. The diffusion models follow:

$$\frac{dy(t)}{dt} = [S - y(t)] \cdot f(t) \qquad (1)$$

Where ***y(t)*** is the penetration for time ***t***, ***S*** is the saturation level and function ***f(t)*** is the coefficient.[2][3]

B. Gompertz Model

Two types of Gompertz models are described by the following:

$$y(t) = S \cdot e^{e^{f(t)}} \qquad (2)$$

$$y(t) = S \cdot e^{e^{f(t)}} + c \qquad (3)$$

In (2), which is called Gompertz I and (3), which is called Gompertz II, ***y***(***t***) is the estimated diffusion level at time *t*, parameter ***S*** is the saturation level. Parameter ***c*** corresponds to a constant. [1][2][3]

C. Logistic Diffusion Model

In general the form of the logistic model is:

$$y(t) = \frac{S}{1 + e^{f(t)}} \qquad (4)$$

In (4), ***y***(***t***) is the penetration of a product in a market, at time ***t***, ***f***(***t***)=***a***+***b·t*** is function of time and ***a, b*** are constants. [1][2][3]

D. BassModel

Bass model has been also presented in [1], [2] and [3]. In this model, the adoption of a new technology has two major categories: innovators and imitators.

The adoption of a new technology ***y***(***t***) is presented in (5):

$$y(t) = \frac{A - C \cdot e^{-B \cdot t}}{1 + D \cdot e^{-B \cdot t}} \qquad (5)$$

In (5), parameter ***A*** depicts the initial purchasers of the new product. Parameter ***B***=***p***+***q***, where ***p*** is innovators coefficient and ***q*** is imitators coefficient. Parameter **C=r·p**, where ***r*** is constant and ***D***=***r***· (***q/A***)**.** [3] [5]

III. Modified Genetic programming method

Koza introduced GP in [7]. In [7] each solution of a problem corresponds to a program which is called chromosome. In GP method the chromosome is presented as a tree, while in GA the chromosome is a string of numbers. The terminology of trees has been presented in [2] and [3]. The main terms are the node, which is a function and leaf which correspond to variables or constants of the solution [8] [9].

A. Flowchart of Modified GP Method

In this paper the modified hGP [3] will be implemented in a dataset. The modified hGP consists of three parts, the regression analysis, the genetic algorithm process and the model selection. The Fig.1 shows the parts of the hGP.

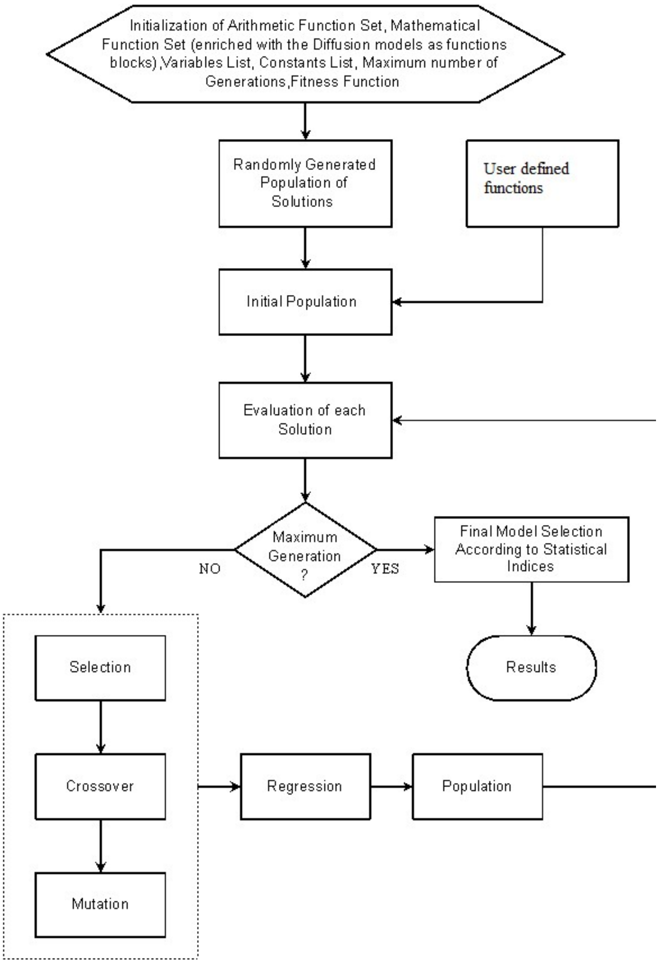
Initialization of Arithmetic Function Set, Mathematical Function Set (enriched with the Diffusion models as functions blocks),Variables List, Constants List, Maximum number of Generations,Fitness Function
Randomly Generated Population of Solutions
User defined functions
Initial Population
Evaluation of each Solution
Maximum Generation ?
NO
YES
Final Model Selection According to Statistical Indices
Results
Selection
Crossover
Mutation
Regression
Population

Fig. 1. Flowchart of the modified hGP

B. Initial population

In [3] the format of the randomly created chromosomes in the initial population of the solutions was introduced. So the blocks form for the population' construction was:

<Constant of S set><variable of M set><arithmetic function of F_A set><mathematical function of F_M set><variable of M set>

In the block F_A={+,-,*,/}, F_M={*exp, ln, log, sin, cos, Logistic, GompertzI, GompertzII, Bass*}, *S*={1, *random from -100 to 100*}, *M*={*t, GDPpC, CPI*}.

In generation 0 or initial population, one can see random chosen functions (F_A-arithmetic, F_M - mathematical, diffusion models), variables, constants and user defined blocks. The models are optimized by regression analysis [2].

C. Solutions Representation

Each chromosome is a string of characters [8]. In [3], the inner representation has the abstract syntax tree of Python Programming Language as parse trees. For example, the two chromosomes, namely, **5/(2·Exp(1-0.2·t))** and **25·GDPpC·GompertzI(5-0.2·*t*)** are presented in Fig.2 and in Fig.3 as strings and parse trees, respectively [3].

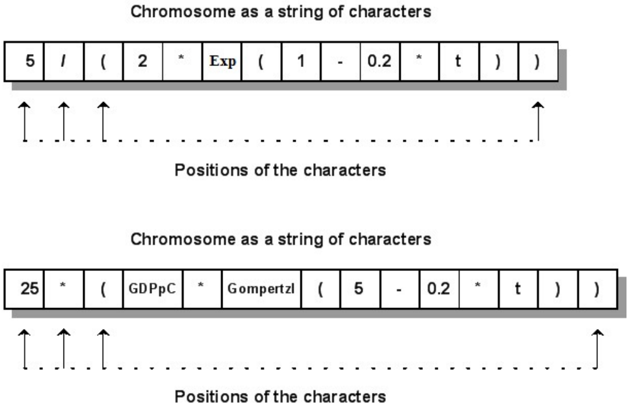
Chromosome as a string of characters
5 / (2 * Exp (1 - 0.2 * t))
Positions of the characters
Chromosome as a string of characters
25 * (GDPpC * GompertzI (5 - 0.2 * t))
Positions of the characters

Fig. 2. Representation of chromosomes in modified GP as strings

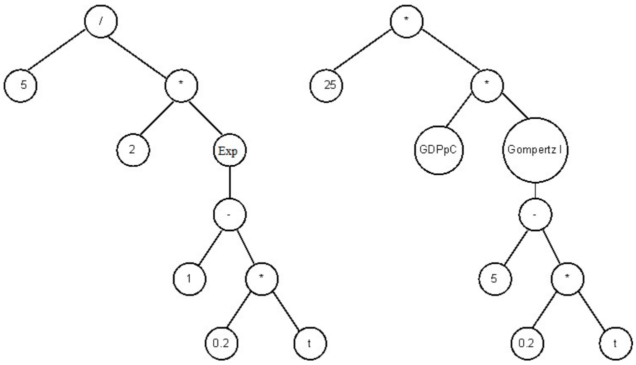

Fig. 3. Representation of chromosomes in modified GP as parse trees

D. Fitness Function and Evaluation

The selection of the best solution becomes according (6) for fitting and (7) for forecasting purposes. These are the fitness functions.

$$SSE = \sum_{t=1}^{T} [r(t) - y(t)]^2 \qquad (6)$$

$$wSSE = \sum_{t=1}^{T} w_t [r(t) - y(t)]^2 \qquad (7)$$

In (6), the sum of squared error (SSE) is over the time period ***t*=1, 2, 3…, T**. Also, ***r*(*t*)** are the dataset observed data for time ***t*** and finally, ***y*(*t*)** corresponds to the model's value [2]. In forecasting process, the fitness function is a combination of the weighted sum of squared error (wSSE) and Mean Absolute Percentage Error (MAPE), as in (7) and (8), respectively. In wSSE function, last data have greater weight factor than the initial data of the dataset.

The whole process is separated into training and forecasting processes. During training process, the wSSE function is used as fitness function, while the forecasting estimation is performed by MAPE. A sorted list with the evaluated solutions is created according the fitness function's value. Every one solution that is not satisfying the precision limit criterion is being removed. The remaining solutions are being sorted according to their fitness value and they are candidates for the next operation, the Crossover or Mutation [3].

E. Crossover and Mutation

In the crossover operation, two solutions (parents) are selected, according to the Tournament Selection process, from the sorted list. In the Tournament selection, a number of solutions from the list are randomly selected and then the best of those is chosen for the Crossover or Mutation [3].

In each parent characters string, one Crossover point is randomly chosen. The substring of each parent which begins at the crossover point is interchanged between two parents and the children are generated (see Fig.4).

In the mutation process, one string's point, which corresponds to a function, is randomly chosen. The process replaces the function, with a new random function [3] (see Fig.5).

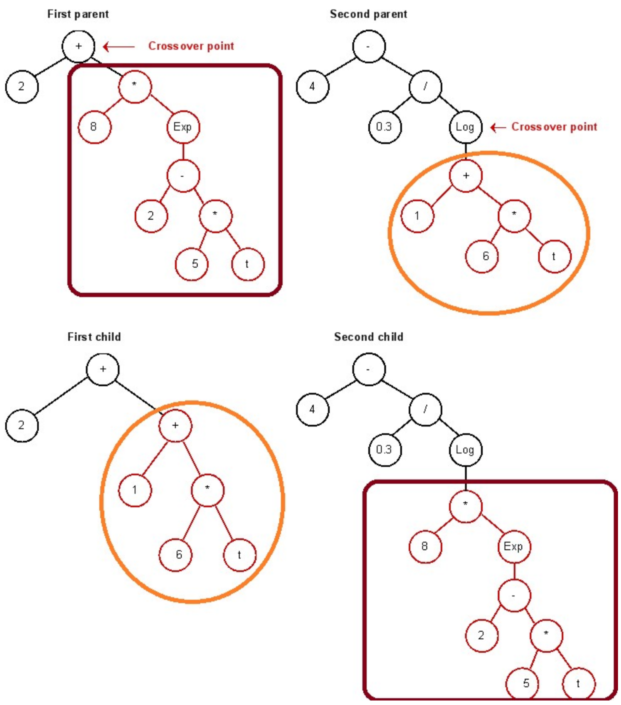

Fig. 4. Crossover Representation in modified GP as parse trees

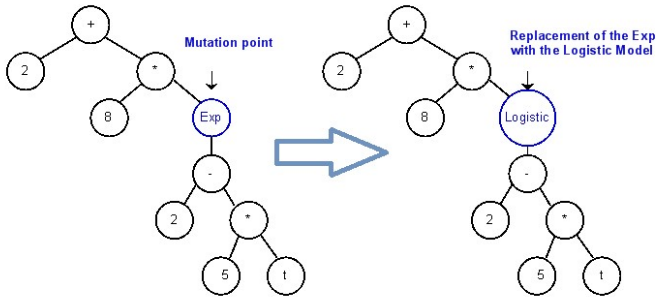

Fig. 5. Mutation Representation in modified GP as parse trees

F. Statistical Indices

Three statistical indices have been used in this study. The overall evaluation method uses Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE) and Mean Absolute Error (MAE).

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{r(t)-y(t)}{r(t)}\right| \tag{8}$$

$$MSE = \sum_{t=1}^{T}\frac{[r(t)-y(t)]^2}{T} \tag{9}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|r(t)-y(t)\right| \tag{10}$$

In (8), (9) and (10), ***y*(*t*)** corresponds to the model's value for time ***t=1, 2, 3…, T***. Also, ***r*(*t*)** are the data of the dataset.

G. Macroeconomic Indicators

In this study, macro-economic indices of GDPpC and CPI will be used, as in [3]. The GDPpC corresponds to the productivity of a country and not of personal income.

Gross Domestic Product (GDP) is a measure of the value of the services, goods produced in a country [10].

CPI corresponds to an average of basic consumer goods prices [10]. The CPI is normalized on the value of the year 2010.

H. Dataset

In this study, the proposed method has been implemented on OECD countries dataset. The dataset presents the overall OECD fixed broadband penetration. The data is derived from the OECD portal [10][11], which presents the total fixed broadband penetration per 100 inhabitants in OECD countries, until June 2015. Specifically, the dataset concerns the time period from fourth quarter (Q4) of the year 2003 until second quarter (Q2) of the 2015. The dataset is comprised by 24 data points.

IV. Results

In this study the SSE is the statistic indicator that has been used for the fitting process. In this section, the fitting performance of a generated GP model is presented.

A. *Fitting Performance of the GP method*

Table.I contains the program execution parameters of the GP method concerning the data sets in OECD countries.

TABLE I.

Fitting	**Program Parameters of modified GP**		
	Maximum Number of Generations	***Evaluation Function***	***Precision for candidates***
value	1000	SSE	0.005[a]

[a] Upper limit of the precision for solutions in Crossover and Mutation.

The fitting performance of the first modified GP model for the broadband penetration percentage in overall OECD countries, according to its fitness value (SSE), is presented in Fig.6.

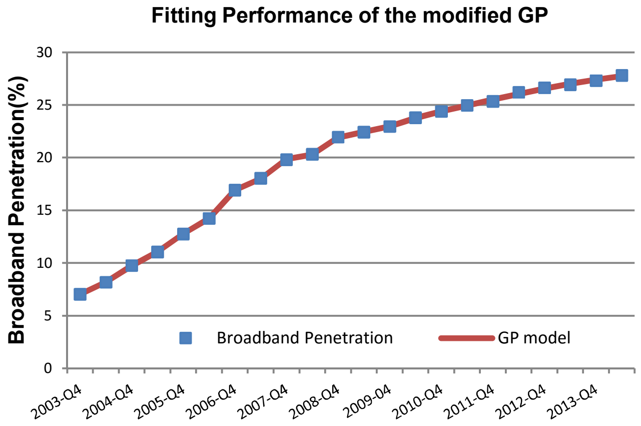

Fig. 6. The performance of curve fitting with GP method

The relative statistical indices SSE, MAPE, MSE, and MAE of the modified-hGP models are presented in Table.II

TABLE II.

Fitting	**Statistical Indices for modified GP model**			
	SSE	***MAPE***	***MSE***	***MAE***
GP model	0.0902217934	0,0021756192	0,004101	0,0473676

The fitting model follows:

119.55024823*Exp(-28.0502424865*Exp(-0.290518716343*t))-
(37.0440157582*Exp(-2.5079520116*Exp(-0.405611372254*t))-(-
50648.2421729-(-50594.7748172-6.54888900507*Exp(-15.8929379299*t))-
(-3318.48073403*Exp(-
0.30325385647*t))/(1+Exp(15.3513028353+0.410321219531*t))/(1+Exp(-
(3830.34908417/GDPpC)+1.088911209*LN(GDPpC/CPI)))))

As one can see, this method combines different variables like GDPpC or CPI with the independent variable of time. In Table.II, the modified-GP method achieves excellent statistical performance, as it shows a SSE value of 0.0902217934.

B. *Forecasting Performance of the GP method*

The program execution parameters of GP are presented in Table.III, for the forecasting process. The forecasting method for a future window of 3-semester ahead, uses 21 data points, as one can see in Fig.7.

TABLE III.

Forecasting	**Program Parameters of modified GP**		
	Maximum Number of Generations	***Evaluation Function***	***Precision for candidates***
Training	500	wSSE	0.5[b]
Forecast	1000	MAPE	0.9[b]

[b] Upper limit of the precision for solutions in Crossover and Mutation.

Also, the forecasting performance of the first modified GP model for the broadband penetration percentage in overall OECD countries is presented in Fig.7.

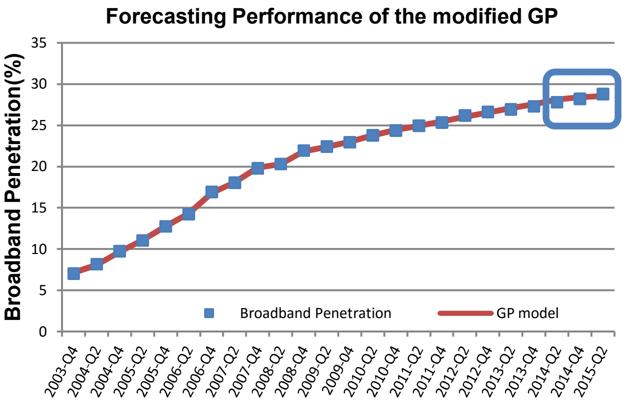

Fig. 7. The performance of forecasting with GP method

The relative statistical indices wSSE, MAPE, MSE, and MAE of the modified-hGP models are presented in Table.IV

TABLE IV.

Forecasting	**Statistical Indices for modified GP model**			
	wSSE	***MAPE***	***MSE***	***MAE***
Model GP	0,263935767	0,005326	0,0158441	0,0992763

The proposed modified-hGP model follows:

-4.76326403494*Exp(-157.366719463*Exp(-0.356124701112*t))-(-
35.0886035005*Exp(-2.90353082668*Exp(-0.150769230567*t))-(-
5116313.17117-(-5109202.53067-6.54888900507*Exp(-15.8929379299*t))-
(-21368481.293*(CPI/GDPpC)-621.316490609*LN(GDPpC/CPI))))

This model combines macro-economic variables with the variable of time. In Table.IV the modified-GP method achieves excellent statistical performance, as it shows a wSSE value of 0.263935767 and MAPE 0.0053264. So for the 2nd Quarter of 2015 the forecasting value is 28.791 and the estimated from OECD is 28.796.

C. *Forecasting Scenarios*

In this section, we investigate three scenarios derived from the forecasting model. The forecasting period is 3 years, from the 4th quarter of 2013 (the 21 data point of the dataset) until 4th quarter of 2016.

The first scenario presents 2.5% GDP and 1% CPI increase. In the specific scenario the forecasting model predicts a value of 29.78 per 100 inhabitants of fixed broadband penetration for the 4th quarter of 2016.

In second scenario, the growth of GDPpC and CPI is 4% and 2%, respectively. The forecasting model predicts a value of 30.66 for the 4th quarter of 2016.

Finally, in the third, worst scenario, the growth for GDPpC and CPI is 0%. The forecasting model predicts a value of 29.698 per 100 inhabitants of fixed broadband penetration for the 4th quarter of 2016.

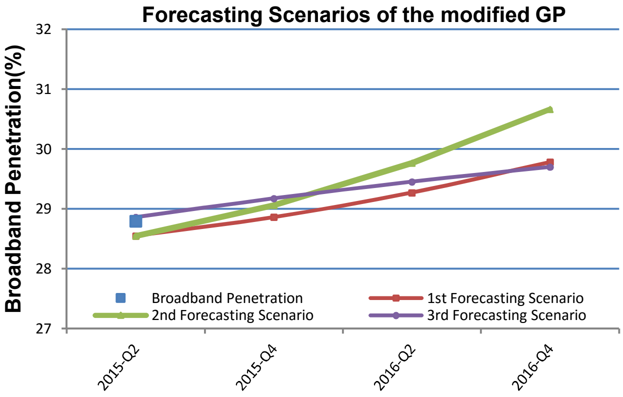

Fig. 8. Forecasting scenarios for the GP method

V. Conclusion

In this paper the modified hGP method [3] is implemented with an updated dataset of the fixed broadband penetration in OECD total. It achieves to produce forecasting models with one or more variables. The fitting and forecasting performance of the modified hGP are presented and the method presents satisfactory statistical indices. The combination of diffusion models and GP with this method indicates that could be a robust forecasting method, especially for no big datasets.

The method represents a forecasting framework that could produce time dependant models and/or causal models for short or long-term forecasting, with more than one variable.

References

[1] Salpasaranis Konstantinos and Stylianakis Vasilios, “A New Empirical Model for Short-Term Forecasting of the Broadband Penetration: A Short Research in Greece,” Modelling and Simulation in Engineering, vol. 2011, Article ID 798960, 10 pages, 2011. doi:10.1155/2011/798960

[2] Konstantinos Salpasaranis and Vasilios Stylianakis, “A Hybrid Genetic Programming Method in Optimization and Forecasting: A Case Study of the Broadband Penetration in OECD Countries,” Advances in Operations Research, vol. 2012, Article ID 904797, 32 pages, 2012. doi:10.1155/2012/904797

[3] Konstantinos Salpasaranis, Vasilios Stylianakis, and Stavros Kotsopoulos, “Combining Diffusion Models and Macroeconomic Indicators with a Modified Genetic Programming Method: Implementation in Forecasting the Number of Mobile Telecommunications Subscribers in OECD Countries,” Advances in Operations Research, vol. 2014, Article ID 568478, 20 pages, 2014. doi:10.1155/2014/568478

[4] OECD. (2015), *OECD Digital Economy Outlook 2015*, OECD Publishing,Paris. DOI: http://dx.doi.org/10.1787/9789264232440-en

[5] Bass F. M., “A new product growth for model consumer durables,” Management Science, vol. 15, no. 5, pp. 215–227, 1969

[6] Holland, J. H., “Adaptation in Natural and Artificial Systems”, University of Michigan Press, 1975

[7] Koza J.R., “Genetic Programming: On the programming of Computers by Means of Natural Selection”, The MIT Press (1992)

[8] Koza J.R., “Genetic programming for economic modeling”, Statistics and Computing 4(2), 187–197 (1994)

[9] Collective Intelligence of Genetic Programming for Macroeconomic Forecasting P. Jędrzejowicz et al. (Eds.):ICCCI 2011, Part II, LNCS 6923, pp. 445–454, Springer-Verlag Berlin Heidelberg, 2011

[10] OECD (2014),OECD Factbook 2014: Economic, Environmental and Social Statistics, OECD Publishing. http://dx.doi.org/10.1787/factbook-2014-en

[11] OECD, Broadband Portal, www.oecd.org/sti/broadband/oecdbroadbandportal.htm, February 2016