Testing the Dinosaur Hypothesis under Empirical Datasets

Michael Kampouridis[1], Shu-Heng Chen[2], and Edward Tsang[1]

[1] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, CO4 3SQ, UK
[2] AI-Econ Center, Department of Economics, National Cheng Chi University, Taipei, Taiwan 11623

Abstract. In this paper we present the Dinosaur Hypothesis, which states that the behaviour of a market never settles down and that the population of predictors continually co-evolves with this market. To the best of our knowledge, this observation has only been made and tested under artificial datasets, but not with real data. In this work, we attempt to formalize this hypothesis by presenting its main constituents. We also test it with empirical data, under 10 international datasets. Results show that for the majority of the datasets the Dinosaur Hypothesis is not supported.

Key words: Dinosaur Hypothesis, Genetic Programming

1 Introduction

The Dinosaur Hypothesis (DH) is inspired by an observation of Arthur [1]. In his work, Arthur and his group conducted the following experiment under the Santa Fe Institute Artificial Stock Market. They first allowed the market evolve for long enough. They then took the most successful agent with his winning predictor[3] out of this continuously evolving market, "froze" him for a while, and then returned the agent back to the market. They found that the early winner could not perform as well as he used to do in the past. His predictors were out of date, which had turned him into a *dinosaur*. This is quite an interesting observation, because it indicates that *any successful predictor or trading strategy can only live for a finite amount of time.*

In addition, Chen and Yeh [3] also tested the existence of this non-stationary market behaviour in their artificial stock market framework; their results verified Arthur's observation. Furthermore, they observed that a dinosaur's performance decreases monotonically.

Based on these observations, Chen [2] suggested a new hypothesis, called the Dinosaur Hypothesis. The DH states that the market behaviour never settles down and that the population of predictors continually co-evolves with this market.

[3] Predictor is the model that the agents use for forecasting purposes. In Arthur's work, predictor is a GP parse tree. In this work, predictors are Genetic Decision Trees (see Sect. 3 for more details). We also refer to them as trading strategies.

In this paper, we first formalize the DH by presenting its main constituents. In addition, motivated by the fact that both Arthur, Chen and Yeh made their observations under an artificial stock market framework, we want to examine whether the same observations hold in the 'real' world. We thus test the hypothesis with empirical data. We run tests for 10 international markets and hence provide a general examination of the plausibility of the DH. Our tests take place under an evolutionary environment, with the use of GP [7]. One goal of our empirical study is to use the DH as a benchmark and examine how well it describes the empirical results which we observe from the various markets.

The rest of this paper is organized as follows: Section 2 elaborates on the DH, and Section 3 briefly presents the GP algorithm that is going to be used for testing the DH. Section 4 then presents the experimental designs, Section 5 addresses the methodology employed to test the DH, and Section 6 presents and discusses the results of our experiments. Finally, Section 7 concludes this paper.

2 The Dinosaur Hypothesis

Based on Arthur's work, we can derive the following statements which form the basic constituents of the DH:

1. The market behaviour never settles down
2. The population of predictors continuously co-evolves with the market

These two statements indicate the non-stationary nature of financial markets and imply that strategies need to evolve and follow the changes in these markets, in order to survive. If they do not co-evolve with the market, their performance deteriorates and makes them ineffective.

However, as we said earlier, these observations were made in an artificial stock market framework. What we thus do in this paper is to test the above statements against our empirical data. We propose the following *Fitness Test*:

The average fitness of the population of predictors from future periods should

1. Not return to the range of fitness of the base period (P1)
2. Decrease continuously, as the testing period moves further away from the base period (P2)

As we can see, there is a population of predictors, which in our framework these are Genetic Decision Trees (GDTs); what we do in this work is to monitor the future performance of these GDTs in terms of their fitness, in accordance with Arthur's and Chen and Yeh's experiments. More details about the testing methodology can be found at Sect. 5.

Statement P1 is quite straightforward and is inspired by Arthur [1]. The term 'range of fitness' is also explained in Sect. 5. Statement P2 is inspired by the observation that Chen and Yeh made [3], regarding the monotonic decrease of a predictor's performance. However, in our framework we do not require the performance decrease to be monotonic. This is because when Chen and Yeh tested

for the Dinosaur Hypothesis (they did not explicitly use this term), they only tested it over a period-window of 20 days, which is relatively short, hence easy to achieve monotonic decreasing. Thus, requiring that a predictor's performance decreases monotonically in the long run would be very strict, and indeed hard to achieve. For that reason, statement P2 requires that the performance decrease is continuous, but not monotonic. It should also be mentioned that we are interested in qualitative results, meaning that we want to see how close the real market behaves in comparison with what is described by the DH.

Finally, in order to make the reading of this paper more comprehensive, we present two definitions, inspired by Arthur's work: *Dinosaur*, is a predictor who has performed well in some periods, but then ceased performing well in the periods that followed. This means that his predictor may or may not become effective again. If it does, then it is called a *returning dinosaur*.

3 GP Algorithm

Our simple GP is inspired by a financial forecasting tool, EDDIE [6], which learns and extracts knowledge from a set of data. This set of data is composed of the daily closing price of a stock, a number of attributes and signals. The attributes are indicators commonly used in technical analysis [5]: Moving Average (MA), Trader Break Out (TBR), Filter (FLR), Volatility (Vol), Momentum (Mom), and Momentum Moving Average (MomMA). Each indicator has two different periods, a short- and a long-term one, 12 and 50 days respectively.

The signals are calculated by looking ahead of the closing price for a time horizon of n days, trying to detect if there is an increase of the price by $r\%$. For this set of experiments, n was set to 1 and r to 0. In other words, the GP tries to forecast whether the daily closing price will increase in the following day.

Furthermore, Fig. 1 presents the Backus Naur Form (BNF) (grammar) of the GP. The root of the tree is an If-Then-Else statement. Then the first branch is a Boolean (testing whether a technical indicator is greater than/less than/equal to a value). The 'Then' and 'Else' branches can be a new GDT, or a decision, to buy or not-to-buy (denoted by 1 and 0). Thus, each individual in the population is a GDT and its recommendation is to buy (1) or not-to-buy (0). Each GDT's performance is evaluated by a fitness function presented below.

Depending on what the prediction of the GDT and the signal in the training data is, we can define the following 3 metrics:
Rate of Correctness

$$RC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Rate of Missing Chances

$$RMC = \frac{FN}{FN + TP} \tag{2}$$

Rate of Failure

$$RF = \frac{FP}{FP + TP} \tag{3}$$

<Tree> ::= If-then-else <Condition> <Tree> <Tree> | Decision
<Condition> ::= <Condition> “And” <Condition> |
<Condition> “Or” <Condition> |
”Not” <Condition> |
Variable <RelationOperation> Threshold
<Variable> ::= MA_12 | MA_50 | TBR_12 | TBR_50 | FLR_12 |
FLR_50 | Vol_12 | Vol_50 | Mom_12 | Mom_50 |
MomMA_12 | MomMA_50
<RelationOperation> ::= “>” | “<” | “=”
Decision is an integer, Positive or Negative implemented
Threshold is a real number

Fig. 1. The Backus Naur Form that the simple GP uses to construct trees

We use these metrics to define the following fitness function:

$$ff = w_1 * RC - w_2 * RMC - w_3 * RF \tag{4}$$

where w_1, w_2 and w_3 are the weights for RC, RMC and RF respectively. The weights are given in order to reflect the preferences of investors. For our experiments, we chose to include GDTs that mainly focus on correctness and reduced failure. Thus these weights have been set to 0.6, 0.1 and 0.3 respectively, and are given in this way in order to reflect the importance of each performance measure for our predictions.

4 Experimental Designs

Experiments were conducted for a period of 17 years (1991-2007) and the data was taken from the daily closing prices of 10 international market indices: CAC 40 (France), DJIA (USA), FTSE 100 (UK), HSI (Hong Kong), NASDAQ (USA), NIKEI 225 (Japan), NYSE (USA), S&P 500 (USA), STI (Singapore) and TAIEX (Taiwan). For each of these markets, we run each experiment for 10 times.

Each year was split into 2 halves (January-June, July-December), so in total, out of the 17 years, we have 34 periods[4]. The GP system was hence executed 34 times. Table 1 presents the GP parameters for our experiments. The behavior of each GDT can be represented by its series of market timing decisions over the entire trading horizon. Thus, the behaviour of each rule is a binary vector of 1s and 0s (buy and not-to-buy). The length or the dimensionality of these vectors is then determined by the length of the trading horizon, which in this study is 6 months, i.e., 125 days long; hence, the market timing vector has 125 dimensions.

Here we should emphasize that the GP was only used for creating and evolving the trading strategies. No validation or testing took place, as it happens in

[4] At this point the length of the period was chosen arbitrarily to 6 months. We leave it to a future research to examine if and how this time horizon can affect our results.

Table 1. GP Parameters. The GP parameters for our experiments are the ones used by Koza [7]. Only the tournament size has been changed (lowered), and the reason for that was because we have observed premature convergence under a larger tournament size. Other than that, the results seem to be insensitive to these parameters.

GP Parameters	
Max Initial Depth	6
Max Depth	17
Generations	50
Population size	500
Tournament size	2
Reproduction probability	0.1
Crossover probability	0.9
Mutation probability	0.01

the traditional GP approach. The reason for this is because we were not using the GP for forecasting purposes; instead, we were interested in using the GP as a *rule inference engine* which would evolve profitable trading strategies for a certain period of time. The GP was thus used for each of the 34 periods to create and evolve trading strategies. After the evolution of the strategies under a specific period, these strategies are not tested against another set. This approach is consistent with the Lo's Adaptive Market Hypothesis [8], as it states that the heuristics of an old environment are not necessarily suited to the new ones. Our no-testing approach is also consistent with the well-tested *overreaction hypothesis* [4], which essentially states that top-ranked portfolios are outperformed by bottom-ranked portfolios during the next period. Thus, after evolving a number of generations (50 in this paper), what stands (survives) at the end (the last generation) is, presumably, a population of financial agents whose market-timing strategies are financially rather successful. This population should, therefore, interest us in spirit of Arthur's adaptive market process; therefore, we use them to test how those competitive strategies perform in the future periods.

5 Testing Methodology

This section presents the testing methodology. But before we do this, let us first present some frequently used terms:

- *Base period*, is the period during which GP was used to create and evolve GDTs that are going to be used for testing the DH
- *Future period(s)*, is a period(s) which follow the base period (in chronological order)

We are interested in observing how the average fitness of the population of GDTs changes throughout time. As we have already seen, we used a simple GP system to generate and evolve trading strategies for each one of the 34 periods.

After this step, we apply this evolved population of GDTs to the future periods' data. In order to better explain this, let us use an example. Let us suppose that the period we trained the GDTs (base period) was the first semester of 1991 (1991a); we can then calculate the average fitness of the population of these trees for this period. From this point on, we will be calling this 'average fitness of the population of GDTs' as *population fitness*. We thus have an indication of how well the population performs during the base period. Then, we apply all evolved GDTs to the data of future periods: second semester of 1991 (1991b), first semester of 1992 (1992a),..., second semester of 2007 (2007b) and calculate the population fitness for each one of these periods. As a result, we can observe how this fitness changes over the future periods.

The same procedure is followed for all periods until 2007a, so that all of them act as a base period. This means that when 1991b is the base period, the GDTs that were created and evolved during 1991b will be applied to all future periods. After 1991b, 1992a takes over as the base period and the same procedure happens again. We do this until 2007a. We obviously cannot do this for 2007b, since there are no data available after this year. The reader should also bear in mind that we only apply the evolved GDTs to future periods; for instance, when the base period is 2000a, we do not apply the GDTs backwards in time, only forwards. We are not interested in looking what happens in the past; we are only interested in observing how the fitness of the GDTs is affected in the future.

Given a base period, the population fitness of all periods is normalized by dividing those population fitnesses by the population fitness in the base period. Hence, each base period has its normalized population fitness equal to 1 and a returning dinosaur is a population of strategies from future periods that has its normalized population fitness 'close to 1'. At this point, we need to define the term 'close to 1'. Strictly speaking, this means that this population's normalized fitness is greater or equal to 1. However, in our opinion, other future periods which do not necessarily satisfy this condition could be considered as returning dinosaurs, too. Let us consider the case of a future period with normalized population fitness very 'close to 1', e.g. 0.99. When this happens, it indicates that there exist those similar market conditions in this future period, as in the base period, so that the dinosaurs can again have high performance. Although this performance may not be exactly equal to 1, we believe that the fact that the normalized population fitness of these strategies (GDTs) is this 'close' to 1, indicates that these GDTs have become successful again, and should thus be considered as returning dinosaurs.

However, defining a specific range of fitnesses for 'close' would be arbitrary; after all, closeness is only a matter of degree. We therefore present in the next section the statistics of fitness observed for each stock market. Besides, as we said in Sect. 2, we are interested in qualitative results; we want to see how close the 10 empirical markets behave in comparison by what is described by the DH.

If DH holds, we should observe two things: firstly, the normalized population fitness of the future periods has decreased and does not return to the range of fitness of the base period (P1), and secondly, this decrease is continuous (P2).

6 Results

6.1 Statement P1

According to P1, the future periods' population fitness will not return to the range of fitness of the base period. As we saw earlier, we test this statement for one period at a time. The subject period forms our base period.

In order to examine how often dinosaurs return, we iterate through each base period and calculate the maximum fitness among its future periods. Let us give an example. If 1991a is the base period, then there is a series of 33 population fitness values for its future periods. We obtain the maximum value among these 33 values, in order to check how close to 1 this future period is. This process is then repeated for 1991b and its 32 future periods, 1992a, and so on, until base period 2007a. We thus end up with a 1×33 vector, which shows the potential returning dinosaur per base period. The graph of this vector is presented in Fig. 2. Each line represents the results on a different dataset and they have been divided in four separate subfigures: CAC40-DJIA-FTSE100 (top-left), HSI-NASDAQ-NIKEI (top-right), NYSE-S&P500 (bottom-left), and STI-TAIEX (bottom-right).

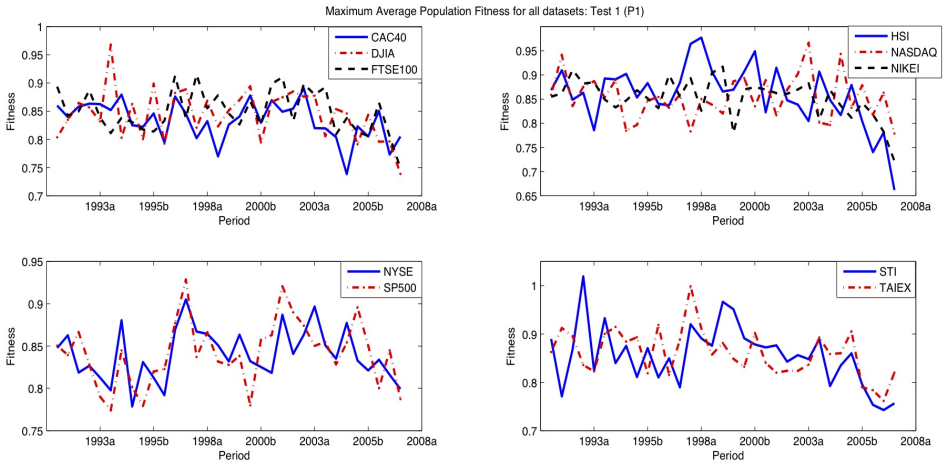

Fig. 2. Fitness Test, P1: The maximum normalized population fitness among all future periods for each base period. Each line represents a single dataset. Results have been divided in 4 subfigures.

What we can see from this figure is that only STI has a base period (1992b) with a maximum normalized population fitness exceeding 1. This indicates returning dinosaurs and goes against statement P1. We cannot observe any more periods that reach the threshold of maximum population fitness greater or equal to 1. Nonetheless, all of our datasets seem to have quite high population fitness

values, which many times exceed 0.9 or even 0.95 (e.g. DJIA-1993b, HSI-1998b, NASDAQ-2003a, TAIEX-1997b). Therefore, although we cannot strictly talk about a returning dinosaur, we should also not neglect the fact that this is an indication that the market environment actually can create conditions that are very similar to the past and as a result, successful strategies from the past do not necessarily have a finite lifetime (as the DH implies), but can again be successful in the future. Thus, our results do not support statement P1.

6.2 Statement P2

To show a continuous decrease in the population fitness, we calculate the sum of the fitness values of all those future periods that are 1 period away from the base period, then the sum of those future periods that are 2 periods away, and so on, up to a period difference between future and base period of 33. In order to do this, we first need to create a table of distances, like the one in Table 2(a). Each row of this table presents the distance of the future periods from their base period. For instance, if 91a is the base (first row), then future period 91b has distance equal to 1, future period 92a has distance equal to 2, and so on. Table 2(b) shows the series of population fitness values for the future periods of each base period. For example, when the base period is 91a (first row), the normalized population fitness starts from 1 in 91a, then drops to 0.66 (91b), then goes to 0.72 (92a), and so on, until it reaches fitness equal to 0.74 in future period 07b. Let us now denote the sum of fitnesses we mentioned at the beginning of this section by $\sum_{|i-j|=m} Fit(i,j)$, where i,j are the base and future period respectively, $|i-j|$ is their absolute distance, as presented in Table 2(a), and m is the distance from the base period and takes values from 1 to 33. We divide this sum by the number of occurrences where $|i-j|=m$. This process hence returns the average of the normalized population fitness, and allows us to observe how it changes, as the distance m from the base period increases. We call this metric D_m and it is presented in (5).

$$D_m = \frac{\sum\limits_{|i-j|=m} Fit(i,j)}{\{\#(i,j), |i-j|=m\}} \tag{5}$$

Let us give an example: if we want to calculate D_{32}, we need to sum up the population fitnesses that have distance $m=32$. This happens with $Fit(91a, 07a)$ (fitness of GDTs from base period 91a, when applied to future period 07a) and $Fit(91b, 07b)$ (fitness of GDTs from base period 91b, when applied to future period 07b). Therefore D_{32} would be equal to the sum of these two fitness rates divided by 2, as there are only 2 periods that can have $m=32$.[5] By calculating

[5] The distance $m=32$ can also be found in 07a91a and 07b91b. However, we do not take them into account because, as we said earlier in Sect. 5, we are not interested in applying the evolved GDTs of a base period (here 07a and 07b) backwards in time (91a and 91b, respectively).

Table 2. (a) Distance of future periods from their base period, over the 17 years 1991-2007. The further away we move from a period, a single unit of distance is added. (b) Series of future population fitnesses per base period. Each base period's series is presented as a horizontal line of this table. Fitness values have been normalized, so that the average fitness in the base period is always equal to 1.

(a)

		j					
		91a	91b	92a	92b	...	07b
	91a	0	1	2	3	...	33
	91b	1	0	1	2	...	32
i	92a	2	1	0	1	...	31
	...	...	...	...	...	...	...
	07b	33	32	31	30	...	0

(b)

		j					
		91a	91b	92a	92b	...	07b
	91a	1	0.66	0.72	0.78	...	0.74
	91b		1	0.76	0.72	...	0.70
i	92a			1	0.74	...	0.77
	...	...	...	...	...	...	...
	07b					...	1

D_m for all m values, we can have a clear idea of how the average of the population fitness changes when we move from periods that are close to the base period (low m), to periods that are further away (high m), and thus observe whether there is a continuous decrease. Figure 3 presents the results for all datasets. Each line represents again a single dataset, similar to that in Fig. 2.

What we observe from this figure is that there are upwards and downwards movements of the D_m metric. This is consistent for all datasets. We do not observe a continuous, or any kind of decrease in general, in the metric. This therefore does not validate the P2 statement.

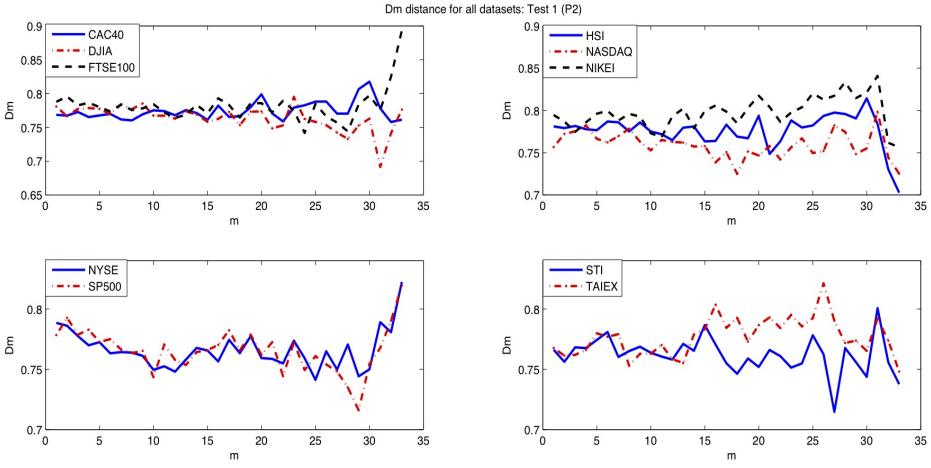

Fig. 3. Fitness Test, P2: D_m values for all m from 1 to 33. Each line represents a single dataset. Results have been divided in 4 subfigures.

7 Conclusion

This paper presented and formalized the Dinosaur Hypothesis. The DH says that the behaviour of a market never settles down and that the strategies in this market continuously co-evolve with it. This was an observation first made by Arthur [1] and later by Chen and Yeh [3]. However, these two works made these observations under an artificial stock market. In this paper, we were interesting in examining whether these observations could also hold in the real world and thus tested the hypothesis with empirical data. For our experiments, we used a *fitness test*, where we created and evolved trading strategies with a GP system. Results showed that 1 of the 10 datasets tested demonstrated the existence of returning dinosaurs; having a returning dinosaur is of course contradicting with statement P1. However, it would not be accurate to say that the remaining 9 datasets fully support P1. This is because all of population strategies have had future periods' average fitness values that are close to the fitness of the base period; in fact, there were many occasions were this fitness was even more than 90% closer to the population fitness of the base period. Therefore, although there is no normalized population fitness among these 9 datasets that reaches 1, we can argue that trading strategies from the past can still be applied to the market and perform satisfactory, even if many years have passed. Markets can thus have a number of 'typical states', where past rules may become useful again. *Returning dinosaurs hence exist.* Finally, regarding statement P2: we did not observe any continuous decrease in the average population fitness of any of the 10 datasets tested, and we can thus argue that P2 is not supported by the empirical data in this work. Overall, we can conclude that the empirical evidence that can support the Dinosaur Hypothesis is quite weak.

References

1. Arthur, B.: On learning and adaptation in the economy (1992), working paper 92-07-038, Santa Fe Institute
2. Chen, S.H.: Financial Applications: Stock Markets, pp. 481–498. Wiley Encyclopedia of Computer Science and Engineering, John Wiley & Sons, Inc (2008)
3. Chen, S.H., Yeh, C.H.: Evolving traders and the business school with genetic programming: A new architecture of the agent based artificial stock market. Journal Of Economic Dynamics & Control 25, 363–393 (2001)
4. De Bondt, W., Thaler, R.: Does the stock market overreact? Journal of Finance 40, 793–805 (1985)
5. Edwards, R., Magee, J.: Technical analysis of stock trends. New York Institute of Finance (1992)
6. Kampouridis, M., Tsang, E.: EDDIE for investment opportunities forecasting: Extending the search space of the GP. In: Proceedings of the IEEE Conference on Evolutionary Computation. Barcelona, Spain (2010), Accepted for Publication.
7. Koza, J.: Genetic Programming: on the programming of computers by means of natural selection. Cambridge, MA: MIT Press (1992)
8. Lo, A.: The adaptive market hypothesis: market efficiency from an evolutionary perspective. Journal of Portfolio Management 30, 15–29 (2004)