NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme 'HSE University and University of London Double Degree Programme in Data Science and Business Analytics'

UDC _____

Research Project Report

on the topic _____

(interim, the first stage)

Fulfilled by the Student:
group #БПАД___ _____ _____
Signature Surname, First name, Patronymic, if any

_____
Date

Checked by the Project Supervisor:

_____
Surname, First name, Patronymic (if any), Academic title (if any)

_____
Job

_____
Place of Work (Company or HSE Department)

Date_____2021 _____ _____
Grade according to 10-point scale Signature

Moscow 2021

Contents

1. Introduction

1.1 Abstract: relevance and statement of purpose

Nowadays it is nearly impossible to operate in a dynamic, data-related environment, such as stock exchange or any other constantly changing financial platform, without tracing the trends of the stocks and other valuable assets, which are being described with time series[1](1), which, along with increasing amount of data, become arduous to analyze and forecast.

The aim of this research is to try to extract trends of financial time series and estimate their persistence, by using different mathematical and statistical concepts and models, and verify their effectiveness by implementing the data-analyzing algorithms, using such programming languages as Python or any applicable C-syntax language. The obtained derivations are to be further leveraged, in order to build a precise prognostic data-driven model for an analysis of some specific time series.

1.2 Project objectives

The work on the research includes three general stages:

1st. Study and examine the method of trend derivation, which we have been provided with.

– Understand and explicitly describe the models, theory and algorithms underlying the given method of analysis of time series.

– Describe the key differences between the given model and some simple enough, widely known model for time series analysis, for example, AR-model[2] [ARM]. Explain in what sense, theoretically, the given method may turn out to be more efficient, precise or accurate, than the simpler one (time-efficiency, preciseness-efficiency, etc.).

2nd. Apply the trend derivation method to some known dataset.

– Implement the algorithm, which applies the given method to a time series. Use some programming language, for example Python or some C-syntax language. Implementation of AR-model may also be done, in case specific libraries for its appliance do not yet exist, or are hard to operate with.

– Take a dataset of a time series of some known public stock or any other financial asset, describe why exactly this dataset was picked, and apply these two models (the given one and AR-one) to the very same data array. The dataset should be drawn from

(1) Please, look at the “List of definitions” section. Further on “blue” references are readdressing the reader to the “List of definitions” section of the report.

some past period, so that one would be able to compare the subsequent behavior of the time series with the behavior of forecasted models.

– Using the obtained algorithms, extract the trends of the time series, and then try to build precise prognoses of the time series' behavior.

– Compare the empirical results of forecast, provided by the given model and the AR-model. Compare both models to the original time series, part of which was used as the basis for the modeling algorithm.

– Verify, whether the theoretical conclusion on the effectiveness of these models matches the empirical results.

3rd. Based on the obtained model, draw the conclusion on the presence of persistence of any significant trends in a time series and verify the credibility of the model's prognosis.

1.3 Project members and their responsibilities

As far as I know, there are two people in our research group, except for me. They are *Maria A. Timonina,* and *Vadim V. Dudkin*, who are both learning on the second course of the bachelor's program "Applied Mathematics and Informatics" of the Faculty of Computer Science in HSE. All of us have same responsibilities that is we have to follow the very stages, stated above; however, each of us must apply a unique method to analyze the data. Maria has to apply the continuous entropy method for time series analysis, extracting and tracing the trends' persistence, whilst Vlad is responsible for analyzing the time series, applying fractal Brownian motion method to forecast the data behavior, while, in addition applying Hurst's exponent to the data in order to make the prognostic model for a time series more precise.

1.4 Responsibilities of mine

My responsibilities are to follow the steps, described above (at "1.2 Project objectives" section), applying the SSA[3] [Singular Spectrum Analysis] method to analyze the data and create a forecasting model, which will further be used in order to extract the trends of some specific time series(2). I have to implement the algorithm performing the SS analysis, which is to be compared with a simple AR-model for predicting the time series' trending. The algorithm, as well as all additional components, models, graph drawers and table parsers will be designed, using Python as main

(2) I have not yet decided which exact data set should I pick, as far as I want to take the one, which would be the most interesting to deal with, and which will actually have some strong trends, being explicit enough to be described.

programming language [Reasons, underlying such choice will be described further in "3. Review and comparative analysis of the sources" section]. C++ may be used as well, in case, some libraries in python will not be able to satisfy the needs of mine, during the implementation of the prognostic model.

Implementation of ARM will be omitted (only the program, using it as a method will be present in my work), as far as there are several different libraries in python, which may be used for that very purpose. So far I have only read several books and articles, regarding the SSA method application for analyzing and extracting the trends of a time series, the next step for me would be to implement the algorithm for SSA and AR-based models.

The main aims of mine are to reach the final prognosis that will be drawn by the model, which I have implemented; to verify the credibility of this prognosis, comparing it with the initial time series; and to conclude on the persistence of trends in the picked financial time series, comprising of the results, obtained by the SSA model.

2. Methods, algorithms, and models for research project implementation

The main method, which is going to be used for the trend analysis and further forecast of a time series, is the Singular spectrum analysis (or SSA) method, which is based on two main source components: the method of principal components or Principle component analysis (PCA), and the singular value decomposition method (SVD). The main purpose of SSA is to decompose the series into the sum of the interpreted components such as trend, periodic components, and white noise[4]. Parametric shape of these components is not strictly required to be known.

Nina Golyandina has precisely described the SS (singular spectrum) method for data analysis in her study guide for Saint-Petersburg State University(3), whilst the basics of SVD and PCA methods (as components of the SS analysis) have been described[4] by Rasmus Elsborg Madsen, et al. Further information on implementation of SVD in an abstract C syntax language can be found(5) in the third edition of "Numerical recipes" by William H. Press, et al.

Right now, I am only going to briefly describe the main steps of the SS method for the time series analysis, since the whole precise model description, including the theory and linear algebra concepts, is going to be present in the thesis itself, as far as it will take a lot of time to explain it there.

Considering the real-valued time series $\mathbb{X} = (x_1, \ldots, x_N)$ of length N, such that $L \in (1, N)$, to be an integer called the "window length" and K to be such integer, that $K = N - L + 1$.

1st step: for the time series $\mathbb{X}$, the $L \times K$ trajectory Hankel matrix[5] $(x_{ij})_{i,j=1}^{L,K} = = \underline{X}$ is built.

2nd step: perform the Singular Value Decomposition of the obtained matrix $\mathbb{X}^H$, so that, taking square roots of the eigenvalues[6] of $S = \underline{X}\underline{X}^T$, i.e. $\sqrt{\lambda_i} : [i \in (1, L)]$ we provide the singular spectrum of $\underline{X}$, and then the specific vectors, designed by the roots of eigenvalues are taken as vectors of principle components.

3rd step: rewrite SVD expansion of $\underline{X}$ as such a decomposition: $\underline{X} = \underline{X}_{I,1} + \ldots + \underline{X}_{I,m}$, where $I_1, \ldots I_m$ are disjoint subsets of eigen-indicies.

4th step: Each matrix $\underline{X}_{I,j}$ of the grouped decomposition is hankelized and then the obtained

(3) Golyandina, Nina. *"Metod «Gusenitsa»-SSA: analiz vremennykh ryadov*". SPb, SPSU, 2004: 5-14.
(4) Madsen, Rasmus Elsborg, Lars Kai Hansen, and Ole Winther. "Singular value decomposition and principal component analysis." *Neural Networks 1*, 2004: 1-4.
(5) Press, William H., et al. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007: 65-75.

Hankel matrices are transformed into new series, each of length N using the one-to-one correspondence between Hankel matrices and time series. The reconstructed series (number of which now is exactly m) now are representing the initial time series, such that the initial time series is the sequential sum of m reconstructed subseries.

This decomposition is the main result of the SSA algorithm, as every i^{th} subseries may now be classified as either a component of the trend in a time series, some periodic component of a time series, or a constant of the white noise.

3. Review and comparative analysis of the sources

The sources, which are to be leveraged for the research purposes of mine, can be divided into three main groups:

3.1 Technical and programming sources

As far as there are many programming languages at the moment, which may all be very useful for various purposes, the decision was made to employ one of the most easy-to-use and intuitive in syntax programming language, which is at the same time a universal tool for data analysis and specific algorithms implementation, namely Python 3.7 'PyCharm Community Edition' v.2019.2.3 is used as an IDE for Python, as far as the researcher is already experienced in using it, and since it allows one to install and use various libraries for project's development directly inside an IDE.

Nowadays there exists a wide range of specific technical and programming sources, available for usage in the field of time series and data analysis. In the scope of this very research, it would be needed to operate with huge amounts of data (usually being present in .csv format files), while at the same time being able to change it or to leverage it in some particular manner, that is for example to create addendums of white noise (ε_t) to the file with raw financial time series data. It was decided to use 'Pandas' [v1.2.1] library as main one to operate with raw data, since it is easier to operate with huge .csv files using pandas, rather than for example MS Excel, as far as it is hard for it handle the work with huge flows of data (megabyte-scaled).

Also during the process of SSA method implementation such useful Python libraries for easing the process of data analysis are going to be used: 'NumPy' [v21.0.1], a fundamental package for scientific computing; 'SciPy' [v1.5.4], basically used to perform high-order mathematical operations, and implement harder statistical and mathematical models; 'statsmodels' [v0.12.2], a useful library, which involves algorithmical concepts of various statistical models, and containing model classes of several widely used methods for time series analysis as well [`statsmodels.tsa`]; 'scikit-learn' [v0.24.1], being used as extensive library for dealing with huge databases, as well as applying different statistical and mathematical operations and models to the data; 'matplotlib' [v3.3.4], a very convenient library for visualization of data and different statistical series.

3.2 Information sources

The information (theoretical) sources, which are going to be used in the research process, include the descriptions of SSA method, its theoretical concepts, underlying the method's credibility, and mathematical procedures and algorithms, which should inevitably be performed during the

implementation of SSA.

The articles and books, which are used as theoretical basis for this research are:

♦ Madsen, Rasmus Elsborg, Lars Kai Hansen, and Ole Winther. "Singular value decomposition and principal component analysis." *Neural Networks 1*, 2004.

– Brief introduction to SVD and PCA, explanation of linear algebra principles, singular value decomposition is nesting on.

♦ Bishop, Christopher M. *Neural networks for pattern recognition*. Oxford university press, 1995. (Chapter 8)

– About dimensionality reduction and extension for raw data preprocessing, input normalization, explanation of multi-step-ahead prediction approach. Explanation of approaches to reduce the error accumulation for multi-step-ahead predictor systems (specific data subset selections).

♦ Press, William H., et al. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007. (Chapter 2, subsection 6)

– Explanation of how to implement an algorithm, performing an SVD of a matric, using abstract C syntax language.

♦ Jolliffe, Ian T. "Graphical representation of data using principal components." *Principal component analysis* (2002): 10-59, 299-308.

– Explanation of the PC analysis, description of principles on which it lies, suggestions on how to use it for statistical analysis, and how is it present in use for SSA method.

♦ Golyandina, Nina. *"Metod «Gusenitsa»-SSA: analiz vremennykh ryadov*". SPb, SPSU, 2004.

– Basic explanation of steps to be performed for explicit application of an SSA algorithm. Description of how dimension extension should be performed to the data array, extracted from a time series, and how singular value decomposition and PCA should be applied to transformed matric-like data in order to analyze its trends and their persistence.

3.3 Data sources

Since it has not been decided yet, which exact time series to pick in order to use its data to make a precise analysis and prognosis of that time series, using the SSA method, it is reasonable so

far to provide several financial platforms, such that it is possible to get some datasets of different financial time series.

The sources for the raw historical data were chosen as most representative and informative sources, these are:

- NASDAQ (the second greatest US stock exchange)
- NYSE (New York stock exchange)
- QUANDL (huge financial and economic datasets storing service and provider)

4. List of definitions

[1] *Time series* – a sequence of some specific numerical data points listed in successive time order.

[2] *AR-model* – a time series analytical model, in which it is assumed that the values of the time series at a given moment are linearly dependent on the previous values of the very same series.

[3] *Singular Spectrum Analysis* – a method of time series analysis based on the transformation of a one-dimensional time series dataset into a multidimensional series applying the singular value decomposition method to the matrix representation of the series, with the subsequent application of the principal component analysis to the obtained multivariate time series.

[4] *The white noise* – a random signal, whose samples are a sequence of unrelated, random variables with no mean and limited variance.

[5] *Hankel matrix* – a square matrix in which all diagonals, which are perpendicular to the main one, consist of equal elements.

[6] *Eigenvalue* – such a constant (λ), that for some linear operator φ, defined on some vector space $\mathbb{V}$ over the field $\mathbb{F}$, $\varphi(v) = \lambda v$, where $v \in \mathbb{V}$ is an eigenvector of φ (s.t. $v \neq \underline{0}$ [zero matrix of size = $\dim(\mathbb{V})$]).

5. List of sources and bibliography

1. Bibliographical references [MLA]:

[1] Madsen, Rasmus Elsborg, Lars Kai Hansen, and Ole Winther. "Singular value decomposition and principal component analysis." *Neural Networks 1*, 2004.

[2] Bishop, Christopher M. *Neural networks for pattern recognition*. Oxford university press, 1995: 295-332.

[3] Press, William H., et al. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007: 65-75.

[4] Jolliffe, Ian T. "Graphical representation of data using principal components." *Principal component analysis* (2002): 10-59, 299-308.

[5] Golyandina, Nina. *"Metod «Gusenitsa»-SSA: analiz vremennykh ryadov*". SPb, SPSU, 2004: 5-14, 49-54.

2. Other references:

[6] NASDAQ, National Association of Securities Dealers Automated Quotation, www.nasdaq.com/index. Accessed 8 February 2021.

[7] NYSE. The New York Stock Exchange, www.nyse.com/market-data/historical#. Accessed 8 February 2021.

[8] QUANDL, The premier source for financial, economic, and alternative datasets, www.quandl.com/. Accessed 8 February 2021.

[9] Maton, Nathan. “Time Series Analysis Tutorial Using Financial Data”. Towards data science. towardsdatascience.com/time-series-analysis-tutorial-using-financial-data-4d1b846489f9/. Accessed 8 February 2021.