Hybridisation of ARIMA and Non-linear Machine Learning Models for Time-series Forecasting

Piotr Arendarski

University of Warsaw

December 5, 2012

Table of contents

1. Introduction
2. Research Objectives
3. Data
4. Methodology
5. Results
6. Conclusions

Linear vs. Non-Linaer Models

- Conventional statistical methods, the autoregressive integrated moving average (ARIMA) is extensively utilized in constructing a forecasting model
- ARIMA cannot be utilized to produce an accurate model for forecasting nonlinear time series
- Machine Learning algorithms have been successfully utilized to develop a nonlinear model for forecasting time series
- Determining whether a linear or nonlinear model should be fitted to a real-world data set is difficult

Linear Approach - ARIMA Model

- Box and Jenkins (1976) developed the autoregressive moving average to predict time series
- The ARIMA model is used for prediction non-stationary time series when linearity between variables is supposed

Linear Approach - ARIMA Model

- Box and Jenkins (1976) developed the autoregressive moving average to predict time series
- The ARIMA model is used for prediction non-stationary time series when linearity between variables is supposed
- However, in many practical situations supposing **linearity is not valid**

Linear Approach - ARIMA Model

- Box and Jenkins (1976) developed the autoregressive moving average to predict time series
- The ARIMA model is used for prediction non-stationary time series when linearity between variables is supposed
- However, in many practical situations supposing **linearity is not valid**
- For this reason, ARIMA models do not produce effective results when used for **explaining and capturing nonlinear relations** of many real world problems, what results in increased forecast error

Linear Approach - ARIMA Model

- Box and Jenkins (1976) developed the autoregressive moving average to predict time series
- The ARIMA model is used for prediction non-stationary time series when linearity between variables is supposed
- However, in many practical situations supposing **linearity is not valid**
- For this reason, ARIMA models do not produce effective results when used for **explaining and capturing nonlinear relations** of many real world problems, what results in increased forecast error

Linear Approach - ARIMA Model

- Box and Jenkins (1976) developed the autoregressive moving average to predict time series
- The ARIMA model is used for prediction non-stationary time series when linearity between variables is supposed
- However, in many practical situations supposing **linearity is not valid**
- For this reason, ARIMA models do not produce effective results when used for **explaining and capturing nonlinear relations** of many real world problems, what results in increased forecast error

Non-linear Approach - Machine Learning Methods

To deal with this problem, various non-linear approaches have been suggested in the literature.

Kernel-based machine learning

Support vector machines

Evolutionary algorithm

Genetic Programming

Others machine learning methods

Artificial neural network (ANN) and others algorithms

Support vector machines (SVM) vs. other methods

- ANN is known to overfit data unless cross-validation is applied while SVM **does not overfit data** and curse of dimensionality is avoided
- Unlike neural network methods, the SVM approach does not attempt to control model complexity by **keeping the number of features small**
- SVM training **always finds a global minimum**

Forecasted subject:	**Author:**	**Year:**	**SVM outperformance:**
Financial TS	Tay	2001	Back-propagation NN
Stock price index	Kim	2003	BPNN
Futures contracts	Cao	2003	BPNN
Nikkei255 directions	Huang	2005	BPNN
SP500	Lahmiri	2011	Probabilistic NN

Genetic Programming vs. other methods

1. SVM and GP are powerful methods
2. The empirical comparison is hardly found in the literature

Forecasted subject:	**Author:**	**Year:**	**GP outperformance:**
Stock index	Saski	1999	ANN
Insurance industry	Sanz	2003	SVM
Egyptian stock market	El-Telbany	2004	ANN
Software reliability	Zhang	2006	ANN
GDP China, US, Japan	Li	2007	ARIMA
Homes prices	Kaboudan	2007	ANN
Wave forecasting	Gaur	2008	ANN
Stock market	Rajabioun	2008	ANN
Industrial production	Klucik	2009	ARIMA
Transport energy demand	Forouzanfara	2012	ANN

Hybrid Models vs. single models

Proposed Hybrid:	**Author:**	**Year:**	**Hybrid outperformance:**
SARIMA - BPNN	Tseng	2000	SARIMA, BPNN
ARIMA - ANN	Zhang	2003	ARIMA, ANN
ARIMA - SVM	Pai	2005	ARIMA, SVM
SARIMA - SVM	Chen	2005	SVM, SARIMA
ARIMA - ANN	Diaz	2008	ANN, ARIMA
ARIMA - ANN	Robles	2008	ARIMA, ANN
ARIMA - ANN	Valenzuela	2008	ARIMA, ANN
ARIMA - ERNN	Aladag	2009	FFNN, ARIMA
ARIMA - ANN	Flores	2009	ARIMA, ANN
ARIMA - ANN	Faruk	2009	ARIMA, ANN
ARIMA - ANN	Khashei	2010	ARIMA, ANN

Hybrid Models vs. single models cont.

1. Lee (2010) used non-polynomial activation function for Genetic Programming

Proposed Hybrid:	**Author:**	**Year:**	**Hybrid outperformance:**
ARIMA - ANN	Areekul	2010	ARIMA, ANN
ARIMA - GP	**Lee**	**2010**	**ARIMA, ANN, GP**
ARIMA - RBFN	Shafie-khah	2011	ARIMA, ARIMA-Walewet
ARIMA - RBFN	Khashei	2011	ARIMA, ANN
SVR - ARIMA	Chen	2011	ARIMA, BPNN
ARIMA - BPNN.GA	Wang	2012	ARIMA, BPNN
ARFIMA - FNN	Aladag	2012	ARFIMA, FNN
ARIMA - SVM	Nie	2012	ARIMA and SVMs

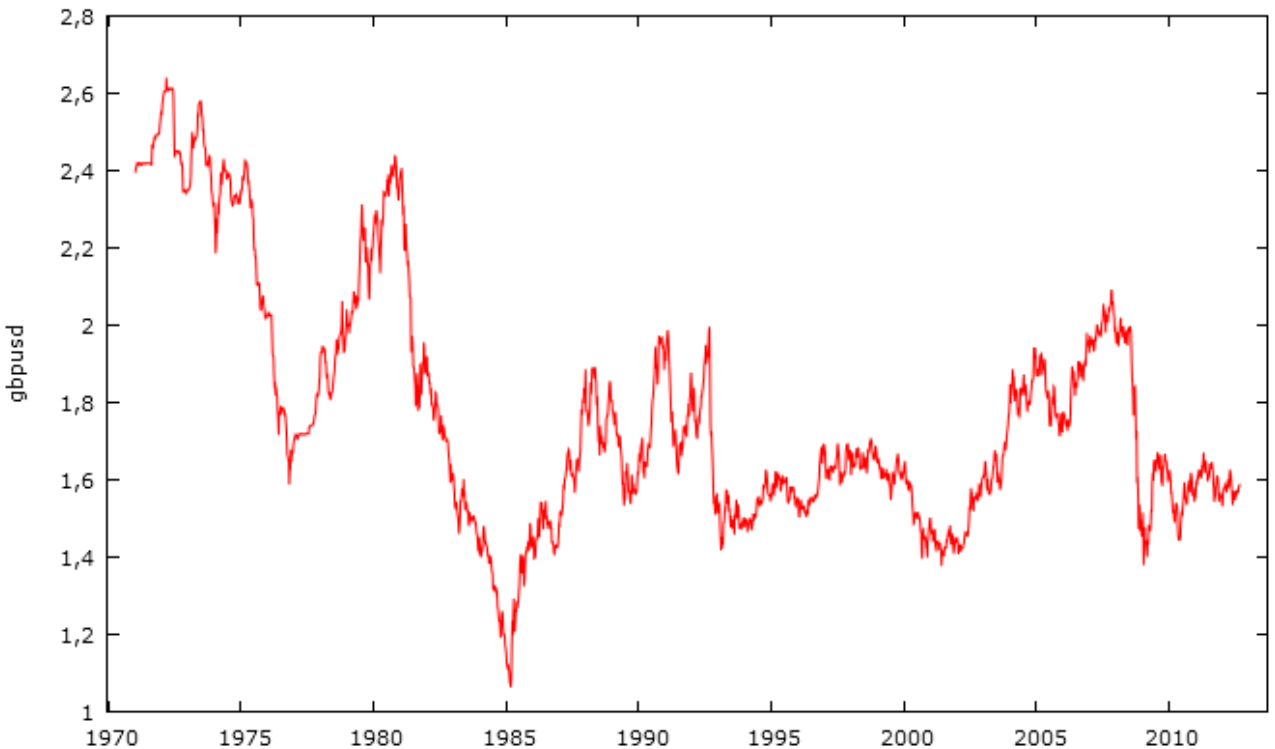
Research Objectives

- The proposed model
  1. Hybrid ARIMA - Genetic Programming, ARIMA - SVM
- The benchmark models
  1. Hybrid ARIMA
- Contribution
  1. Genetic Programming for Polynomial Regression - hybridisation with ARIMA
  2. Different time series modeling, universality of the proposed hybrid forecasting model

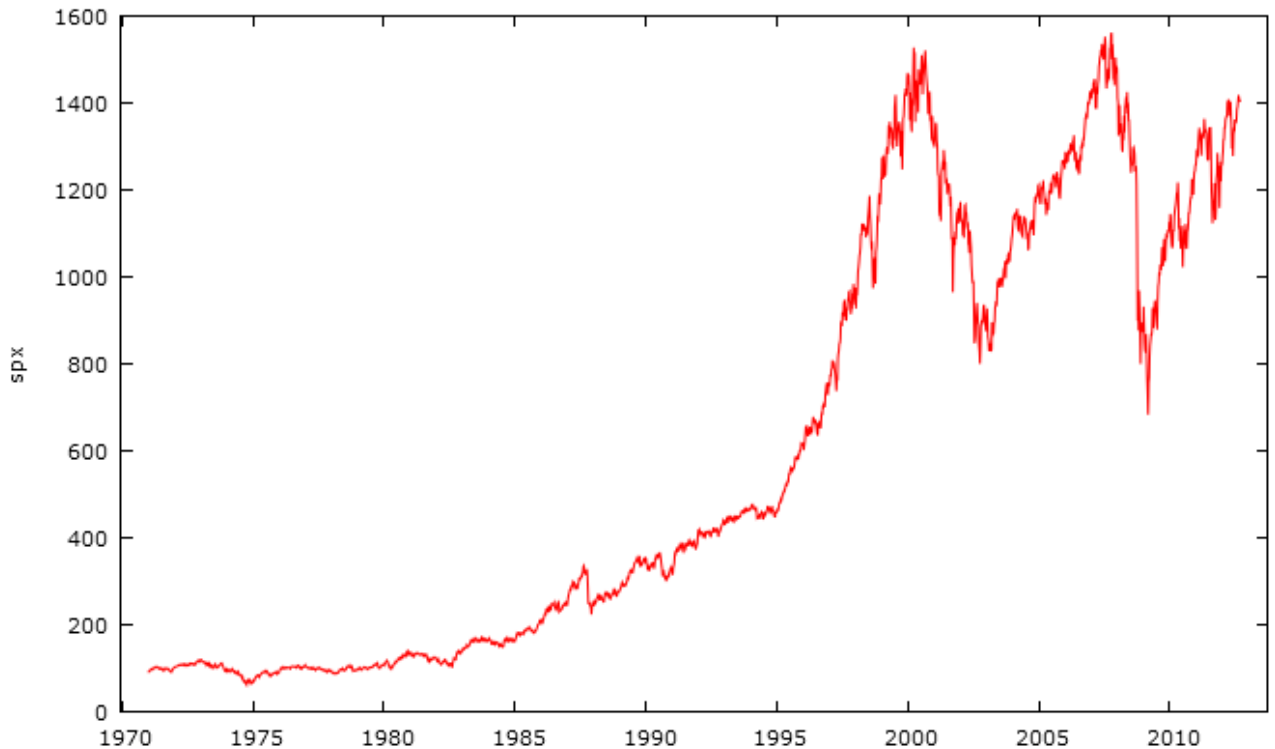
Data

- Financial time series
  - British Pound / US Dolar (GU)
  - S&P 500 index futures
  - Wheat futures
- Time frame
  - Weekly data
  - 01.01.1970 - 30.09.2012 giving a total 2176 weekly observations
    - In-sample data (1976 observations)
    - Out-of-sample data (200 observations)

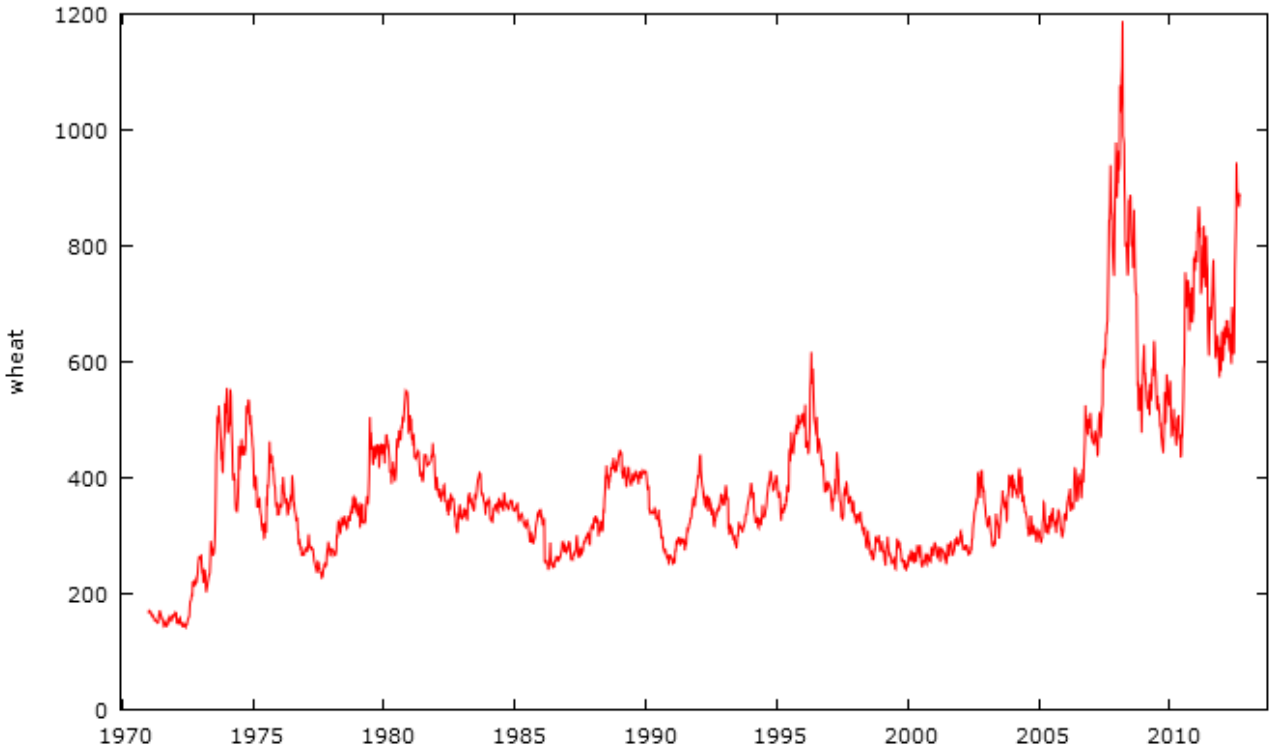
GBPUSD data


gbpusd
2,8
2,6
2,4
2,2
2
1,8
1,6
1,4
1,2
1
1970
1975
1980
1985
1990
1995
2000
2005
2010

S&P 500 data


spx
1600
1400
1200
1000
800
600
400
200
0
1970
1975
1980
1985
1990
1995
2000
2005
2010

Wheat data


wheat
1200
1000
800
600
400
200
0
1970
1975
1980
1985
1990
1995
2000
2005
2010

A hybrid model

A hybrid model: $y_t = L_t + N_t$ (1)
where:

- y_t represents the original positive time series at time t;
- L_t represents the linear component
- N_t is the nonlinear component of the model

The residuals can be obtained using the ARIMA model:
$r_t = y_t + \hat{L}_t$ (2)
where

- r_t is estimated using such nonlinear methods as GP or SVM

A hybrid model cont.

- $\hat{L_t}$ is the forecasted value of L_t and is estimated using the ARIMA model
- The residual can be rewritten as follows:
  $r_t = f(r_{t-1}, r_{t-2}, ..., r_{t-n}) + \varepsilon_t$ (3)
  where
  - $f(r_{t-1}, r_{t-2}, ..., r_{t-n})$ represents the nonlinear function that is constructed using GP or SVM
  - ε_t is the random error term.

The hybrid model for forecasting time series is: $\hat{y}_t = \hat{L_t} + \hat{N}_t$ (4)

A hybrid model step-by-step

1. Step1. The ARIMA model is used to model the linear component of time series. That is, L_t is obtained by using the ARIMA model.
2. Step2. From Step 1, the residuals from the ARIMA model are obtained. The residuals are modeled by the GP and SVM models in equation(3). That is, N_t is the forecast value of equation (3) by using GP and SVM
3. Step3. Using equation(4), the forecasts of the hybrid model are obtained by adding the forecasted values of linear (ARIMA) and nonlinear (GP and SVM components).

ARIMA introduction

- The basic idea of this forecasting approach is to predict future values of time series y_t using:
  - past values of the time series, y_{t-k}
  - the past residuals from the model, ϵ_{t-k}
- ARIMA models form an important part of the Box-Jenkins approach to time-series modelling.
- When one of the three terms is zero, it's usual to drop "AR", "I" or "MA". For example, ARIMA(0,1,0) is I(1), and ARIMA(0,0,1) is MA(1).

ARIMA setup

- Model identification
  - The appropriate ARIMA(p, d, q) model is obtained by applying the Akaike Information Criterion (AIC) rule
- Parameter estimation
- Modeling diagnosis
  - Tests for white noise residuals indicate whether the residual series contains additional information that might be utilized by a more complex model

Representations of ARIMA models

GBPUSD

ARIMA(1,1,3)

S&P 500

ARIMA(3,1,2)

Wheat

ARIMA(1,1,2)

SVM introduction

- Basic idea of Support Vector Machines
  - Optimal hyperplane for linearly separable patterns
  - Extend to patterns that are not linearly separable by transformations of original data to map into new space Kernel function

SVM setup

- Sigma estimation
  - Automatic sigma estimation (sigest)
- Default parameters
  - Kernel - Radial Basis kernel "Gaussian"
  - Cost of constraints violation - 0.1
  - Epsilon in the insensitive-loss function - 0.1

Genetic programming intro

Genetic Programming (GP) tackles learning problems by means of searching a **computer program space** for the most probable program given a functionality specification. The search is performed using an Evolutionary Algorithm.

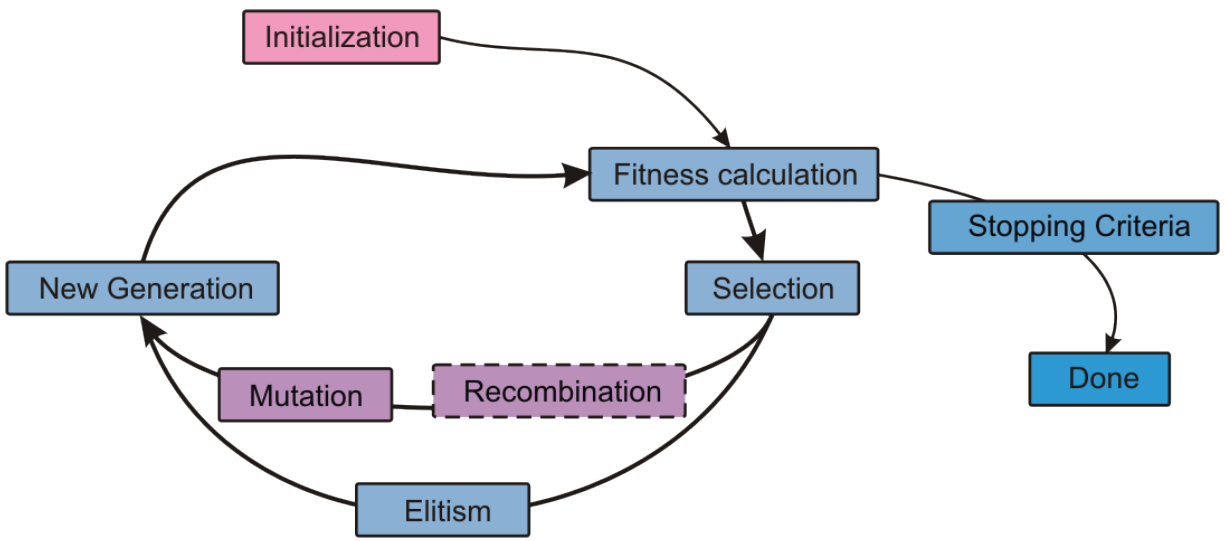


Figure : The Evolutionary Algorithm cycle

Genetic Programming of polynomial models

- The polynomial regression problem can be formulated as follows: given training examples $\{(x_1, y_i), \ldots, (x_N, y_N)\}$ of explanatory variables, that is vectors $x_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \in R^d$, and corresponding response values $y_i \in R$
- Genetic Programming of polynomial models is a powerful paradigm for learning well performing non-linear regression models
  - References: Nikolaev,N., and Iba,H. (2002). Genetic Programming of Polynomial Models for Financial Forecasting.

Genetic programming setup

As for the GP model, the input, output variables are:

- $(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-5}, y_{t-6}, y_{t-7}, y_{t-8})$ and y_t, respectively.

To reduce the forecast error of nonlinear component of ARIMA - GP model, the **fitness function** of GP is expressed as:

$$\sum_{i=1}^{N} \frac{(\hat{r}_t - r_t)^2}{N}$$

where r_t represents the actual residual value, and the $\hat{r}_t$ represents the forecasted value of r_t.

Representation ARIMA-GP Model

Hybrid ARIMA-GP model generated for S&P 500 time series:

$$\hat{y}_t = 0.1387 \cdot y_{t-1} - 0.6742 \cdot y_{t-2} - 0.1295 \cdot y_{t-3} - 0.2285 \cdot \epsilon_{t-1} + 0.7626 \cdot \epsilon_{t-2} + \sin(0.103602645916204 + r_{t-8}) \cdot (0.177070398354227 \cdot r_{t-4} \cdot r_{t-7})$$

Mean Square Error - GBPUSD data

Out-of-sample results

	ARIMA	**ARIMA-SVM**	**ARIMA-GP**
1 month	0.00283	0.00272	0.00219
2 months	0.00260	0.00324	0.00255
6 months	0.00165	0.00179	0.00149
12 months	0.00120	0.00134	0.00090
24 months	0.00078	0.00085	0.00060
All sample	**0.00055**	**0.00061**	**0.00048**

Mean Square Error - S&P 500 data

Out-of-sample results

	ARIMA	**ARIMA-SVM**	**ARIMA-GP**
1 month	3286.8	3866.4	3732.3
2 months	2708.1	2279.9	2280.2
6 months	1765.9	1679.6	1549.3
12 months	706.7	1469.2	1298.1
24 months	1024.0	1151.9	1021.3
All sample	**963.8**	**1110.9**	**943.2**

Mean Square Error - Wheat data

Out-of-sample results

	ARIMA	**ARIMA-SVM**	**ARIMA-GP**
1 month	2478.3	2486.9	2321.2
2 months	2828.4	1431.5	1561.2
6 months	2770.6	976.1	966.1
12 months	2400.5	833.7	811.9
24 months	1118.2	910.1	745.2
All sample	**1198.5**	**1187.2**	**1121.7**

Average weekly returns - GBPUSD data

Out-of-sample results

	ARIMA	**ARIMA-SVM**	**ARIMA-GP**
1 month	0.008	0.016	0.025
2 months	-0.004	0.012	0.023
6 months	0.004	0.006	0.019
12 months	0.010	0.001	0.003
24 months	-0.007	0.000	0.002
All sample	**0.000**	**0.000**	**0.001**

Average weekly returns - S&P data

Out-of-sample results

	ARIMA	**ARIMA-SVM**	**ARIMA-GP**
1 month	0.059	0.000	0.039
2 months	0.044	0.012	0.042
6 months	0.031	0.008	0.020
12 months	0.020	0.005	0.021
24 months	0.007	0.004	0.019
All sample	**0.001**	**0.002**	**0.010**

Average weekly returns - Wheat data

Out-of-sample results

	ARIMA	**ARIMA-SVM**	**ARIMA-GP**
1 month	0.089	0.063	0.071
2 months	0.057	0.038	0.069
6 months	0.056	0.011	0.060
12 months	0.042	0.009	0.059
24 months	0.019	-0.002	0.010
All sample	**-0.001**	**-0.001**	**0.004**

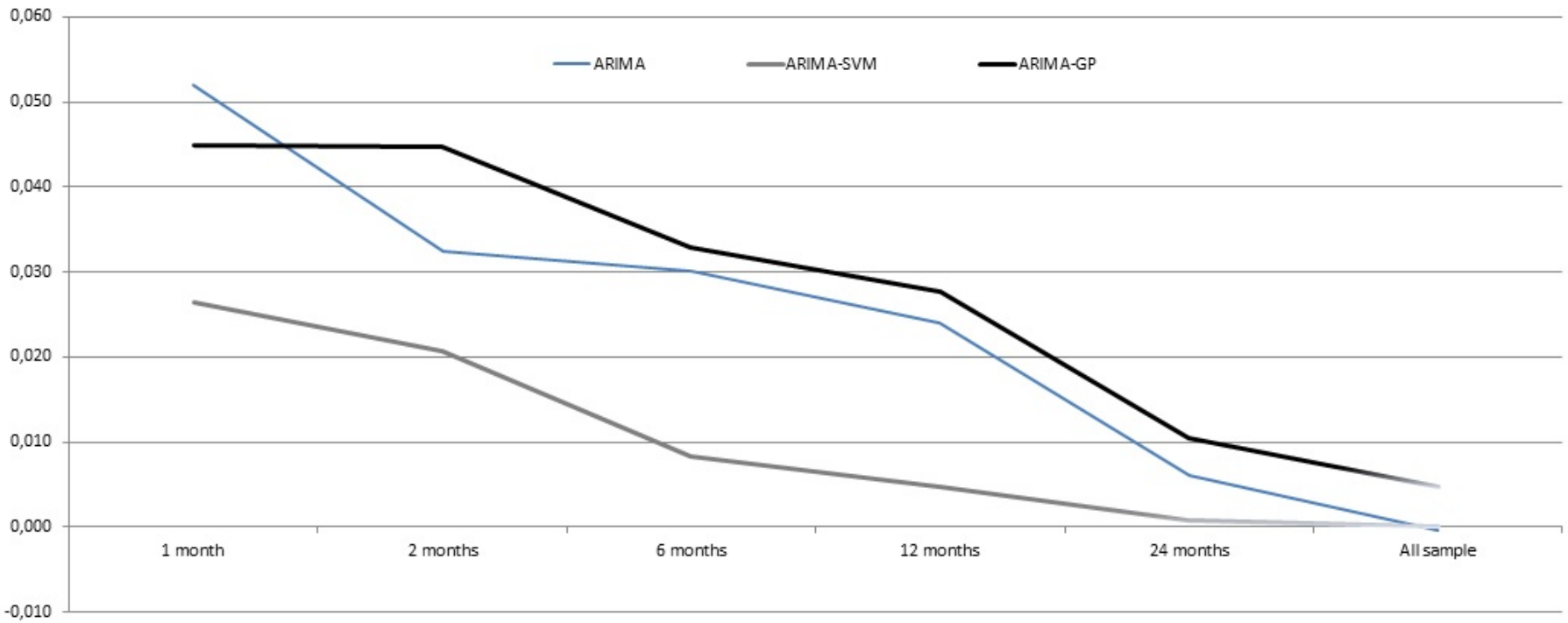
Average weekly returns - All data

Out-of-sample results

	ARIMA	**ARIMA-SVM**	**ARIMA-GP**
1 month	0.052	0.026	0.045
2 months	0.033	0.021	0.045
6 months	0.030	0.008	0.033
12 months	0.024	0.005	0.028
24 months	0.006	0.001	0.011
All sample	0.000	0.000	0.005

Average weekly return - All data

Average weekly return - methods comparison



Prediction Accuracy

- Hybrid structure : ARIMA - GP outperforms ARIMA and ARIMA - SVM across all the time series in terms of MSE

Trading Performance

- Hybrid structure : ARIMA - GP outperforms ARIMA and ARIMA - SVM across all the time series in terms of trading performance.
- In general, weekly average returns decrease as time pass
- There is a need to re-estimated the models