Parallelization of the VAR Algorithm Family to Increase the Efficiency of Forecasting Market Indicators During Social Disaster

Artem Khovrat [1], Volodymyr Kobziev [1], Alexei Nazarov [1] and Sergiy Yakovlev [2]

[1] *Kharkiv National University of Radio Electronics, 14, Nauky, Ave., Kharkiv, 61166, Ukraine*
[2] *Lodz University of Technology, 90-924 Lodz, Poland*

Abstract

Social unrest caused by the crisis makes it difficult to forecast the market performance for the company, and, accordingly, the decision-making process. Today's situation creates the basis for food, migration, political, and energy crises. The complexity of the mathematical representation of the situation, and the generally chaotic nature of changes necessitates the use of a more comprehensive approach to forecasting market indicators. The dimensionality reduction and parallelism principles are traditional approaches to reduce the impact of data volumes on the execution time of an algorithm. For the current study, it was decided to choose a MapReduce technology to implement a parallelization approach. Based on the results of this study, the main approaches to forecasting economic indicators were defined within the current framework. The chronological non-determinism of the target data was declared, and their co-integration was considered accordingly. The obtained data made it possible to define a family of VAR models as the basis for forecasting these indicators. A set of factors as exogenous variables and related parameters were defined. The accuracy of forecasting the specified algorithm family was checked using the e-commerce market data during recent social disasters. The obtained indicator of parallelization efficiency of these algorithms using the MapReduce technology allows us to state the expediency of a similar parallelism form for vector autoregressive forecasting models in practical application.

Keywords [1]

E-commerce, forecasting, parallelization, social disaster, VAR

1. Introduction

In 2020, a pandemic was declared as a reaction to the spread of an acute respiratory disease caused by the SARS-Cov-2 coronavirus [1]. According to the latest data, the total mortality has decreased by 90% and it's already possible to state a general decline in morbidity [2], however, the pandemic managed to cause the world economy losses, which, according to the ILO, amount to more than 3.5 trillion US dollars [3].

On February 24, 2022, the situation worsened as a result of the full-scale invasion of the Russian Federation on the territory of Ukraine. According to various analytical agencies, the war provoked various types of crisis phenomena at the global level: energy, food, migration, etc. [4, 5]. It, in turn, caused social unrest in different parts of the world [6, 7]. This state led to the market situation transformation and market relations in general. As an example, it can be noted that in 2020, the demand for online entertainment increased several times [8]. In order to generalize the definition of the specified problems, we will use the concept of social vulnerability – it's the inability of people, organizations, or society to resist the influence of a stressful situation [9]. A phenomenon that causes social vulnerability will be called a social disaster. The duration of this is determined separately for

Information Technology and Implementation (IT&I-2022), November 30 - December 02, 2022, Kyiv, Ukraine
EMAIL: artem.khovrat@gmail.com (A. 1); volodymyr.kobziev@nure.ua (A. 2); oleksii.nazarov1@nure.ua (A. 3); svsyak7@gmail.com (A. 4)
ORCID: 0000-0002-1753-8929 (A. 1); 0000-0002-8303-1595 (A. 2); 0000-0001-8682-5000 (A. 3); 0000-0003-1707-843X (A. 4)

each case and without significant reference to the event that caused it. For example, the tragedy of September 11, according to some scientific sources, still has a social impact on the USA [10].

During a social disaster, forecasting market activity becomes a complex and atypical task, which without the use of modern information technologies and/or appropriate education is an almost insoluble task. This causes the prevalence of risk forecasting related to consumer behaviour and means of automating these processes. One such tool that is currently gaining popularity is forecasting indicators using autoregression algorithms [11]. They guarantee a relatively accurate forecasting result, but with large amounts of data available, their performance is quite limited. To solve this problem, we can use the principles of parallelism, but their classical forms do not guarantee a significant gain in speed [12]. MapReduce technology is used as a basic alternative to these principles. As part of the current paper, a review of models for forecasting economic indicators based on autoregression, the possibility of modifying these algorithms, and the means of their parallelization will be carried out. The purpose is to determine the effectiveness of using MapReduce technology to parallelize autoregressive models for e-commerce market indicators during social disasters. To achieve this purpose, the following tasks can be distinguished:

- analyze the subject domain;
- carry out a mathematical representation overview of the selected algorithm family;
- develop a modification of the models to take into account the impact of a social disaster on the industry, the economy and the audience;
- build an experimental environment that would allow checking the feasibility of applying the MapReduce technology to modified algorithms;
- conduct an experiment and analyze the results.

2. Domain analysis

The concept of social vulnerability is most often considered in terms of risk analysis, which has three key stages: definition, assessment, and management. In the 20th century, to optimize this process, decision support systems (DSS) were created [13]. Nowadays, these systems are actively implemented in various business environments, where they are divided according to [14] by the following criteria:

- method of support: knowledge-based, document-based, data-based, communication-based, and model-based;
- interaction with the user: cooperative, active, passive.

The current paper will consider passive data-oriented systems because it was decided to concentrate on autoregressive models. An example of such systems can be [15]: Hyperproof; Soterion; Whistic. It is notable that some systems, in addition to the autoregressive approach, implement the Bayesian approach and/or artificial intelligence methods [16, 17]. However, in the framework of the current work, these approaches are not considered, because, because of the available studies, their working time is much higher than autoregression models [18, 19]. During social disasters, which can include both military actions and natural catastrophes, the time of decision-making is the critical indicator.

Not all autoregressive models will be considered either. This is explained by the fact that, in most cases, market indicators, such as the level of demand, prices, and investments, are time series and should be considered complex. The reason for this can be considered the possibility of classical paradoxical situations (Giffen's paradox, Veblen's paradox, etc.) and impossibility of their reflection by a single indicator. Therefore, the current study will consider only vector autoregressive models (abbreviated VAR), in particular:

- vector autoregression of the moving average;
- vector autoregression of the distributed lag;
- vector seasonal autoregression.

Such a choice can be explained by the general volatility of economic indicators and the aggravation of this process during social disasters. In addition, it was decided to slightly modify the input data to take into account qualitative indicators that would describe the state of the disaster.

3. Mathematical representation

Carry out a step-by-step review of the selected models for research, their modifications, and the technology for their parallelization.

3.1. Models overview

Before proceeding to the modification of the basic VAR family algorithms, we will consider their mathematical represantation:

$$\Phi_0 y_t = \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + \Theta_0 u_t + \Theta_1 u_{t-1} + \cdots + \Theta_q u_{t-q}, \quad (1)$$

where y_t.– K-dimensional time series; Φ_i, Θ_j – matrices $K \times K$, $i = \overline{1,p}$, $j = \overline{1,q}$; u_t – K-dimensional vector of white noise with zero mean and nondegenerate covariance matrix $\Sigma = \mathbb{E}(u_t, u_t')$. Φ_0 and Θ_0 – nondegenerate matrices.

It follows from the given formula (1) that the classic models predicts only static variables. To exogenous indicators consider, it was decided to use a modification of error correction (EC). Such an adjustment is necessary if several endogenous variables have a common stochastic trend, which is typical for indicators of the market activity.

The general formula for the modified EC-VAR family of algorithms will have the following form:

$$\Phi_0 \Delta y_t = \Pi y_{t-1} + \Psi_1 \Delta y_{t-1} + \cdots + \Psi_{p-1} y_{t-p+1} + \Theta_0 u_t + \Theta_1 u_{t-1} + \cdots + \Theta_q u_{t-q}, \quad (2)$$

where

$$\Pi = -(\Phi_0 - \Phi_1 - \cdots - \Phi_p), \Psi_i = -(\Phi_{i+1} + \cdots + \Phi_p), i = \overline{1, p-1}.$$

In the selected case, an important place is occupied by exogenous variables, in fact, they will serve as the basis for taking into account the social disasters in the quantitatively presented.

3.2. Models modification

In general, it's accepted that a social disaster affects all market subjects, from individuals to the state. The business management theory indicates that the nature of the impact on individuals is reflected in the behavior of the target audience, changes in the company's work and market regulation e (which is a representation of the state) are combined into the microeconomic profile of the market situation. Given the fact that for each indicator it is necessary to consider the disaster itself, we will begin the modification by analyzing the possibilities of converting its description into a quantitative form. For qualitative indicators to reflect an objective view of the disaster, it was decided to conduct an expert assessment among 5 risk managers of different companies in Kharkiv and Novomoskovsk, 5 sociologists from Kharkiv and Dnipro. The following four indicators were mentioned the largest number of times:

- a general textual description of a social disaster;
- the news is related to a catalyzing phenomenon;
- assessment of the company's employees regarding readiness for unforeseen circumstances;
- the duration of the disaster (from the moment the catalyst begins to act).

The last indicator has a quantitative representation, and the subjective assessment of employees can easily be converted into a quantitative form if a survey is using a scale from 0 to 100, where 0 is "the company is not prepared for unforeseen circumstances". and 100 – "the company is fully prepared for unforeseen circumstances." After the survey, the result will be normalized to 1. For conversion into quantitative form, we will use the principles of content analysis. In general, the algorithm should be as follows:

1. remove non-letter expressions from the text, in particular numbers and punctuation marks;
2. divide the cleaned text into sentences and words;
3. apply the stemming operation – shortening the word to its base;
4. to prevent stemming errors, we carry out lemmatization – bringing the word form to the lemma;
5. unify the created dictionary;

6. remove words with minimal linguistic load, for example, conjunctions;
7. find the TF-IDF frequency response [20];
8. find the polarity indicator of each word using the available word corpora;
9. multiply TF-IDF to polarity to find the frequency-polarity indicator (hereinafter referred to as the FPI);
10. find the amount of the received FPI for each received text (or its part);
11. normalize the value of the sum in the range from 0 to 1, where 0 is the absence of a disaster, and 1 is a significant negative impact.

We can formally describe the social disaster as follows:

$$SSD = \frac{SDO \times SDS \times t}{R}, \quad (3)$$

where *SDS* – content analysis result for disaster description provided by the company; *SDO* – content analysis result for disaster description provided by news; *t* – disaster duration; *R* – readiness for unforeseen circumstances.

Now let's move on to the main indicators of the profile of the market situation and the target audience. For the target audience, by analogy with the social disaster indicator, it was decided to conduct an expert assessment, during which it was found that the following factors are the most important for this indicator:

- general text description of the target audience;
- audience size;
- volume of income divided by total income.

The last two indicators are quantitative, so it is enough to normalize them from 0 to 1. To convert the text description, we will use the principle of cluster analysis. The first 5 steps of the algorithm are similar to those noted in the description of the disaster. The final steps are slightly modified:

6. summarize the audience's description using a set of features proposed by Robert Plutchik: expectation, anger, disgust, sadness, fear, surprise, hope, and trust [21];
7. form the emotional colour of each word from the audience description, giving the value of each feature in the range from 0 to 100;
8. summarize the obtained values, taking into account the sign of the emotion;
9. normalize the values in the range from 0 to 1;
10. consider the general market paradoxical situations mentioned earlier (Veblen, Giffen, snob effects) by introducing the appropriate indicator, which is calculated as the module of the difference between the normalized number indicator and the income share indicator.

We can formally describe the social disaster as follows:

$$\begin{cases} TASD = TAD \times \mu \times |CD - RD|, (RD \geq 0.5 \wedge CD \geq 0.5) \vee (RD < 0.5 \wedge CD < 0.5) \\ TASD = TAD, (RD > 0.5 \wedge CD < 0.5) \vee (RD < 0.5 \wedge CD > 0.5) \end{cases}, \quad (4)$$

where μ – coefficient of market paradoxes normalization, *TAD* – content analysis result for target audience description provided by the company, *CD* – normalized number indicator, *RD* – income share indicator.

It is worth noting that formula (4) does not consider the impact of a social disaster as such, so we introduce the following formula for the problem under consideration:

$$TAOI = TASD^{SSD}, \quad (5)$$

Each modified algorithms of vector autoregression treat external variables as numerical series, not scalars, so the indicator obtained as a result of calculation according to formula (5) must be vectorized. To achieve this, we'll use the obtained result as the centre of the normal distribution, which will serve as an exogenous variable for the selected algorithms of the modified EC-VAR family. Let's move on to the next indicator that needs to consider – the profile of the market situation. According to microeconomic theory, it can include a large number of different indicators. After surveying experts, the following characteristics were identified, which will be taken into account in the current paper:

- volumes of innovative activity;
- financial stability;

- market monopolization;
- the state of the target industry and the world economy;
- social disaster.

To measure market monopolization, we will use the Herfindahl-Hirschman Index:

$$HHI = \sum_{j=1}^{N} s_j^2 \quad (6)$$

where N – number of companies in the selected market, s – market share owned by each company.

Considering the fact that the e-commerce market has the characteristics of monopolistic competition with some regional manifestations of oligopoly, to find the approximate value of the index, it is enough to know the value of the market shares for the 5 largest companies on it. Innovative activity expressed by the relevant indicator is as follows:

$$IAI = \frac{IR}{MT}, \quad (7)$$

where IR – share of income from innovations, MT – total monetary mass of the company.

The resulting social catastrophe is characterized by the previously mentioned SSD indicator. The financial stability of the company is a classic economic indicator (FSI) that reflects the company's ability to meet its obligations in the long- and medium-term perspectives [22]. In general, we can write down the following formalized presentation of the above indicators of the company's activity:

$$MRI = \left(\frac{s_t^2 \times IAI \times FSI}{HHI}\right)^{SSD}, \quad (8)$$

As in the case of the TAOI indicator, a scalar value will be obtained during the calculation according to the formula (8). To vectorize it, we will use the result as the centre of the normal distribution. The obtained numerical series does not consider the state of the target industry and the world economy. To take into account the state of the economy, we will choose several classic indicators:

- world GDP level;
- prices for energy resources level;
- the S&P 500 index.

The indicated data are already numerical series, so to be able to use them as exogenous variables, it is enough to normalize them. We will determine the state of the industry based on the portfolio of shares of the five biggest companies in the e-commerce market. According to [23] the latest data:

- Jingdong Mall (ticker JD);
- Amazon (ticker AMZN);
- Meituan (ticker 3690.HK);
- Alibaba (ticker BABA);
- Pinduoduo (ticker PDD).

As in the previous case, the data are already numerical series, but they need to be normalized for the algorithms to work correctly. Finally, the following series will be considered as external variables for the family of modified EC-VAR models:

- target audience characteristics;
- company's activities on the market characteristics t;
- world economy state;
- industry state.

Let's move on to consider the theoretical basis of MapReduce technology, which will be used to parallelize the selected algorithms.

3.3. MapReduce technology

The essence of the approach, which is embedded in MapReduce technology, is to distribute the total data set to individual nodes. According to [24] the procedure is performed using mapping

functions, followed by the application of selected algorithms, and a reducer that collects data from all nodes and unifies them. In general, this technology is implemented in several frameworks, but within the framework of the current study, it was decided to use MapReduce, which is offered by Hadoop. Each implementation has its own features, so the architecture of this approach can be depicted as shown in Figure 1:

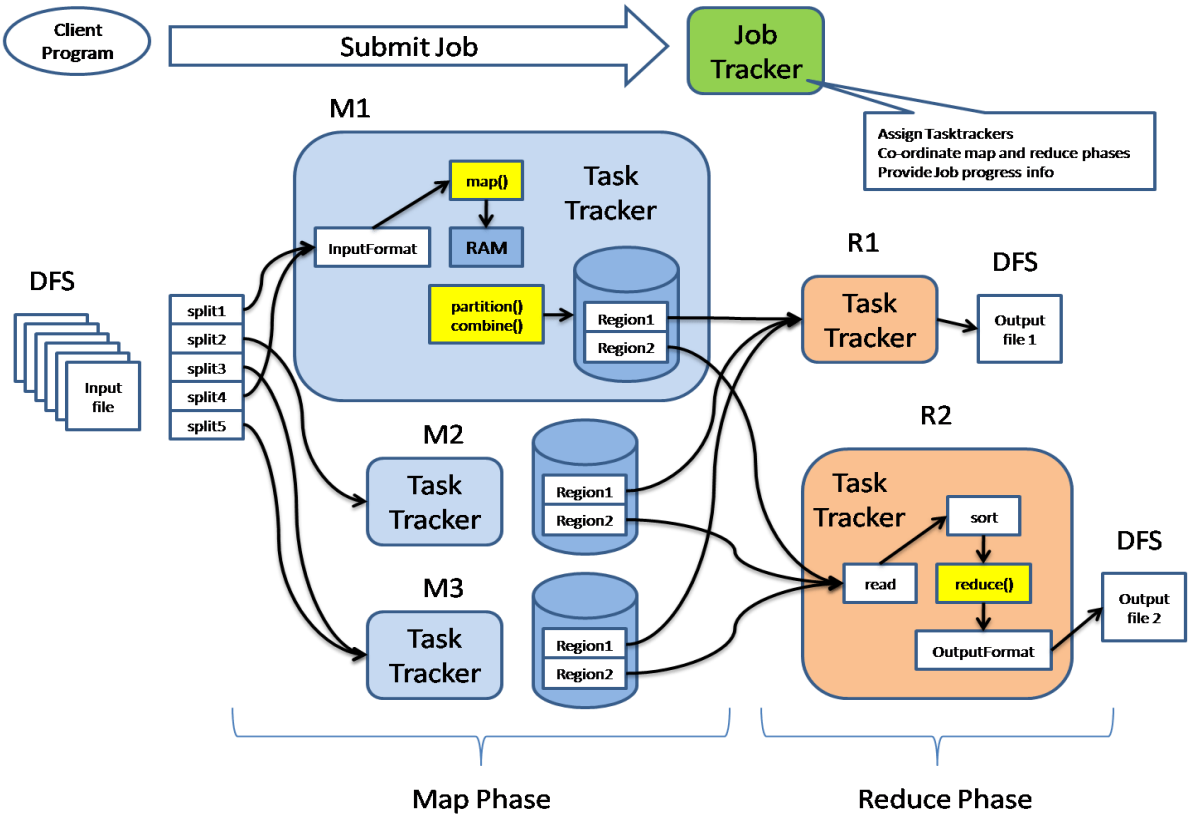


Figure 1: MapReduce in Hadoop

As we can see, in this case, in addition to the actual mapping and reduction functions, there are distribution and combination functions, which are necessary for the response from each node to arrive in a combined form. The results are sorted before the reduction. Among the advantages of the chosen approach, it is possible to single out scalability, relative cheapness, ease of use (although the technology is complicated by the need for configuration settings), and the possibility of monitoring execution. On the other hand, among the disadvantages, it is worth highlighting the need to create a large volume of programme code and the stealth of processing. There is also a need for a lengthy configuration setup.

4. Experimental environment

By the experimental environment we mean the following set of characteristics:

- general conditions;
- efficiency function;
- rule for comparing the efficiency of two models;
- errors and uncertainties.

We will gradually determine each of the specified characteristics of the environment.

4.1. General conditions

Given the peculiarities of the algorithms, it was decided to use the method of a controlled experiment. This became the basis for choosing a permanent and stable execution environment – a physical device based on Ubuntu with the following technical characteristics:

- CPU: Intel Core i5-1135G7;
- RAM: 16 Gb;
- SSD: 512 Gb;
- OS: Ubuntu 21.04.

To determine the execution speed, it was decided to use the Python 3 with libraries for natural language processing (nltk, re, etc.) and working with data (pandas, numpy, etc.). For MapReduce implementation, as mentioned above, it was decided to use Hadoop technology with 3 cores. Two test samples provided by Amazon [25] and Walmart [26] are used to determine accuracy. We will distribute the data according to the Pareto principle in a ratio of 80/20. Figure 2 present fragments of data from Walmart dataset:

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106

Figure 2: Walmart dataset

In Figure 3, the data for Amazon is slightly different from the previous one and contains information about fluctuations in the stock market, unlike the sales levels for Walmart:

	Date	Open	High	Low	Close	Adj Close	Volume
0	1997-05-15	2.437500	2.500000	1.927083	1.958333	1.958333	72156000
1	1997-05-16	1.968750	1.979167	1.708333	1.729167	1.729167	14700000
2	1997-05-19	1.760417	1.770833	1.625000	1.708333	1.708333	6106800
3	1997-05-20	1.729167	1.750000	1.635417	1.635417	1.635417	5467200

Figure 3: Amazon dataset

4.2. Efficiency function

As an efficiency function, to compare the VAR family algorithms with each other within the framework of the current study, we will consider the following function:

$$E = f(\vartheta, MSE, \nu) \tag{9}$$

where ϑ – speed of forecasting execution; *MSE* – mean square error of the forecast; ν – data preparation speed.

All performance metrics will be determined using the time module of the Python 3. We will calculate the *MSE* indicator using the following formula:

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{10}$$

where N – number of predicted values; y_i – real value; $\hat{y}_i$ – predicted value.

4.3. Efficiency comparison rule

To compare the efficiency of two models, we will use the following formula:

$$C_{AB} = \sum normalize(\frac{m_A}{m_B}), \tag{11}$$

where m_A – metric value for model *A*; m_B – metric value for model *B*.

The determined parameter *C* is fuzzy, so we will use the following set of rules to determine the most efficient model:

- $C_{AB} \geq 3.5$: model *A* is more efficient than *B*;
- $C_{AB} \leq 2.5$: model *A* is less efficient than *B*;
- $2.5 < C_{AB} < 3.5$: it's impossible to determine the most efficient model.

4.4. Errors and uncertainties

Within the framework of the proposed task and the described experiment, we can outline the following uncertainties and errors:

- during speed testing: human factor and instrumental error;
- during accuracy check: data problem.

To mitigate the impact of these problems, we set the number of measurements for the performance of parallelized and sequential modified autoregressive models equal to 10. The number of measurements for determining accuracy is equal to 5, and for the speed of preparation also 5.

5. Models implementation

The first step in implementation is to programming modified autoregressive models. To avoid errors in the implementation of successive versions, it was decided to use a library of the Python 3 – statsmodel. To create a MapReduce version, develop two script files for the reducer and the mapper. To simplify the perception, we show in Figure 4 the pseudocode for the mapping function:

```
Input: max window size & input record
Output: Addends of predicted value
for (i = 2 to w) do
    count =0
    Δi= i*(i+1)/2
    while (i ≥ count) do
        k' = j + '.' + (i - (count - 1))
        v' = data8count/Δi
        context.write(k', v')
        count++
    end while
end for #map
```

Figure 4: Pseudocode for mapping function

The next stage of implementation is the content analysis algorithm. As already mentioned, for this it was decided to use the nltk library, which contains a large number of different corpora of words. The Porter stemmer will be used as a base for stemming, and the WordNet lemmatizer will be used for lemmatization. The TF-IDF frequency characteristic will be obtained using the sklearn library. Part of the code that performs content analysis is shown in Figure 5. Data related to stock prices were chosen to be taken from the Alpha Vantage API. They will be combined into special data frames using pandas, a library that makes it easy to work with large amounts of data. Their aggregation and subsequent transformation for transmission as exogenous variables will be carried out at the expense of the already mentioned sklearn library. It was decided to take information related to the world economy as a whole in csv format from the database of the World Bank [27].

6. Experiment results

We present Figure 6 as a confirmation that the EC-modification of the algorithms by adding the processing of the previously given data was necessary. Figure 6 (actual value in orange and predicted value in dashed blue) showing a modified implementation of the moving average vector autoregression algorithm indicates that the data in the basic algorithm "a" doesn't consider the fluctuations, but the improved algorithm "b" provides fairly accurate results. Now we can proceed to the obtained results. Let's start with the execution time of the forecasting algorithms, information on 10 measurements is given in Table 1 (VARS – seasonal autoregression, VARL – distributed lag autoregression, VARMA – moving average autoregression).

```python
def process_text(text):
    stemmer = PorterStemmer()
    lemmatizer = WordNetLemmatizer()
    stop_words = set(stopwords.words("english"))
    sentences = nltk.sent_tokenize(text)
    sentences_processed = []
    for sentence in sentences:
        words_processed = []
        words = nltk.word_tokenize(sentence)
        words = [word for word in words if word not in stop_words]
        for word in words:
            word = stemmer.stem(word)
            word = lemmatizer.lemmatize(word)
            words_processed.append(word)
        words = [word for word in words_processed if word not in stop_words]
        sentences_processed.append(" ".join(words))
    text_processed = " ".join(sentences_processed)
    return text_processed
```

Figure 5: Python code for part of content analysis process

Table 1
Forecasting time

Sequential algorithm			MapReduce		
VARL	VARS	VARMA	VARL	VARS	VARMA
0.094 s	0.103 s	0.127 s	0.027 s	0.042 s	0.055 s
0.169 s	0.197 s	0.214 s	0.057 s	0.067 s	0.084 s
0.129 s	0.143 s	0.159 s	0.049 s	0.059 s	0.065 s
0.253 s	0.273 s	0.299 s	0.096 s	0.049 s	0.117 s
0.229 s	0.245 s	0.276 s	0.083 s	0.093 s	0.103 s
0.117 s	0.134 s	0.163 s	0.048 s	0.061 s	0.056 s
0.204 s	0.217 s	0.241 s	0.084 s	0.074 s	0.091 s
0.133 s	0.153 s	0.189 s	0.067 s	0.051 s	0.066 s
0.203 s	0.219 s	0.251 s	0.099 s	0.089 s	0.091 s
0.099 s	0.107 s	0.131 s	0.029 s	0.038 s	0.052 s

Let's find the average value for each case. We have 0.163 s for the serial VARL, and 0.064 s for the MapReduce version, we will get a speed gain of ~2.55. For VARS sequential we have 0.179 s, based on MapReduce technology – 0.062 s, we get a speed gain of ~ 2.89. For serial VARMA we have ~0.205 s, for the parallelized version – 0.078 s, we get a speed gain of ~ 2.63. So, the average speed gain is 2.69. The obtained value already allows us to state the expediency of using the technology for forecasting algorithms of the VAR family. To additional confirmation, the number of cores was increased to 4, as a result, the average speed gain for the set of models was 3.74. Now let's consider to comparing algorithms of the VAR family with each other.

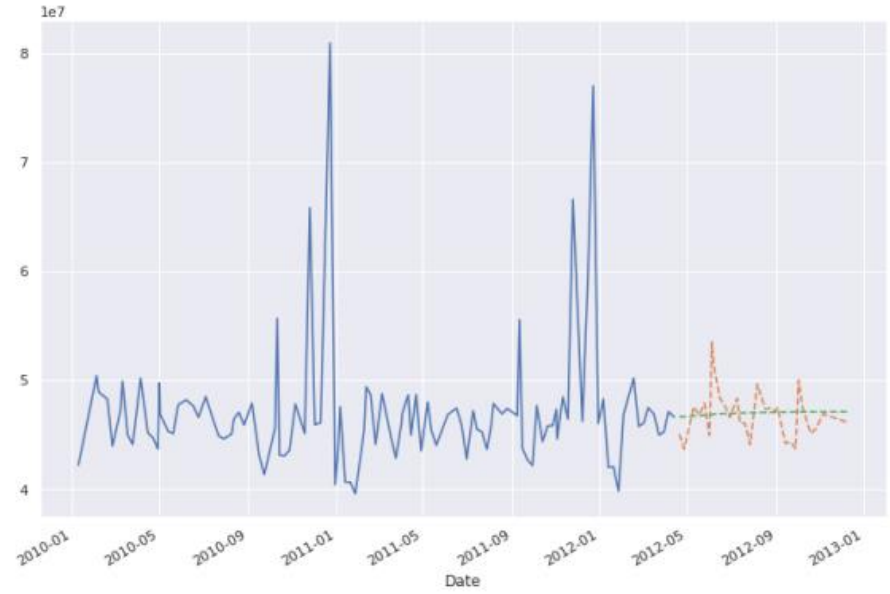
The first step of this is to check the accuracy of the forecast for the above samples. The obtained aggregated result of 5 measurements is shown in Table 2 (the result is rounded to a whole value).

Table 2
Forecasting accuracy

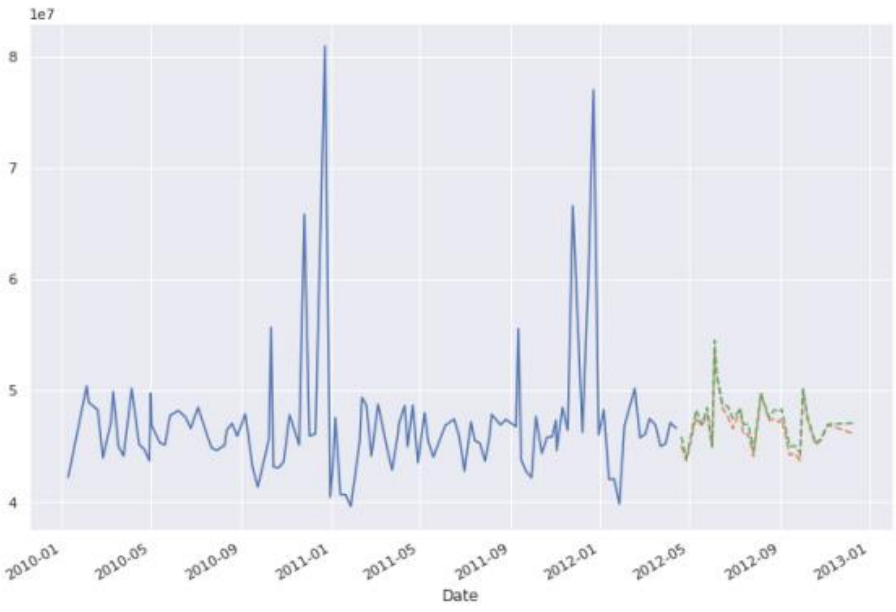
Dataset	VARL	VARS	VARMA
Walmart	87%	90%	93%
Amazon	91%	89%	97%

As we can see from the Table, in both cases, the moving average algorithm is the most accurate, which explains its wide applicability in econometric analysis. The resulting accuracy is based on the

fact that the moving average algorithm uses neighbouring values of the numerical series, this somewhat smoothest the overall forecast, making it more natural.



a



b

Figure 6: Forecasting results for Walmart's dataset

Regarding the time of data preparation, it is about 291 s for all algorithms. To reduce the impact of this time, it is necessary to parallelize the quality indicators processing. As part of the experiment, this was done using MapReduce technology (3 cores were used). As a result, the average preparation time was 104 s, i.e., the gain in speed is ~2.8. With an increase in the number of cores, the time was 78 s, that is, the gain in speed increased to ~3.73. In the case of comparing successive algorithms with each other using formula (11), it was determined that the VARMA algorithm is the most effective when considering the e-commerce market during social disasters (in both cases, the indicator exceeded 3). However, the use of parallelization, especially with 4 cores, indicates a significant time advantage of the parallelized versions, while the prediction accuracy remains unchanged. This allows us to state the expediency of implementing this version within the framework of the proposed task.

7. Conclusion

The purpose of this work was to determine the effectiveness of using MapReduce technology to parallelize autoregressive models for e-commerce market indicators during social disasters. For this purpose, an analysis of the subject area was carried out and the main methods of forecasting market indicators were classified, the reason for choosing vector autoregression models as the basis for further research was indicated. It was found that the nature of the input data requires modification of these models, considering the non-determinism of external factors.

During the theoretical examination of these models, their features were clarified and parallelization possibilities were outlined.

The choice of MapReduce, based on Hadoop technology is justified for the next three algorithms:

- vector autoregression of the moving average;
- vector autoregression of the distributed lag;
- vector seasonal autoregression.

The need for additional modification was determined to take into account exogenous variable indicators characterizing the microeconomic profile of the market situation and the characteristics of the target audience during social upheavals. In accordance with the tasks set, an experimental environment was constructed to test the feasibility of using the MapReduce technology and the created modification.

Based on all the above modifications and using the MapReduce technology, a series of experiments was conducted with the data of the companies Walmart regarding the level of sales, and Amazon regarding indicators of stock market activity. In the course of the experiments, it was found that the highest accuracy (in both cases more than 93%) is guaranteed by the use of vector autoregression of the moving average. At the same time, it was found that the gain in the speed of the algorithms during parallelization can reach 3.73, and in data preparation – 3.73.

Therefore, we can state that the use of MapReduce technology in combination with additional modifications of the basic algorithms, related to taking into account the non-determinism and high volatility of the indicators of the external environment, is expedient.

8. Acknowledgements

The authors would like to thank the Armed Forces of Ukraine for the opportunity to write a valid work during the full-scale invasion of the Russian Federation on the territory of Ukraine.

9. References

[1] N. J. Beeching, T. E. Fletcher, and R. Fowler. "COVID-19. BMJ best practices." BMJ Publishing Group. https://bestpractice.bmj.com/topics/ru-ru/3000201 (accessed Oct. 18, 2022).

[2] "COVID Data Tracker Weekly Review." Center for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html (accessed Nov. 11, 2022).

[3] Impact of the COVID-19 pandemic on trade and development. New York: United Nations Publications, 2020.

[4] J. Tollefson. "What the war in Ukraine means for energy, climate and food." Nature. https://www.nature.com/articles/d41586-022-00969-9 (accessed Oct. 30, 2022).

[5] D. Ratha. "A war in a pandemic - Implications of the Ukraine crisis and COVID-19 on global governance of migration and remittance flows." World Bank Blogs. https://blogs.worldbank.org/peoplemove/war-pandemic-implications-ukraine-crisis-and-covid-19-global-governance-migration-and (accessed Oct. 31, 2022).

[6] K. van der Zwet, A. I. Barros, T. M. van Engers, and P. M. A. Sloot, "Emergence of protests during the COVID-19 pandemic: quantitative models to explore the contributions of societal conditions," Humanities & Social Sciences Communications, vol. 9, 2022, Art. no. 68. Accessed: Nov. 1, 2022. [Online]. Available: https://doi.org/10.1057/s41599-022-01082-y.

[7] M. Tardzenyuy Thomas, "Protests during the 2020 COVID-19 lockdown in South Africa," MoLab Inventory of Mobilities and Socioeconomic Changes, 2021. Accessed: Nov. 1, 2022. [Online]. Available: https://doi.org/10.48509/MoLab.8077.
[8] J. Clement. "COVID-19: Steam user increase 2020." Statista. https://www.statista.com/statistics/1108322/covid-steam-users/ (accessed Oct. 31, 2022).
[9] C. G. Burton, S. Rufat, and E. Tate, "Social vulnerability: conceptual foundations and geospatial modeling," Cambridge University Press, pp. 53–81, 2018.
[10] Analyzing the social impacts of disasters, vol. 1, Methodology. Washington D. C.: GFDRR, 2015.
[11] R. B. Khodaparasti and S. Moslehi, "Application of the varma model for sales forecast: case of urmia gray cement factory," *Timisoara Journal of Economics and Business*, vol. 7, no. 1, pp. 89–101, 2014.
[12] A. Sinha and P. K. Jana, "MRF: MapReduce based Forecasting Algorithm for Time Series Data," Procedia Computer Science, vol. 132, pp. 92–102, 2018.
[13] G. M. Marakas, Decision Support Systems: In the 21st Century, 2nd ed. Hoboken: Prentice Hall, 2003.
[14] K. Srinivas, Process of Risk Management. IntechOpen, 2019. Accessed: Nov. 2, 2022. [Online]. Available: https://doi.org/10.5772/intechopen.80804.
[15] "Best IT risk management software." G2. https://www.g2.com/categories/it-risk-management (accessed Oct. 28, 2022).
[16] J. Li, C. Zhan, W. Sha, W. Jiang, and Y. Guo, "E-commerce Sales Forecast Based on Ensemble Learning," IEEE International Symposium on Product Compliance Engineering-Asia, 2020.
[17] W.-L. Chang and S.-T. Yuan, "A synthesized model of Markov chain and erg theory for behavior forecast in collaborative prototyping," Journal of Information Technology Theory and Application, vol. 9, no. 2, pp. 45–63, 2008.
[18] F. Engström and D. N. Rojas, Prediction of the future trend of e-commerce. Stockholm: KTH, 2021.
[19] A. Faehnle and M. Guidolin, "Dynamic Pricing Recognition on E-Commerce Platforms with VAR Processes," Forecasting, vol. 3, pp. 166–180, 2021.
[20] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," International Journal of Computer Applications, vol. 181, no. 1, pp. 25–29, 2018.
[21] H. Karimova. "The Emotion Wheel: What It Is and How to Use It." PositivePsychology. https://positivepsychology.com/emotion-wheel/ (accessed Nov. 1, 2022).
[22] W. R. Nelson and R. Perli, Selected Indicators of Financial Stability. Division of Monetary Affairs, 2020. Accessed: Jun. 1, 2022. [Online]. Available: https://www.ecb.europa.eu/pub/conferences/shared/pdf/jcbrconf4/Perli.pdf.
[23] "Largest e-commerce companies by market cap." Global ranking. https://companiesmarketcap.com/e-commerce/largest-e-commerce-companies-by-market-cap/ (accessed Nov. 10, 2022).
[24] J. Lin and C. Dyer, Data-Intensive Text Processing with MapReduce. College Park: University of Maryland, 2010. Accessed: Nov. 10, 2022. [Online]. Available: https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf.
[25] J. H. Roveda. "Historical Amazon stock prices." kaggle. https://www.kaggle.com/datasets/josehenriqueroveda/historical-amazon-stock-prices (accessed Oct. 19, 2022).
[26] A. Ahmedov. "Walmart Sales Forecast." kaggle. https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast (accessed Oct. 19, 2022).
[27] "World Bank Open Data." WorldBank Data. https://data.worldbank.org/ (accessed Nov. 1, 2022).