Analysis of Geo-Economic Distribution of Scientific Publications Citation and Self-Citation Standardized Indices Based on Machine Learning

Bohdan Korostynskyi [1], Oleksandr Mediakov [1], Victoria Vysotska [1,2], Oksana Markiv [1] and Michał Duda[3]

[1] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *Osnabrück University, Friedrich-Janssen-Str. 1, Osnabrück, 49076, Germany*
[3] *University of Warmia and Mazury in Olsztyn, Michała Oczapowskiego Street, 2, Olsztyn, 10719, Poland*

Abstract

The article dwells upon the logical order of processing, transformation and synthesis of data windows, their visualization and analysis for geo - economic distribution research of articles authorship numerical characteristics, their citation, estimation, lack of linear and nonlinear relationships between individual parameters of author and percentage of self-citation. The work demonstrates the possibility of using new and classical methods of data visualization to study patterns, relationships between numerical and nominal data as well as methods of using conventional multilayer perceptrons to search for nonlinear relationships between multiple parameters. Open source software designed to build the necessary representations of data and models is the important part of the investigation.

Keywords

Scientific Publications Citation, Self-Citation Standardized Indices, Machine Learning, dataset, neural network, correlation matrix, correlation analysis, regression analysis

1. Introduction

General development of mankind and globalization has caused the fact that most modern advances in various fields of human activity are spreading with incredible speed. The authorship, sale or distribution of articles, books and works are made through a large number of different platforms that are available to anyone. Such systems are the basis of virtuous exchange of knowledge.

Modern scientific works are always based on certain previous results and in accordance with the generally accepted principles of the scientific world, the source of a certain result is always indicated. Accordingly, each quality article has its own list of references, which from the point of view of the author of these works, is his citation.

It is logical that many works cite many other works, forming complex graph-like connections. Using numerical indicators or certain indices, they form ratings of the most influential, popular or most productive authors in the world.

Since the position of the author can be increased by self-citation, there is a cluster of individuals who abuses sit. That is why one of the tasks of this work is to check the relationship between different author numerical and nominal characteristics with the level of his self-citation.

However, the development of the scientific world has different bases, including financial, which can be extended to the geographical and economic distribution of scientific capacity, in the perspective of

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland

EMAIL: bohdan.korostynskyi.sa.2019@lpnu.ua (B. Korostynskyi); Oleksandr.Mediakov.sa.2019@lpnu.ua (O. Mediakov); victoria.a.vysotska@lpnu.ua (V. Vysotska); oksana.o.markiv@lpnu.ua (O. Markiv); michal.duda@uwm.edu.pl (M. Duda)
ORCID: 0000-0002-0545-6264 (B. Korostynskyi); 0000-0002-2580-3155 (O. Mediakov); 0000-0001-6417-3689 (V. Vysotska); 0000-0002-1691-1357 (O. Markiv); 0000-0002-9174-3840 (M. Duda)

CEUR Workshop Proceedings (CEUR-WS.org)

writing articles, citations, indices of authors and more. Thus, another task set in the paper is to study the geographical distribution, as well as economic, by using certain economic and social indicators.

2. Related Works

The first researches and works related to the evaluation of authors citation are related to the derivation of correct index or rating formula. Data set used in the work contains such two indices: h-index, hm-index. First of all, the quality of the author citation represetation using the h-index is debatable, because many scholars and publicists argue the low efficiency and incompetence of this approach [1-4].Since this work aims to study self-citation and by comparing the h-indices of the authors with and without self-citation, the question of study correctness has been emerged. Many papers dwells upon direct citation analysis limited to a specific cluster of authors tied to publishing platforms. In the result of the analysis, data about distribution of citations by person partners as a physical being, as well as by their intellectual characteristics or areas such as education have been obtained [1,3].

The problem of self-citation research is quite popular [5-9]. It leads to the availability of many sources of data related to it. The study of these data is limited to cluster, regression analysis. So this paper describes experiments using neural networks to identify or refute the existence of nonlinear, complex relationships between many possible parameters of the author and his level of self-citation, to identify patterns or limited human resources in certain areas. However, when the issue of citation is extended to authors nationalities, there is a gap in the availability of intelligence analysis of the number of self-citations in parts of the world or countries, as most studies are unidirectional.

3. Methods and Materials

The study of relationships between different parameters, proving the presence or absence of these relationships, as well as establishing their analytical form requires certain mathematical and algorithmic apparatus, such as classical methods, correlation analysis, regression analysis, least squares method for approximation [5], backpropagation algorithm as a method of learning neural networks (including multilayer perceptrons) and Deep Learning method [10-13]. Some algorithms used in the work have a high level of complexity and do not require additional changes or descriptions [14-21], because it does not meet the objectives of this research. They are actuality for content analysis or web resources monitoring based on machine learning [22-30], for example, citation and self-citation analysis in scientific and technical articles for dataset formation [31-36]. Such algorithms include the Adam method - a stochastic algorithm for finding the numerical value of the FBZ gradient. However, some algorithms for generating the required type of data, their visualization or tokenization have been supplemented or changed specifically for this research. Some of them are given as an example based on algorithm supplemented algorithm for filling the correlation matrix and the corresponding heat map:

Step 1. Initiate a single matrix of the attribute’s numbers order. The value that are not on the main diagonal equals to the special NaN value.

Step 2. Choose the type of triangular matrix.

Step 3. Pass cyclic the cells of matrices with incremental parameters i, j

Step 3.1. If $i \neq j$ belongs to the matrix

Step 3.1.1. If the upper triangular matrix is chosen and $i>j$ (in cell i, j), then assign the value of Pearson correlation coefficient [5] between the attributes of numbers i and j.

Step 3.1.2. If the lower triangular matrix is chosen and $i<j$ (in cell), then assign the correlation coefficient value between the attributes by numbers i and j (in the cell i and j).

Step 3.2. If $i = j$, then skip step

Step 4. Choose two color nodes.

Step 5. Interpolate the color nodes values on the interval [-1, 1], for NaN values it is necessary to return transparent color.

Step 6. Fill the matrix with the appropriate color values.

Step 7. Display the color table according to the values of the matrix cells

Another example is tokenization, which is used to convert some string fields of data windows into corresponding numerical vectors. Fig. 1 shows the activity diagram describing this algorithm.

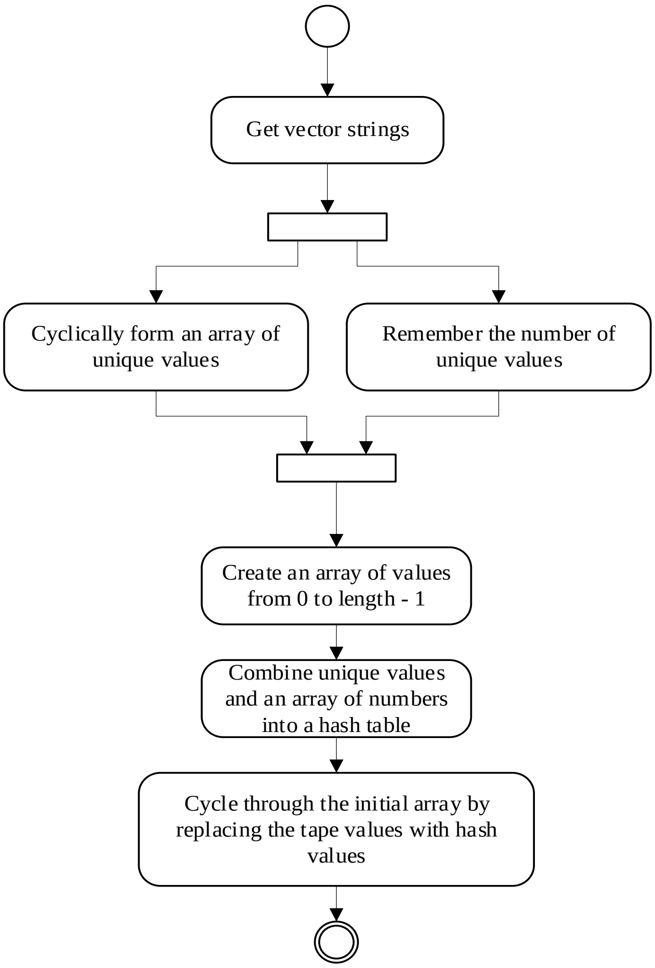

Get vector strings
Cyclically form an array of unique values
Remember the number of unique values
Create an array of values from 0 to length - 1
Combine unique values and an array of numbers into a hash table
Cycle through the initial array by replacing the tape values with hash values

Figure 1: UML diagram of activity and tokenization

4. Experiment

The study of selected parameters distribution is based on several datasets, including data about 160 thousand of unique authors [8], world GDP information in 2020, as well as dataset of the global distribution of corruption indices - CPI, 2019.

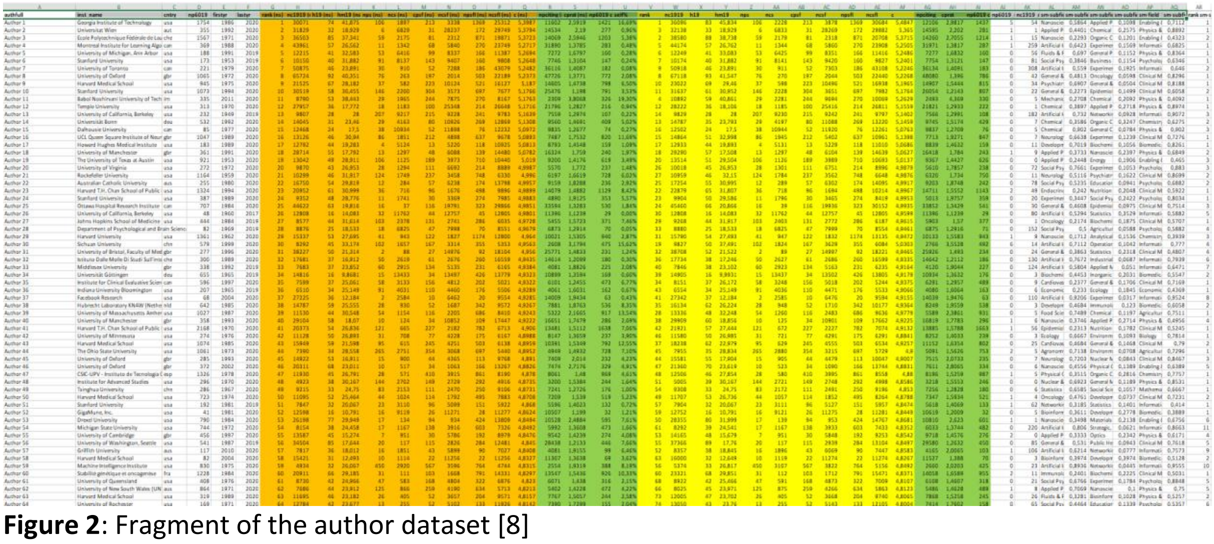


Figure 2: Fragment of the author dataset [8]

The first dataset about authors contains 46 fields and more than 161,441 lines. Each field has an encrypted name, the description of which is given in Fig. 2. However, only some of them were actually used in the study, namely: author country, number of author articles for the period 1960 - 2019, number of author citations for 2019 (including and exclusively with self-citation - Fig. 3), author h-index for 2019 (including and exclusively with self-citation - Fig. 4), as well as the top category of the author (from the categories of ScienceMetrix). For example, for general statistics within 12% of self-citation - 160 thousand; with <1% - about 9700 authors; from> 12% - approximately 65400 people, and from> 30% - approximately 7500 (Fig. 6-7).

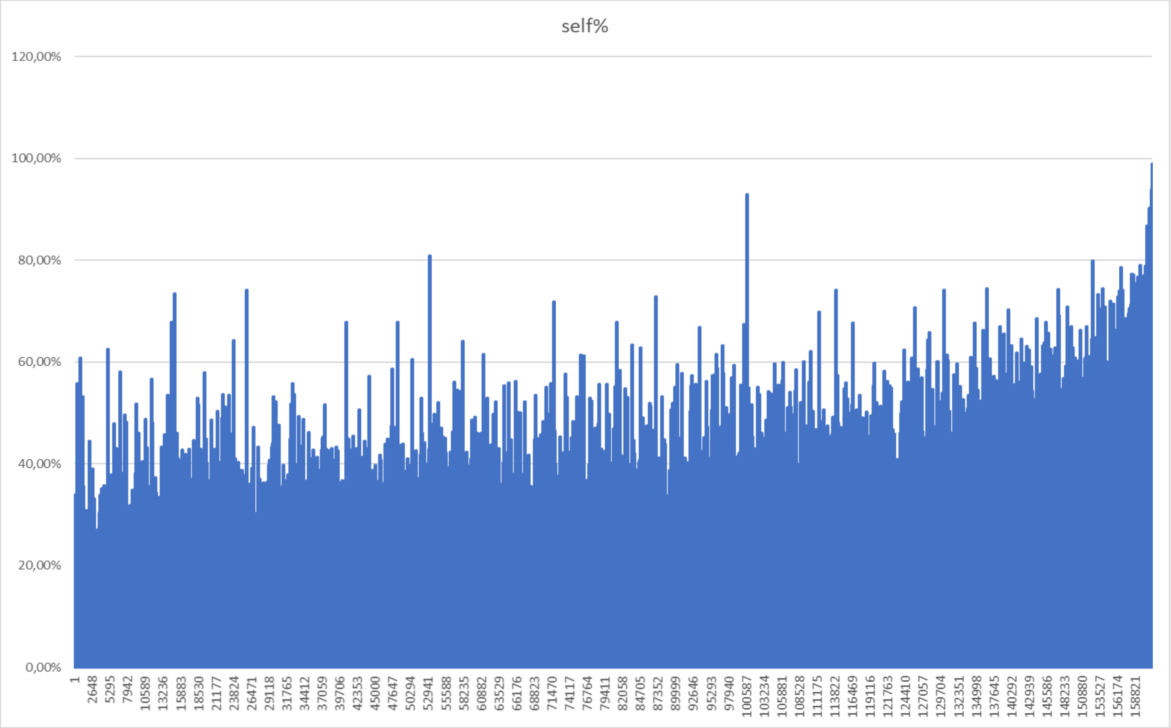


Figure 3: Growth of self-citation in the same year

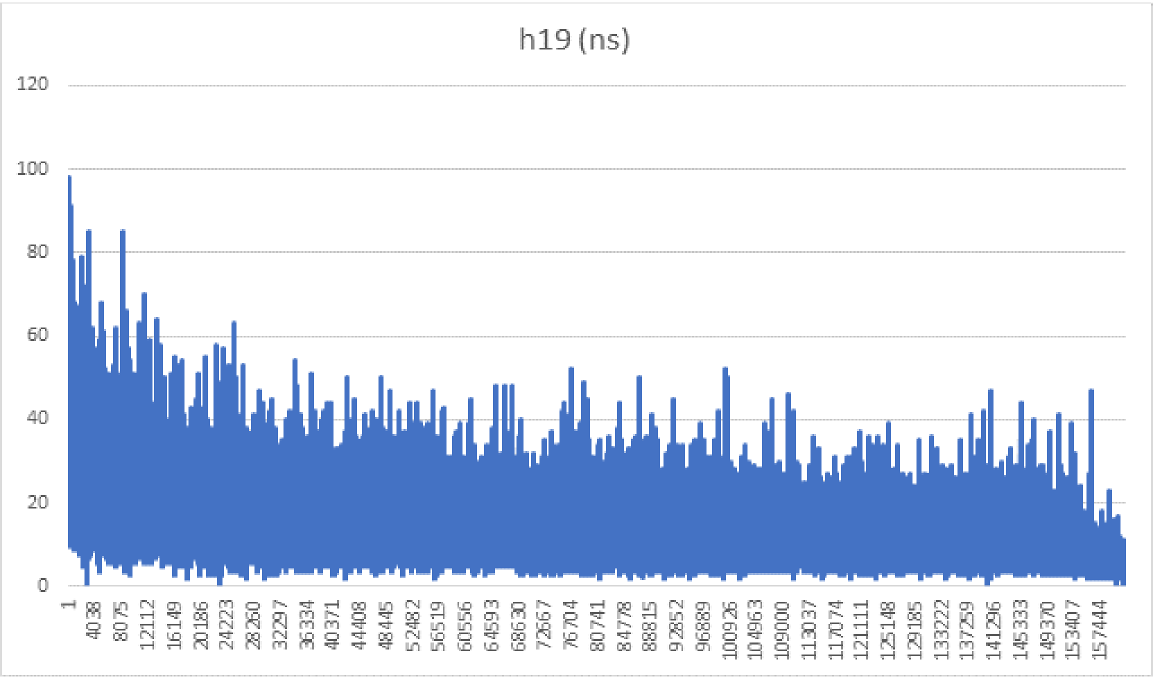


Figure 4: H-index growth at the end of 2019

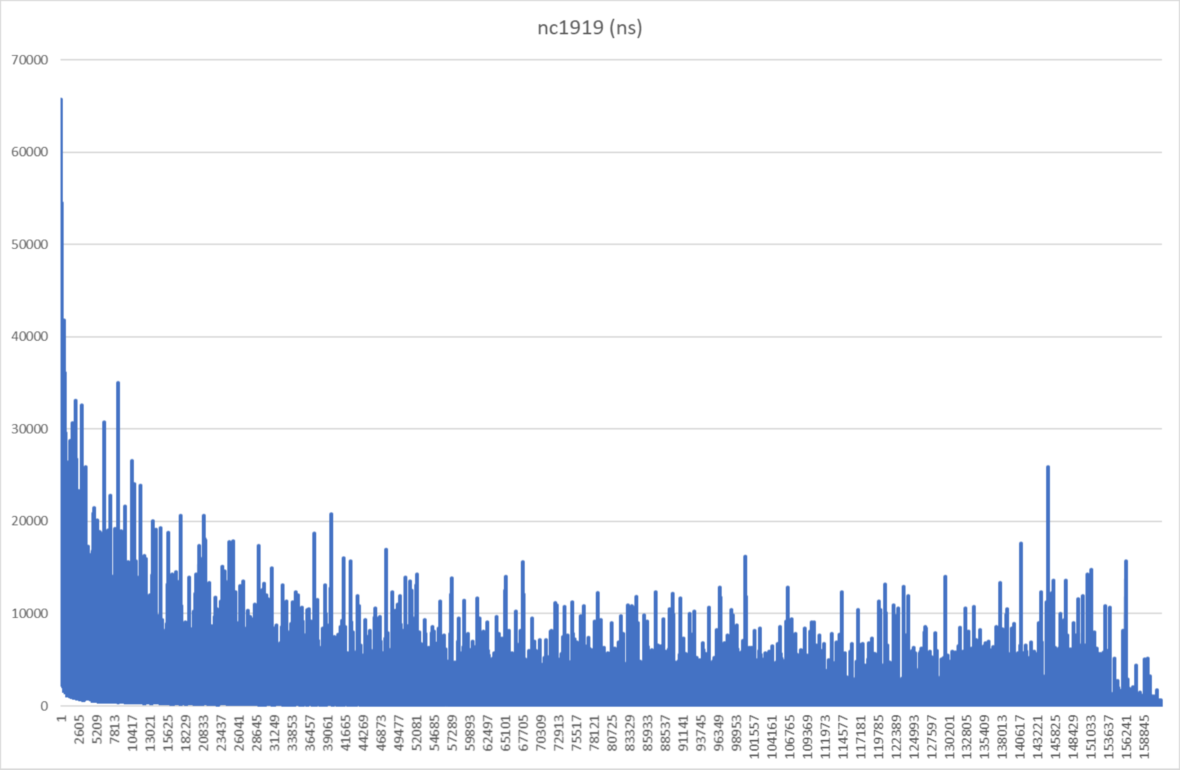


Figure 5: Change of links for the same year

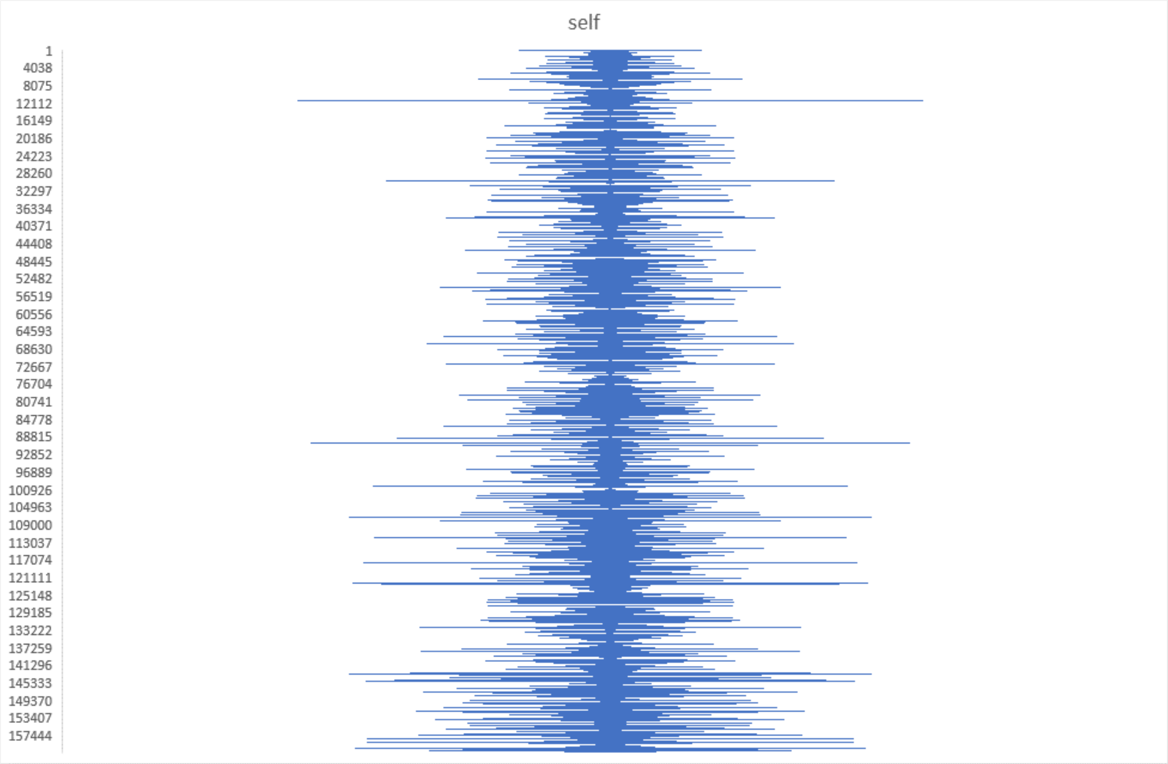


Figure 6: Growth of self-citation in the same year

The next dataset is the data on GDP of countries according to the World Bank (Fig. 8) [6]. From this dataset, two fields were used in the work - the ISO3 representation of the country name and the actual value of GDP.

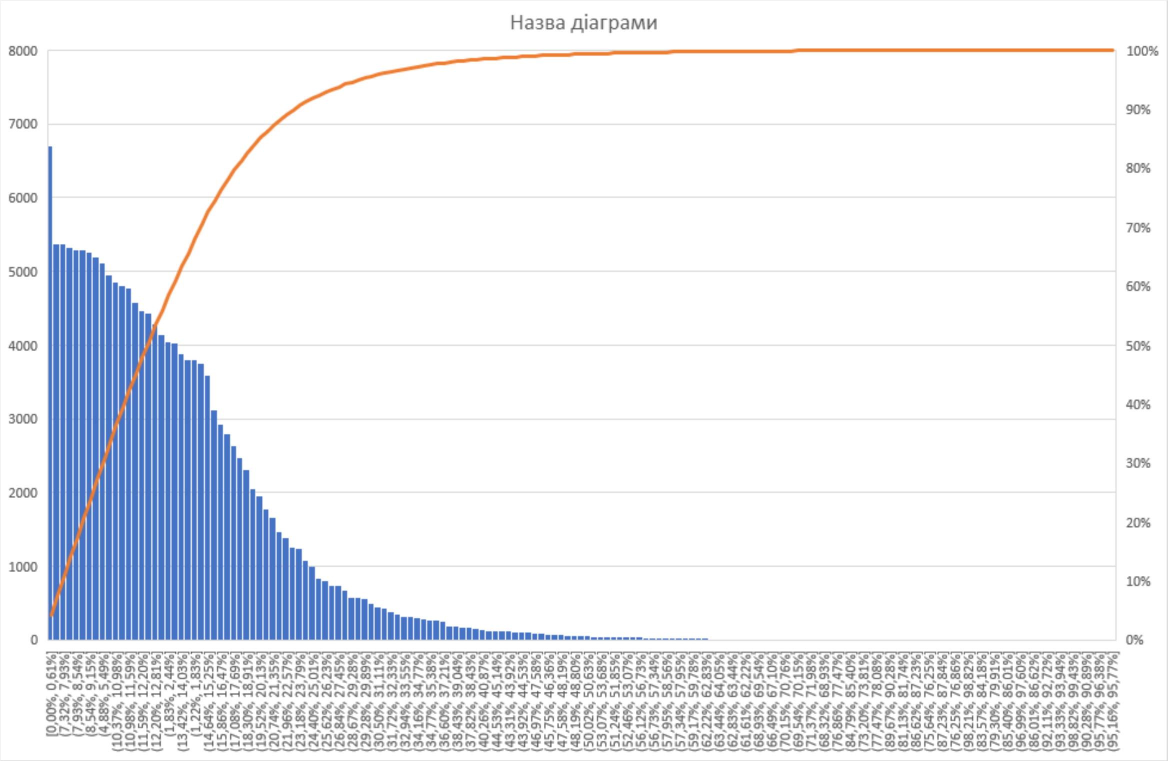
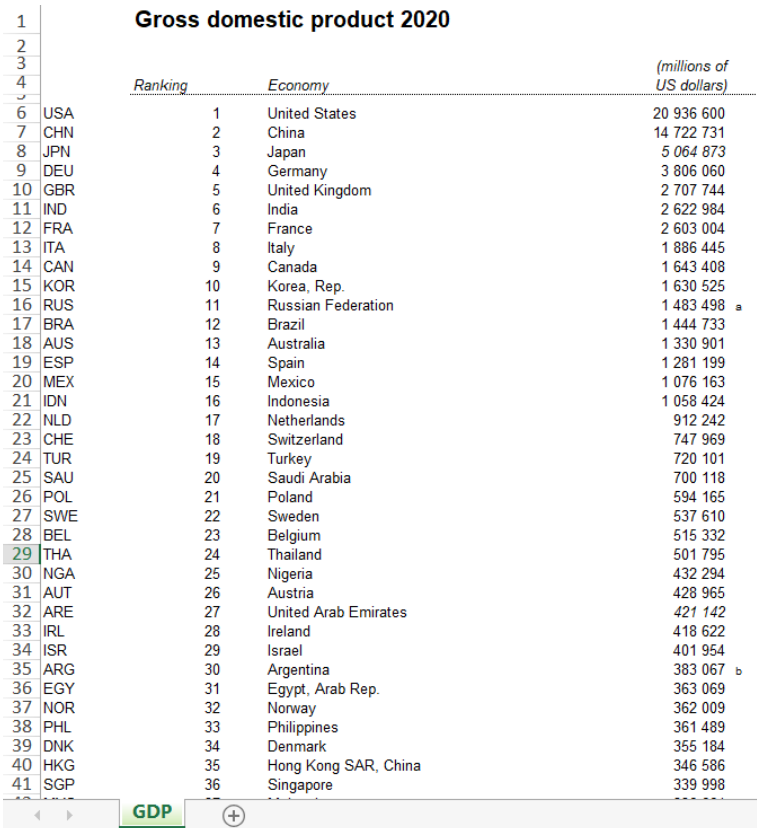


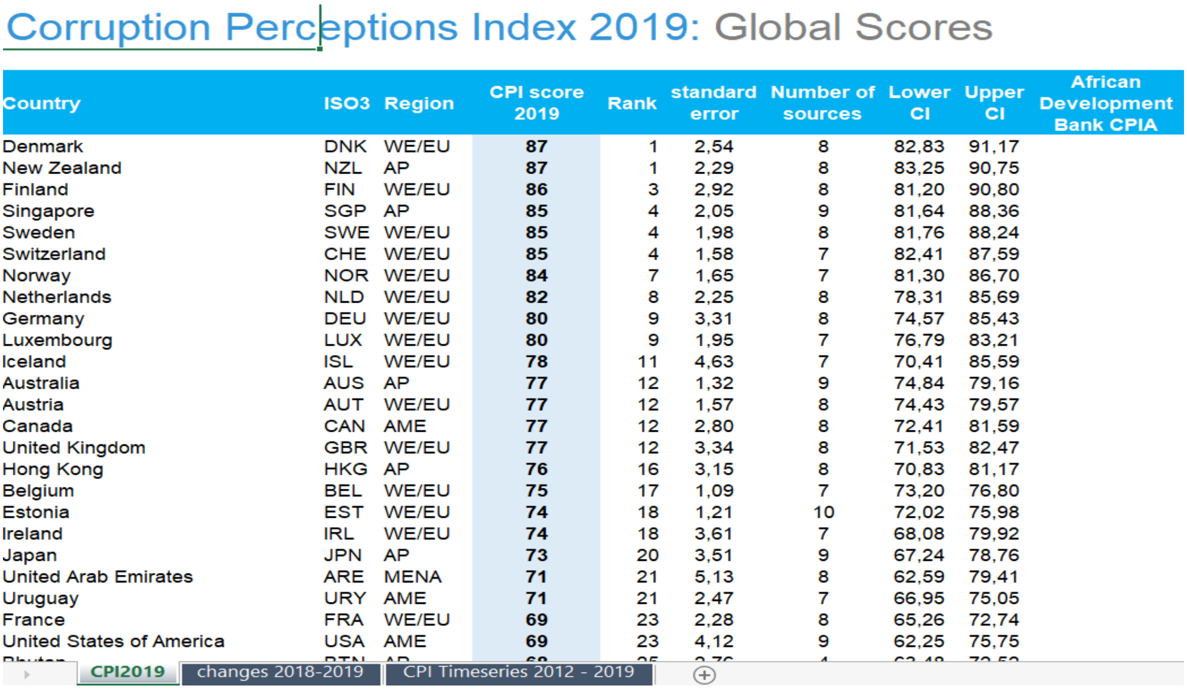
Figure 7: Growth of self-citation in the same year



1		Gross domestic product 2020		
2				
3				(millions of
4		*Ranking*	*Economy*	*US dollars)*
6	USA	1	United States	20 936 600
7	CHN	2	China	14 722 731
8	JPN	3	Japan	*5 064 873*
9	DEU	4	Germany	3 806 060
10	GBR	5	United Kingdom	2 707 744
11	IND	6	India	2 622 984
12	FRA	7	France	2 603 004
13	ITA	8	Italy	1 886 445
14	CAN	9	Canada	1 643 408
15	KOR	10	Korea, Rep.	1 630 525
16	RUS	11	Russian Federation	1 483 498 a
17	BRA	12	Brazil	1 444 733
18	AUS	13	Australia	1 330 901
19	ESP	14	Spain	1 281 199
20	MEX	15	Mexico	1 076 163
21	IDN	16	Indonesia	1 058 424
22	NLD	17	Netherlands	912 242
23	CHE	18	Switzerland	747 969
24	TUR	19	Turkey	720 101
25	SAU	20	Saudi Arabia	700 118
26	POL	21	Poland	594 165
27	SWE	22	Sweden	537 610
28	BEL	23	Belgium	515 332
29	THA	24	Thailand	501 795
30	NGA	25	Nigeria	432 294
31	AUT	26	Austria	428 965
32	ARE	27	United Arab Emirates	*421 142*
33	IRL	28	Ireland	418 622
34	ISR	29	Israel	401 954
35	ARG	30	Argentina	383 067 b
36	EGY	31	Egypt, Arab Rep.	363 069
37	NOR	32	Norway	362 009
38	PHL	33	Philippines	361 489
39	DNK	34	Denmark	355 184
40	HKG	35	Hong Kong SAR, China	346 586
41	SGP	36	Singapore	339 998

Figure 8: Part of the data from the file

The third dataset (Fig. 9) is one used in the work, the information in which is obtained from Transparency International – the international organization against corruption [2]. This dataset is suitable for separate study, but in this paper is used as a source of information to test the hypothesis.



Corruption Perceptions Index 2019: Global Scores

Country	ISO3	Region	CPI score 2019	Rank	standard error	Number of sources	Lower CI	Upper CI	African Development Bank CPIA
Denmark	DNK	WE/EU	**87**	1	2,54	8	82,83	91,17	
New Zealand	NZL	AP	**87**	1	2,29	8	83,25	90,75	
Finland	FIN	WE/EU	**86**	3	2,92	8	81,20	90,80	
Singapore	SGP	AP	**85**	4	2,05	9	81,64	88,36	
Sweden	SWE	WE/EU	**85**	4	1,98	8	81,76	88,24	
Switzerland	CHE	WE/EU	**85**	4	1,58	7	82,41	87,59	
Norway	NOR	WE/EU	**84**	7	1,65	7	81,30	86,70	
Netherlands	NLD	WE/EU	**82**	8	2,25	8	78,31	85,69	
Germany	DEU	WE/EU	**80**	9	3,31	8	74,57	85,43	
Luxembourg	LUX	WE/EU	**80**	9	1,95	7	76,79	83,21	
Iceland	ISL	WE/EU	**78**	11	4,63	7	70,41	85,59	
Australia	AUS	AP	**77**	12	1,32	9	74,84	79,16	
Austria	AUT	WE/EU	**77**	12	1,57	8	74,43	79,57	
Canada	CAN	AME	**77**	12	2,80	8	72,41	81,59	
United Kingdom	GBR	WE/EU	**77**	12	3,34	8	71,53	82,47	
Hong Kong	HKG	AP	**76**	16	3,15	8	70,83	81,17	
Belgium	BEL	WE/EU	**75**	17	1,09	7	73,20	76,80	
Estonia	EST	WE/EU	**74**	18	1,21	10	72,02	75,98	
Ireland	IRL	WE/EU	**74**	18	3,61	7	68,08	79,92	
Japan	JPN	AP	**73**	20	3,51	9	67,24	78,76	
United Arab Emirates	ARE	MENA	**71**	21	5,13	8	62,59	79,41	
Uruguay	URY	AME	**71**	21	2,47	7	66,95	75,05	
France	FRA	WE/EU	**69**	23	2,28	8	65,26	72,74	
United States of America	USA	AME	**69**	23	4,12	9	62,25	75,75	
Bhutan	BTN	AP	**68**	25	2,76	4	63,48	72,52	

Figure 9: Part of excel datasheet

In order to use, convert and process data from datasets using R the best is to translate the obtained data into csv format. Since one of the investigation tasks is to study the percentage of self-citation depending on other parameters, and in the initial data they are given as a tape value, it is necessary to develop a program for converting a column of data, for example:

```
aut.df <- read.csv(file = "authors_dataset.csv", sep = ";", header = T, dec = ",")
aut.df %>% mutate(self.p = as.double(gsub("[,]", ".", gsub(".{1}$", "", self.)))) -> aut.df
```

When working with data approaches and technologies, the tidy verse group has been used. Since the important work is the visualization and representation on graphs of many parameters simultaneously, the basic principles of developing such graphs using ggplot2, ggridges and patchwork have been considered. The main type of graphs is lollipop. The developed template looks like this:

```
# lollipop diagrams
th <- theme_light() + theme(
    panel.grid.major.x = element_blank(),
    panel.border = element_blank(),
    axis.ticks.x = element_blank(),
    axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
    axis.title.x = element_blank())
# F
lp.F.top60 <- ggplot(cit.df, aes(x = cntry, y = F)) +
    geom_segment(aes(x = cntry, xend = cntry, y = 0, yend = F), color = "grey") +
    geom_point(col = "orange", size = 3) + th
```

Replacement of the variable F with any available in the data frame will allow to build different graphs, for example from Fig. 10. Classic point graphs with a regression line also are used. There is an example code to create:

```
# point charts
(ggplot(data.df) +
    geom_point(aes(x = h, y = cit19, colour = self.p), size = 3.5, alpha = 0.8) +
    labs(title = "citation - h.index") + theme_light()) -> h.cit.plot
```

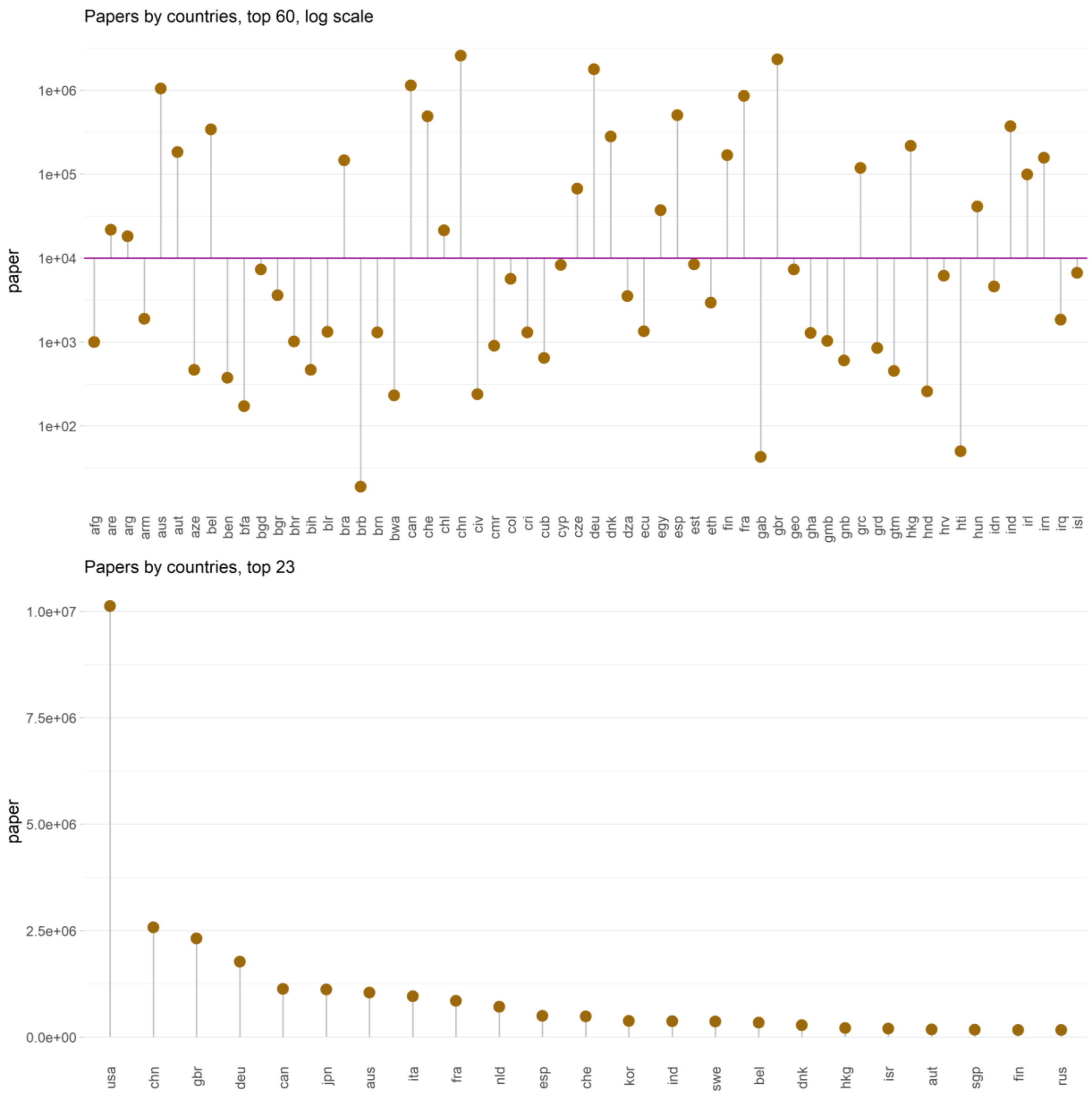


Figure 10: Example of lollipop chart

Another important type of displaying the relationship between data is the thermal map, the algorithm of which was described in the previous section, in particular, the supplemented algorithm can be performed as follows:

```
# lower-triangular, transposed thermal map of the correlation matrix
data.df %>% cor() %>% melt() %>%
    arrange(Var1) %>%
    group_by(Var1) %>%
    filter(row_number() <= which(Var1 == Var2)) %>%
    ggplot(aes(x = Var1, y = Var2, fill = value)) + geom_tile(color = "white") +
      scale_fill_gradient2(low = "#040035", high = "#c50000", mid = "#fffebb",
        midpoint = 0, limit = c(-1,1), space = "Lab", name="Cor coef") +
      labs(x = "", y = "") + geom_text(aes(label = round(value, 3))) +
      theme_bw() -> heat.map
```

Relatively new type of many samples' properties representation, their distribution or relationship are the so-called joyplot or ridgeline. In this paper it is performed using point gradient by the following code:

```
data.df %>%
    melt() %>% # "Melting" to convert to a key-value data window
    ggplot(aes(x = sqrt(value), y = variable, fill = sqrt(..x..), xmin = 0, xmax = 1)) +
    geom_density_ridges_gradient() + # ridgeline з ggridges
    scale_fill_gradient2(low = "#220061", high = "#ff9100", mid = "#e87059",
        midpoint = 0.65, limit = c(0, 1), space = "Lab", name ="Norm.values(sqrt scale)") +
    theme_ridges() +
    theme(axis.title.x = element_blank(),axis.title.y = element_blank(),) -> plot.rid
```

The results of this graph type are shown in Fig.13.

After describing features of building data graphical representation, construction of some algorithms for data processing has been demonstrated. First of all, the method of data conversion of the authors dataset has been considered. It is necessary to combine all records by country and then convert the grouped fields into new ones, for example, the number of published articles by authors of one country should be summed, the level of self-citation should be turned into the average one, and the h-index of country authors should be generalized.

```
# greeting of the original dataset
aut.df <- read.csv(file = "authors_dataset.csv", sep = ";", header = T, dec = ",")
aut.df %>% mutate(self.p = as.double(gsub("[,]",".",gsub(".{1}$", "", self.)))) -> aut.df
# h-index calculation function
h.index <- function(x) {
    x <- sort(x, decreasing = T)
    sum(x >= seq_along(x)) %>% return
}
# field sampling, grouping and transformation
aut.df[c("cntry", "np6019", "nc1919..ns.", "nc1919", "self.p", "h19..ns.", "h19")] %>%
group_by(cntry)  %>%  summarise(sum(np6019),  sum(nc1919..ns.),  sum(nc1919),  mean(self.p),
h.index(h19..ns.), h.index(h19)) %>%
data.frame() -> countr.df
# rename fields
names(countr.df) <- c("cntry", "paper", "cit19.ns", "cit19", "self.p", "h.ns", "h")
countr.df %>% arrange(desc(h)) %>%
mutate(cntry=factor(cntry, levels=cntry)) -> countr.df
write.csv(countr.df, "citation.csv", row.names = F, quote = F)
```

After creating the new data window, it is possible to merge it with two others that have a common field - country names. For croquet combination without forming NA values, firstly, it is needed to find the intersection of three sets of existing countries in each dataset and then select only those countries that are part of the new set:

```
# read three datasets
gdp.df <- read.csv("gdp.csv", dec = ",", sep = ";")
cit.df <- read.csv("citation.csv")
cpi.df <- read.csv("cpi19.csv", dec = ",", sep = ";")
# intersection of sets of countries by iso3
intersect(cit.df$cntry, gdp.df$cntry) %>% intersect(tolower(cpi.df$ISO3)) -> cntrs
gdp.df[gdp.df$cntry %in% cntrs,] %>% arrange(cntry) -> gdp.df
cit.df[cit.df$cntry %in% cntrs,] %>% arrange(cntry) -> cit.df
cpi.df[tolower(cpi.df$ISO3) %in% cntrs,] %>% arrange(ISO3) -> cpi.df
# pasting a new data window
cbind(cit.df, gdp.df, cpi.df) %>% data.frame() -> data.df
```

In the part of the work responsible for the study of the considered properties geo-economic distribution properties, the deviation to the exponential regression has been performed. If to use the built-in function lm and compose the formula as y ~ exp (x), then there will be a problem that the best parameters for the regression function f (x) = c + exp (x) will be selected.

However, in the general case, the exponential function has the form of a * b ^ x, which is why the formula ln (y) ~ x is chosen to perform the correct regression analysis, which is equivalent to the function ln (f (x)) = c1 + c2x, which is equal to f (x) = exp (c1) exp (c2x), where a = exp (c1), and b = exp (c2):

```
# exponential regression model
h_cit19.model <- lm(log(data.df$cit19) ~ data.df$h)
# using the model to calculate the approximation f-tion
exp(predict(h_cit19.model, newdata = list(data.df$h)))
```

The last important part of the experiments descriptions is the neural networks development. The paper represents four different in architecture and set of activation functions, single and multilayer perceptrons, i.e. fully connected networks (Fig. 11).

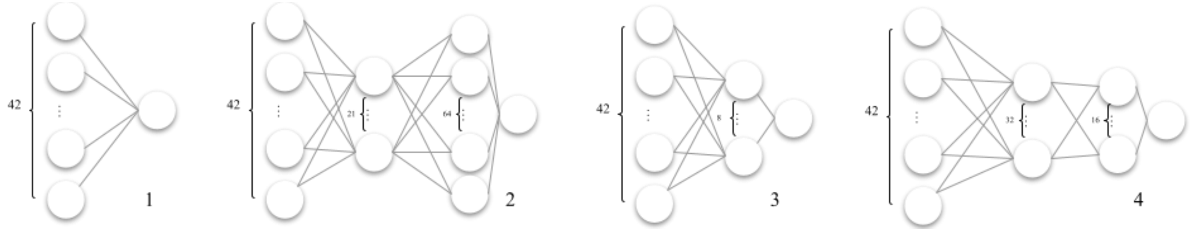


Figure 11: Topology of four perceptron models

TensorFlow was used as a library for python to build perceptrons and the model itself was built using the Functional API:

```
def get_model(type: int, shape: int) -> tf.keras.Model:
    inputs = tf.keras.Input(shape = (shape))
    if type == 0:
        output = tf.keras.layers.Dense(1, activation="sigmoid")(inputs)
        return tf.keras.Model(inputs, output)
    if type == 1:
        x = tf.keras.layers.Dense(21, activation="sigmoid")(inputs)
        x = tf.keras.layers.Dense(64, activation="sigmoid")(x)
        output = tf.keras.layers.Dense(1, activation="sigmoid")(x)
        return tf.keras.Model(inputs, output)
    if type == 2:
        x = tf.keras.layers.Dense(8, activation="relu")(inputs)
        output = tf.keras.layers.Dense(1, activation="sigmoid")(x)
        return tf.keras.Model(inputs, output)
    if type == 3:
        x = tf.keras.layers.Dense(32, activation="selu",
          kernel_initializer='lecun_normal')(inputs)
        x = tf.keras.layers.Dense(16, activation="relu")(x)
        output = tf.keras.layers.Dense(1, activation="sigmoid")(x)
        return tf.keras.Model(inputs, output)
self_models = []
for i in range(4):
    self_models.append(get_model(i, 36))
    self_models[i].compile(optimizer = tf.keras.optimizers.Adam(),
        loss = "mse", metrics = ['accuracy'] )
for i in range(4):
    print(i)
    self_models[i].fit(train_ds, epochs=3)
for i in range(4):
    print(i)
    self_models[i].evaluate(test_ds)
```

To build a dataset for learning and testing models from the initial worksheet about the authors developed code:

```
library(dplyr)
aut.df <- read.csv(file = "authors_dataset.csv", sep = ";", header = T, dec = ",")
aut.df %>% mutate(self.p = as.double(gsub("[,]", ".", gsub(".{1}$", "", self.)))) -> aut.df
# removal of NA-local columns
aut.df[, c(-1, -7, -21, -22, -39, -41, -43, -44, -45, -46)] -> aut.df
# tokenization
l <- unique(aut.df$inst_name)
aut.df$inst_name <- as.numeric(factor(aut.df$inst_name, levels=l))
l <- unique(aut.df$cntry)
aut.df$cntry <- as.numeric(factor(aut.df$cntry, levels=l))
l <- unique(aut.df$sm.subfield.1)
aut.df$sm.subfield.1 <- as.numeric(factor(aut.df$sm.subfield.1, levels=l))
l <- unique(aut.df$sm.subfield.2)
aut.df$sm.subfield.2 <- as.numeric(factor(aut.df$sm.subfield.2, levels=l))
l <- unique(aut.df$sm.field)
aut.df$sm.field <- as.numeric(factor(aut.df$sm.field, levels=l))
write.csv(aut.df, "self citation/self_data.csv", row.names = F)
```

5. Results

The general relationship between the parameters is the first step in considering the data. It was carried out with the help of class correlation analysis, by constructing thermal map of the correlation matrix (Fig. 12).

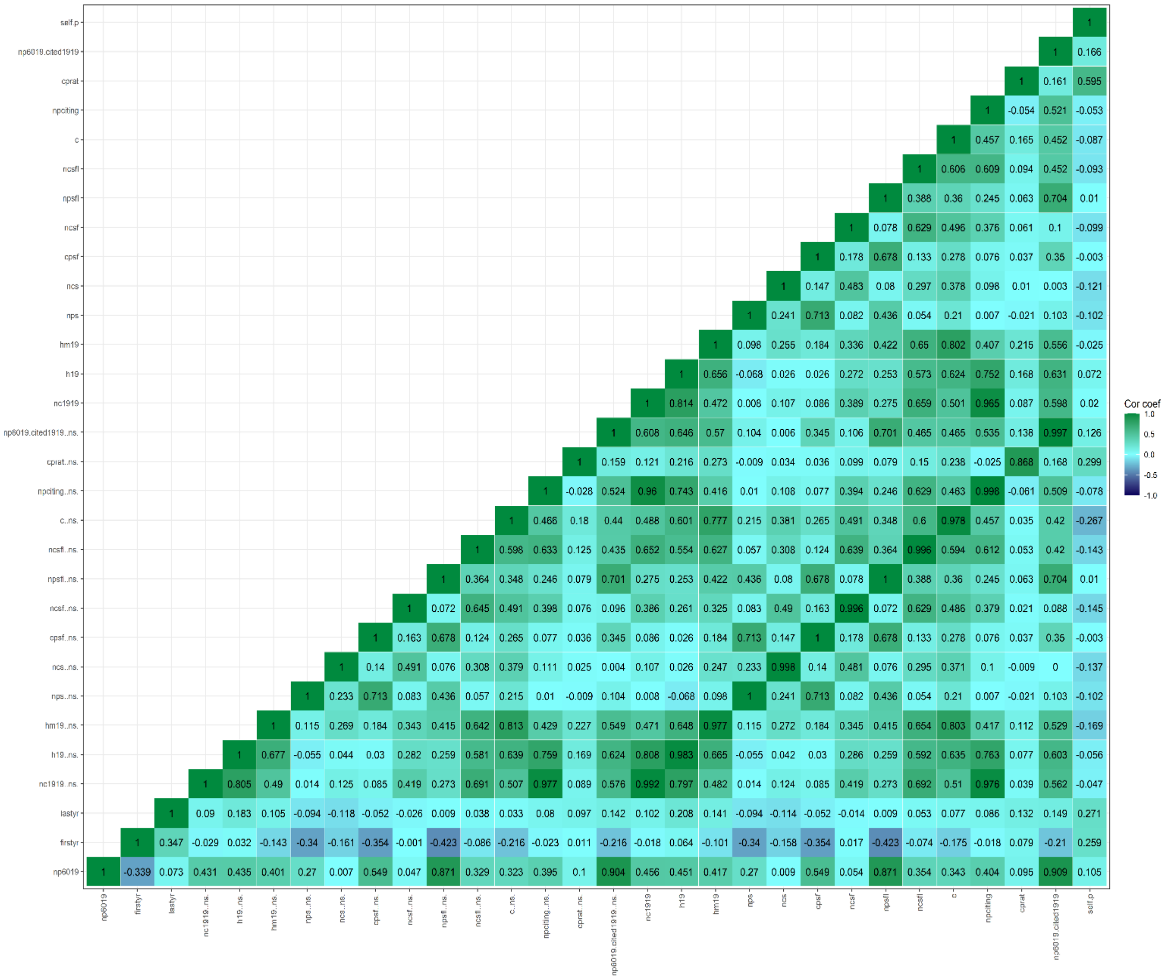


Figure 12: Thermal map of the correlation matrix

The first important aspect which should be highlighted is the second sub-diagonal presence with a correlation coefficient is closed to 1. This linear data cloud is a correlation between citation parameters (number, h-index, hm-index, etc.) with and without self-citation. The presence of such correlation means that in the general case, self-citation does not have a strong impact on the performance of authors and therefore is not used as self-plagiarism or as a source of increasing the author rating. The next important aspect is the self-citation percentage correlation to other parameters. The last column of the map from Fig. 7 is almost all painted according to zero correlation. This means that at least there is no linear relationship between self-citation and parameters presented in the dataset. Proving of the hypothesis that there is no connection between the parameters available in the data set and self-citation will be done by building neural networks and analyzing of their results.

The next stage of data review is the distribution of citation parameters and articles publication by countries. At the beginning, it is necessary to create a data window according to the algorithm described in the previous paragraph of this paper. Firstly, the statistical distributions of the fields selected from the dataset are considered, in particular, they are built as joyplot. Because the ranges and data intervals between these articles, h-index, self-citation percentage and citations are very different, all column data values have been normalized. The diagram in Fig. 13 shows a certain similarity in the distribution of these parameters, in particular, all have positive asymmetry and are higher than normal excess.

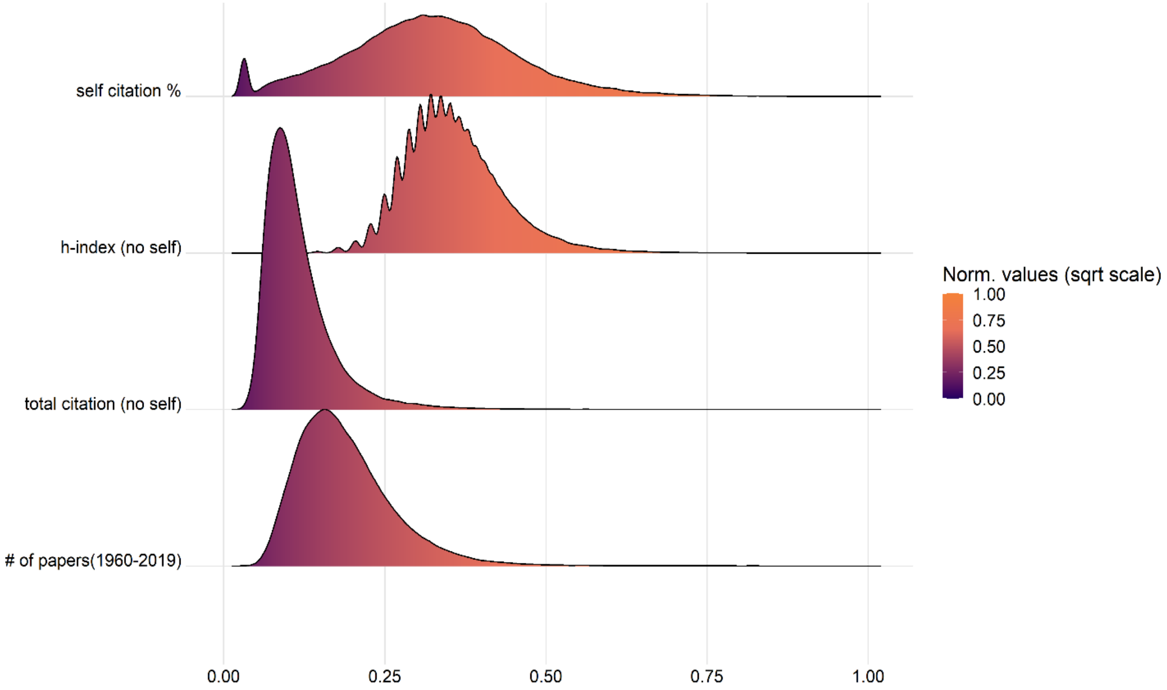


Figure 13: Joyplot chart

It is important to notice that the correlation coefficient between the average percentage of self-citations from the dataset, as well as the calculated percentage grouped according to the data is 0.95 respectively and can be considered as a correct calculation of the average percentage. The next stage is to consider individual distributions of parameters relative to the country. In Fig. 14. the top 60 (unsorted values) and 23 countries are shown in terms of the total number of articles. Part of the chart with the top 23 countries shows that the top is occupied by economically developed countries or large economies, in particular, this observation suggests the existence of correlation between economic development and the number of articles. GDP was chosen as an indicator of economic development. In contrast to GDP, an indicator of corruption in the states was added to obtain information on the dependence of the level of corruption, in particular in the educational and scientific spheres, on the abuse of self-citation among authors. The next considered parameter is the total number of the country authors citation. The graph in Fig. 15 shows the top 60, top 23 and the position of Ukraine. Each lollipop has two labels, the larger one is responsible for quoting with self-citation and without the smaller one. This graph already shows the presence of high correlation between these parameters.

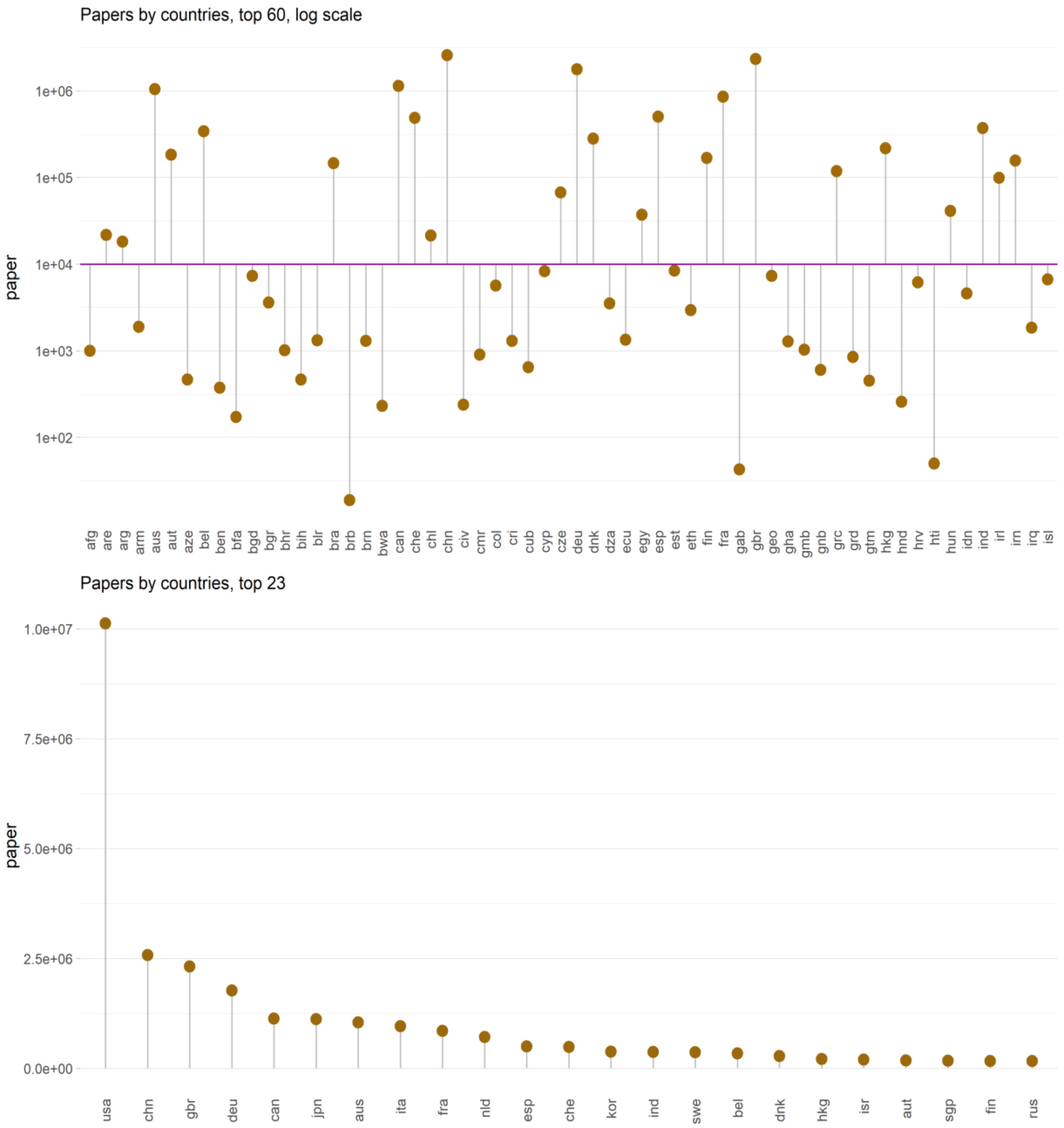


Figure 14: Lollipop diagram of the article distribution methods published for the period 1960 - 2019 by countries

Ukraine ranks 49th in the number of articles, which is equal to 10029, such a small number is due to the incompleteness of the dataset, because it includes only the most popular authors in the world. Again, the first place is behind the United States, second - China, third - Great Britain and so on. Ukraine ranks 65th along with Hungary and Peru. The distribution of the percentage of self-citations by countries is given below, the graphical representation of the data is shown in Fig. 15. Unfortunately, according to this parameter, Ukraine is in the top five, ranking fourth, with approximately equal self-citation of 34%. The smallest number of self-citations are in small countries in the Caribbean, such as Barbados and Santa Lucia. In particular, the lack of visual link between self-citation and the economic development of the state suggests that there is no correlation between GDP and the percentage of self-citation. The most interesting approach to considering of selected primers for countries is developed for the h-index. Since the h-index data is on the controversial boundary between nominal and relative data types, it is incorrect to estimate the country by the average value, so it was decided to calculate the h-index of authors h-indices in a particular country. The algorithm for constructing such a parameter is described in the previous section.

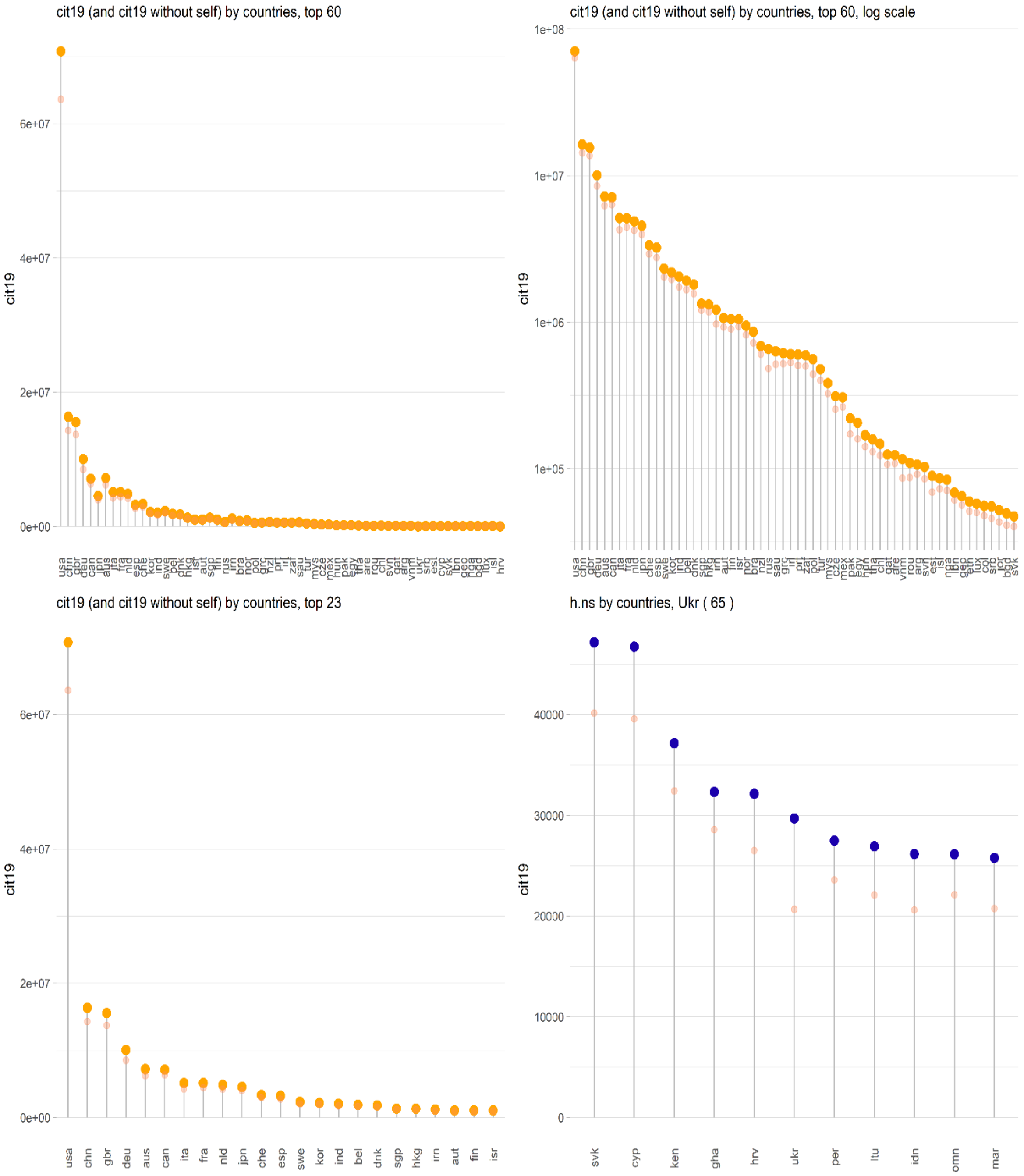


Figure 15: Lollipop chart of the authors citations number distribution for 2019 by countries

The first graph from Fig. 16 is sorted by the number of citations for 2019, and it can be seen that for the first 60 countries there is a correlation between these fields. Again, the United States ranks first, with an h-index of 55. Ukraine ranks 68th, with an index of 16, along with a number of other countries.

To display the global picture of the geographical distribution of these parameters, a graph is constructed Fig. 17.Deciphering graph values is quite intuitive, countries are sorted by number of articles, the number is indicated by the color of lolipop, the value of lolipop consists of two parts, most of them are responsible for the h-index of the country on the h-index of authors including self-citation, the size of the lolipop indicates the percentage of self-citation. Some interesting points can be seen from the graph: Russia is in the top 23 countries, but the percentage of self-citation of this country clearly stands out among others, and it falls out of the picture and the h-index, because it is quite low compared to neighboring countries.

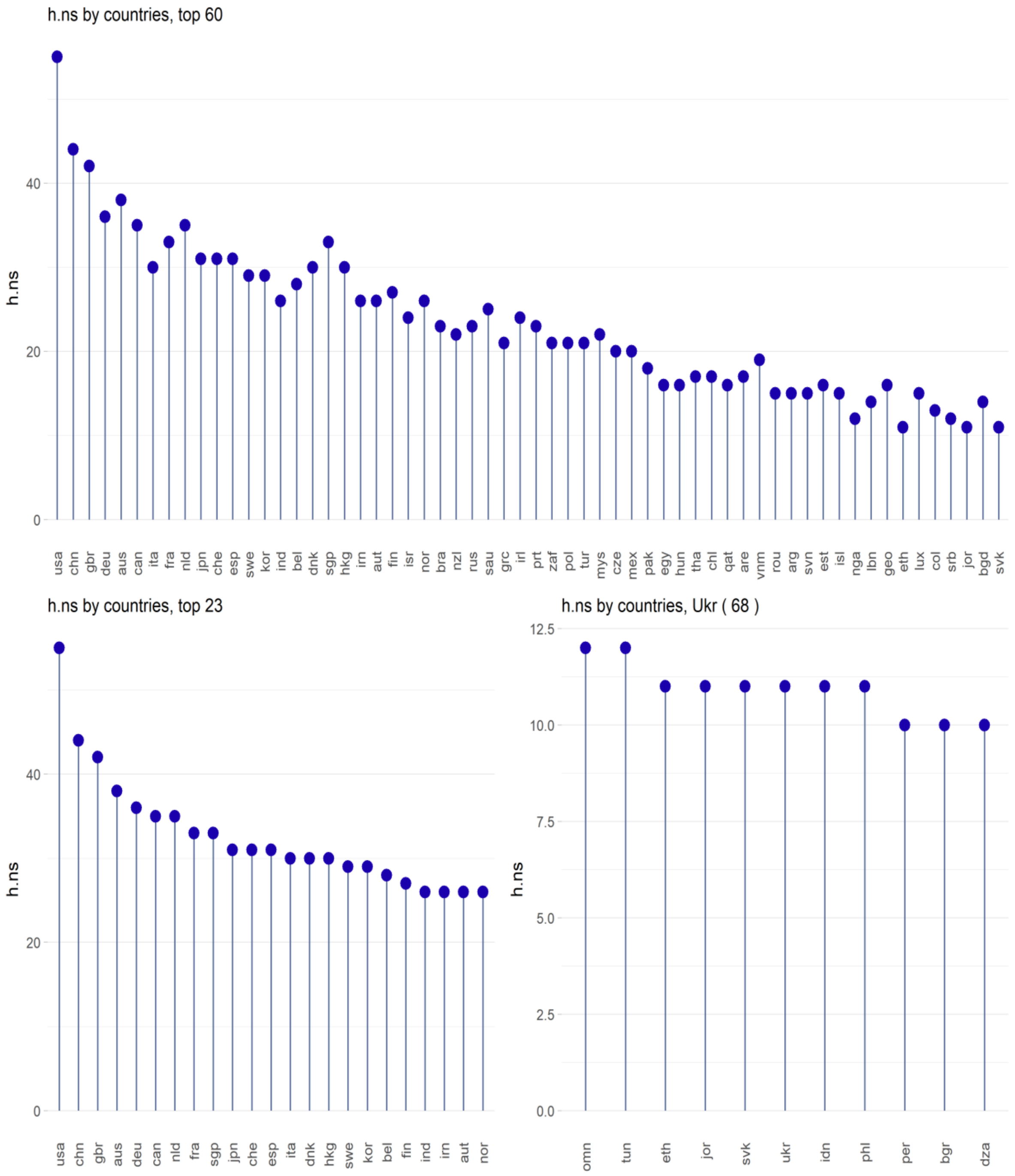


Figure 16: Top countries by h-index of authors h-indices (excluding self-citation)

After conducting graphical intelligence analysis, the values of GDP and CPI of countries were added to the data. The first step is correlation analysis. Fig. 18 shows thermal map of the correlation matrix. This map shows the presence of linear relationship between the h-indices of countries, which confirms one of the previous assumptions. The correlation matrix also confirms a fairly obvious relationship between the number of articles and citations (with and without self-citation). All correlation coefficients with self-citation are negative. In particular, the largest modulus correlation is between self-citation and the corruption rate, which allows a weak linear relationship between them, the higher the CPI (i.e. the more transparent the various spheres of government are) the lower the level of self-citation abuse is, but it is impossible to insist on the straight relationship.

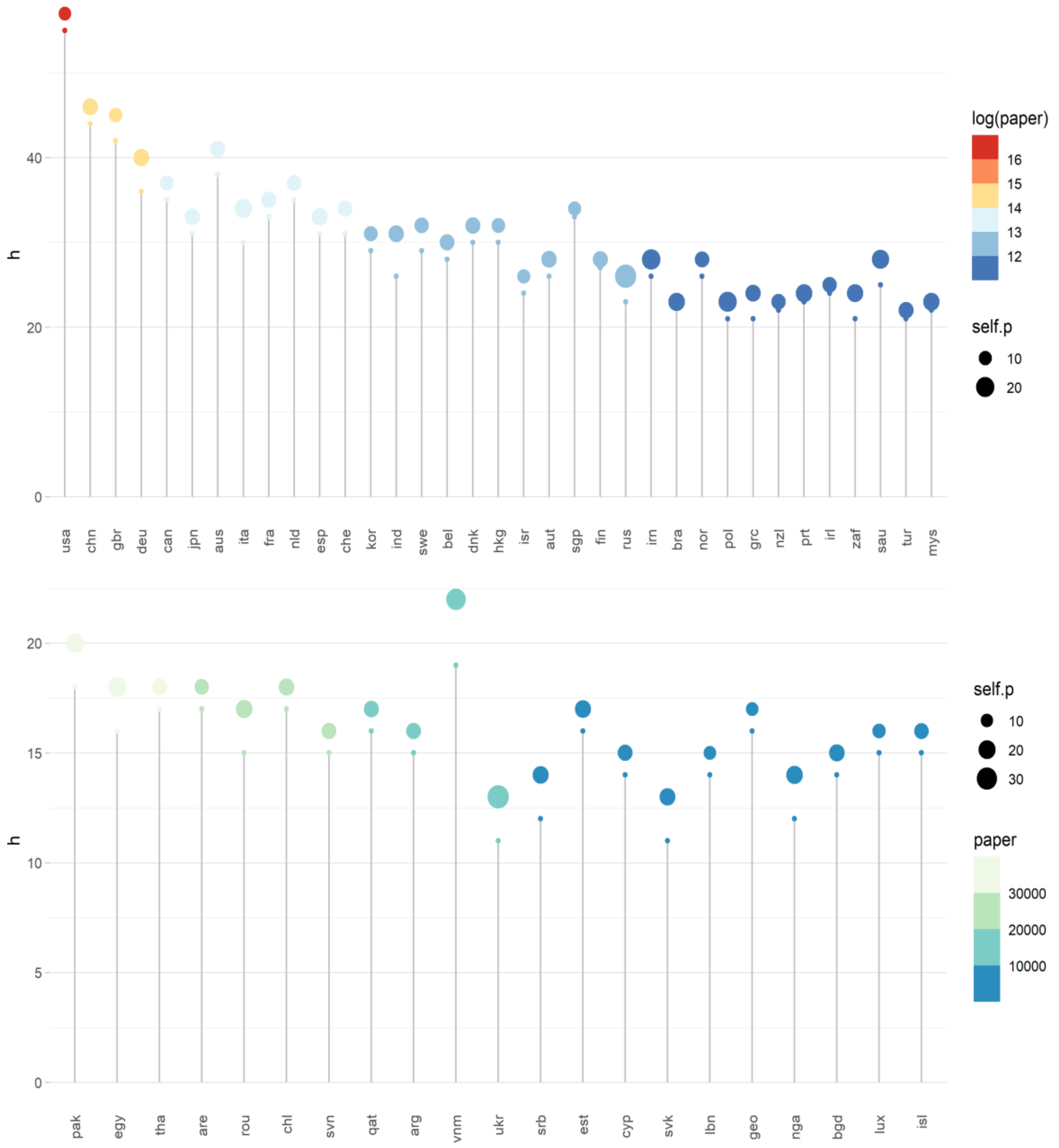


Figure 17: General representation by countries parameters

Another assumption concerns the country GDP and the number of articles. The map shows that there is a linear relationship between GDP and the number of citations. However, a little deeper analysis was done to confirm the connection. Thus, scatter plots, i.e. correlation fields, were constructed between the value of GDP and the country authors citation numbers for different countries quantities, in particular for 40, 90 and all available ones. Figure 19 shows that with a smaller sample, the correlation coefficient decreases significantly and only with the inclusion of countries such as the United States, China and the United Kingdom, the coefficient increases to the value obtained from the matrix. The fourth graph from Fig. 18 deals with the study of self-citation relationships between different parameters, and is a correlation field between the value of self-citation percentage and GDP. It shows the chaos and lack of specific dependence. Additionally, a study of extraordinary results of the weak linear relationship between the h-index and GDP and CPI. Comparison of correlation fields are shown in Fig. 20. First three graphs in Fig. 19 show a regression line constructed by using the lm-function. Similar additions to the graph are made for scatter charts from Fig. 20.

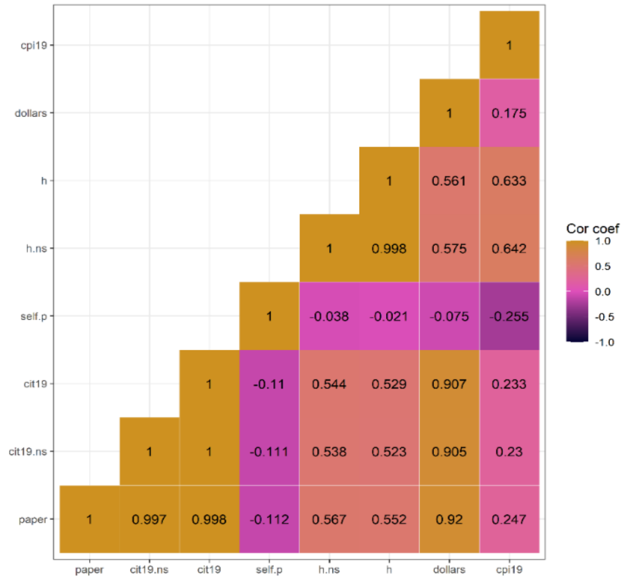


Figure 18: Thermal map of world countries parameters

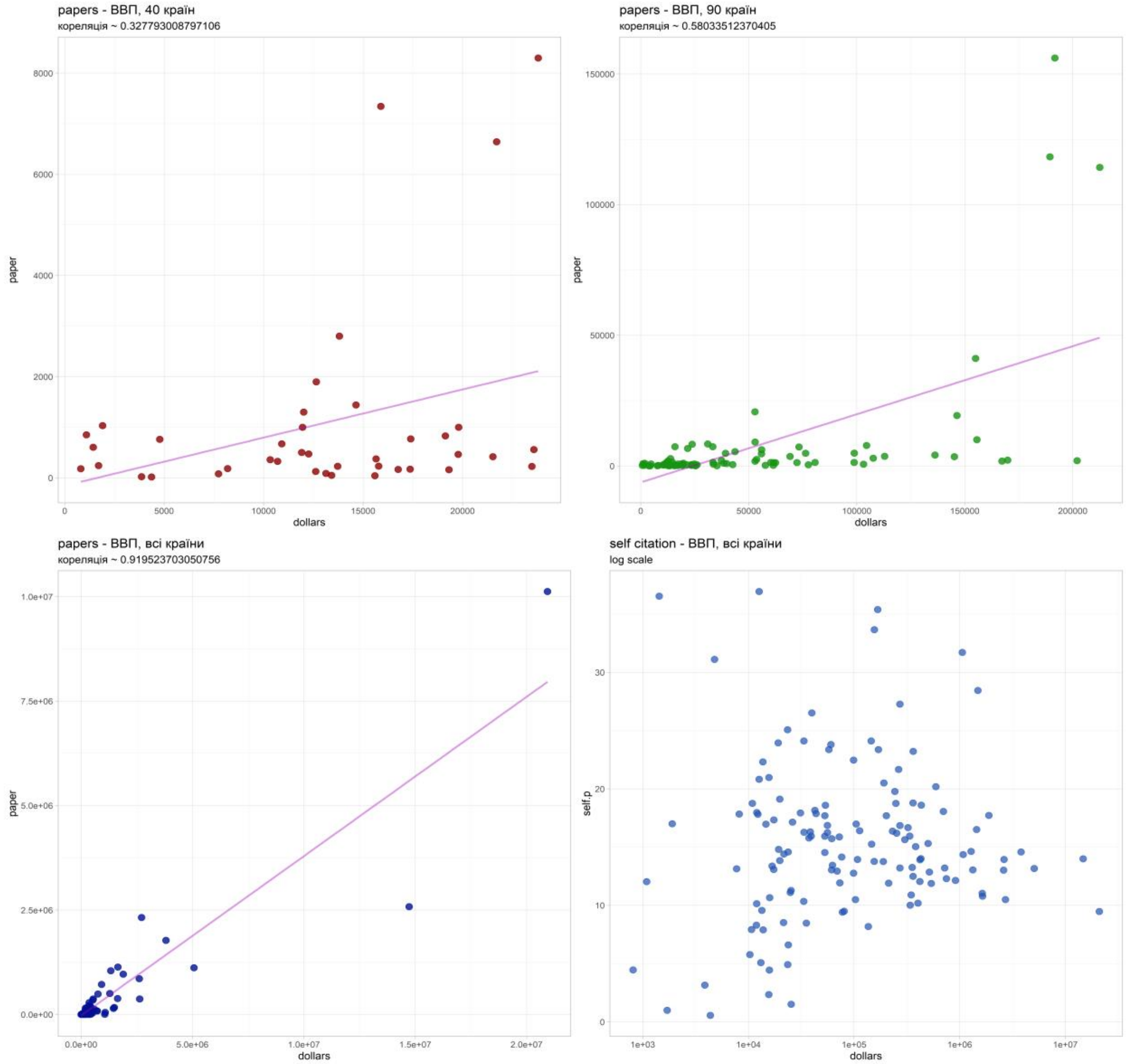


Figure 19: Set of points plots between GDP and (self) citation

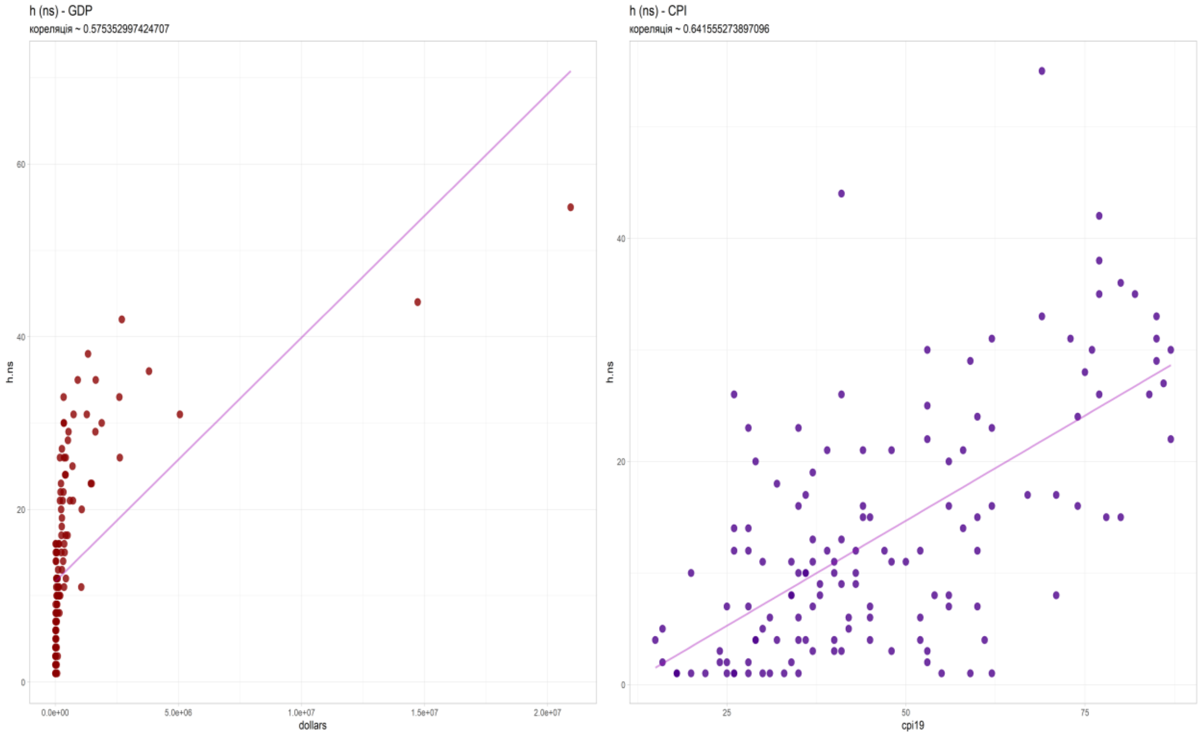


Figure 20: Comparison of correlation fields between the h-index of the country and GDP, CPI

The last mini-study based on the results of the correlation matrix is regression analysis of the relationship between the number of citations and the country h-index. Fig. 21-22 presents a bar chart. It shows that the approximation function must have the form of an exponent. The construction of this model was described in the previous section.

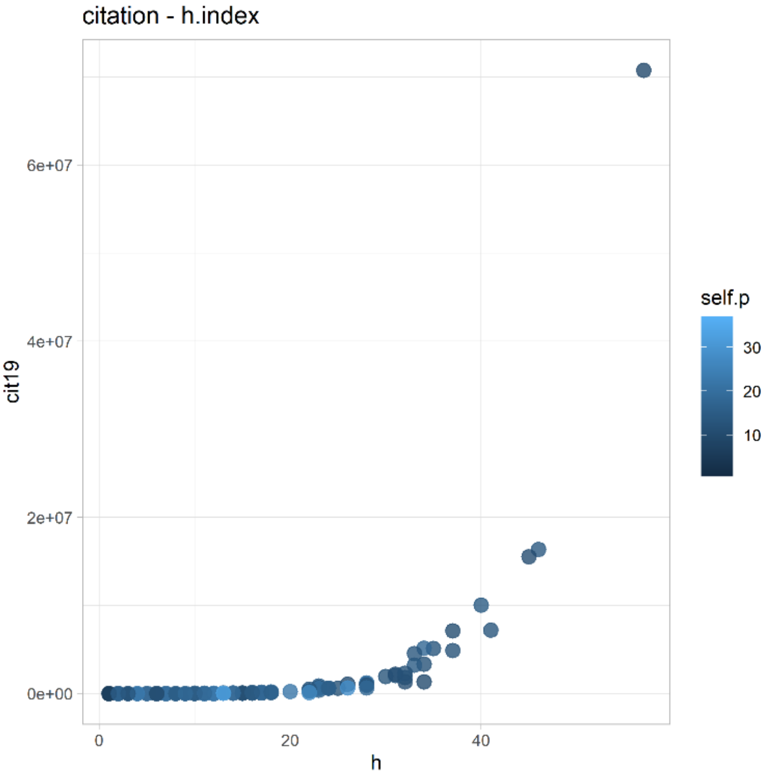


Figure 21: Correlation field

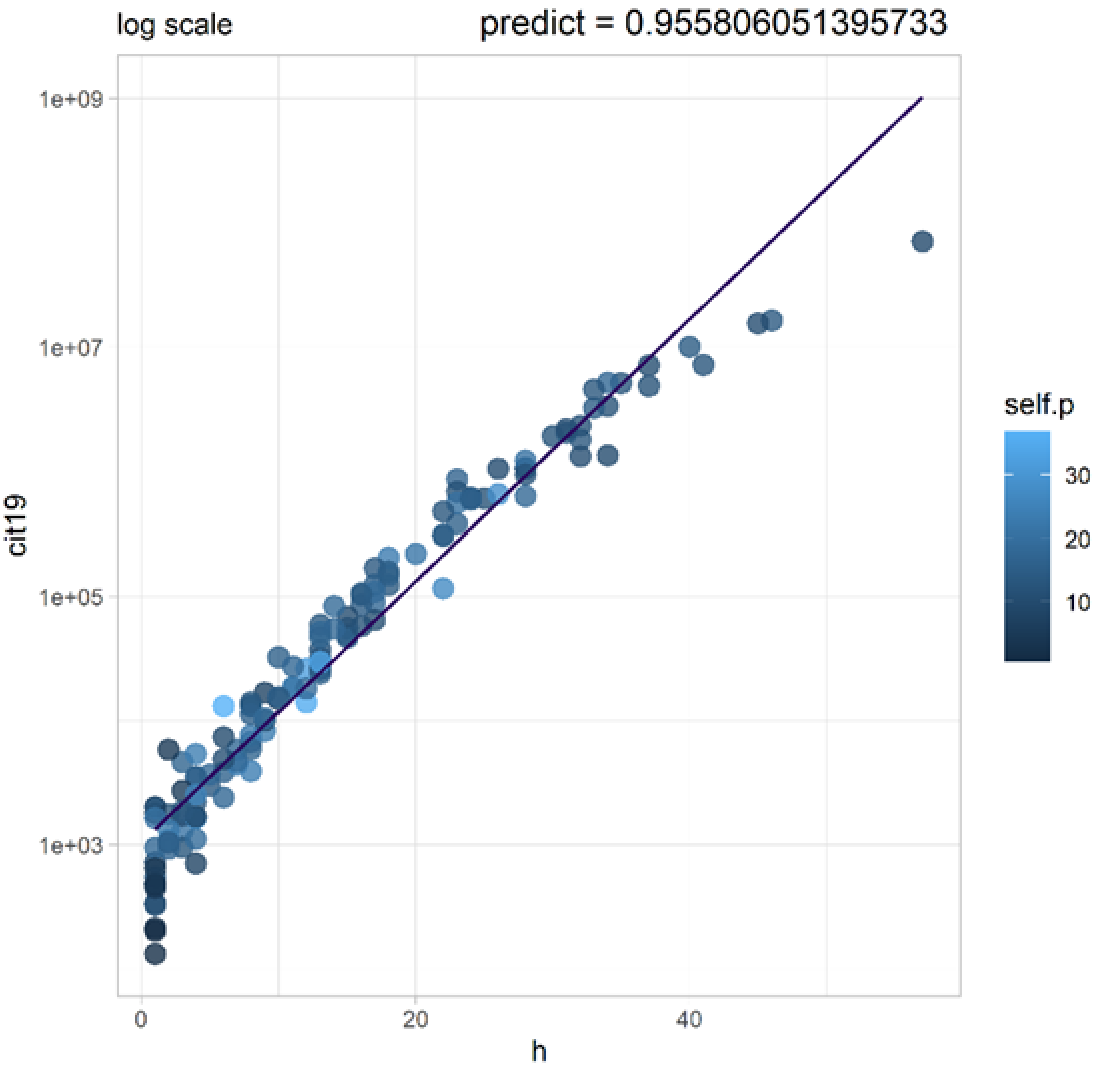


Figure 22: Correlation field with an approximation line for logarithmic transformation of the citations numbers

Before describing the results of neural networks the another observation about the distribution of self-citation has been given but this time through the prism of the field to which the author belongs (top sphere from the dataset).It is similar to the creation of data window with data about countries, performed by grouping data records by category and country. After that, three-dimensional histogram was constructed (Fig. 23), where the third dimension describes the percentage of self-citation and is determined by color. Since the global view of the histogram is not the most comfortable for visual analysis, a subset of countries for which a separate chart is built. Fig.24 shows that the highest level of self-citation is in the field of mathematics and statistics in Mexico, however, if considering Fig. 23, it is clear that the largest number of self-citations predominates in the field of physics and information technology.

The last point of the research is the construction of neural networks to study the existence of nonlinear connections between those available in the first dataset about the world authors and their self-citation. As it was described in the previous section, four different simple single- and multilayer perceptrons were created with different activation functions and randomization distribution of filling the initial values of weights and displacements. After tokenization of the fields denoting the country, institute and three different categories of the author, a new .csv file with a training and test sample has been created (Fig. 25).The resulting file contains 37 columns, respectively 36 properties, and 161,441 records. When forming the training and test sample, the ratio 80/20 was chosen.

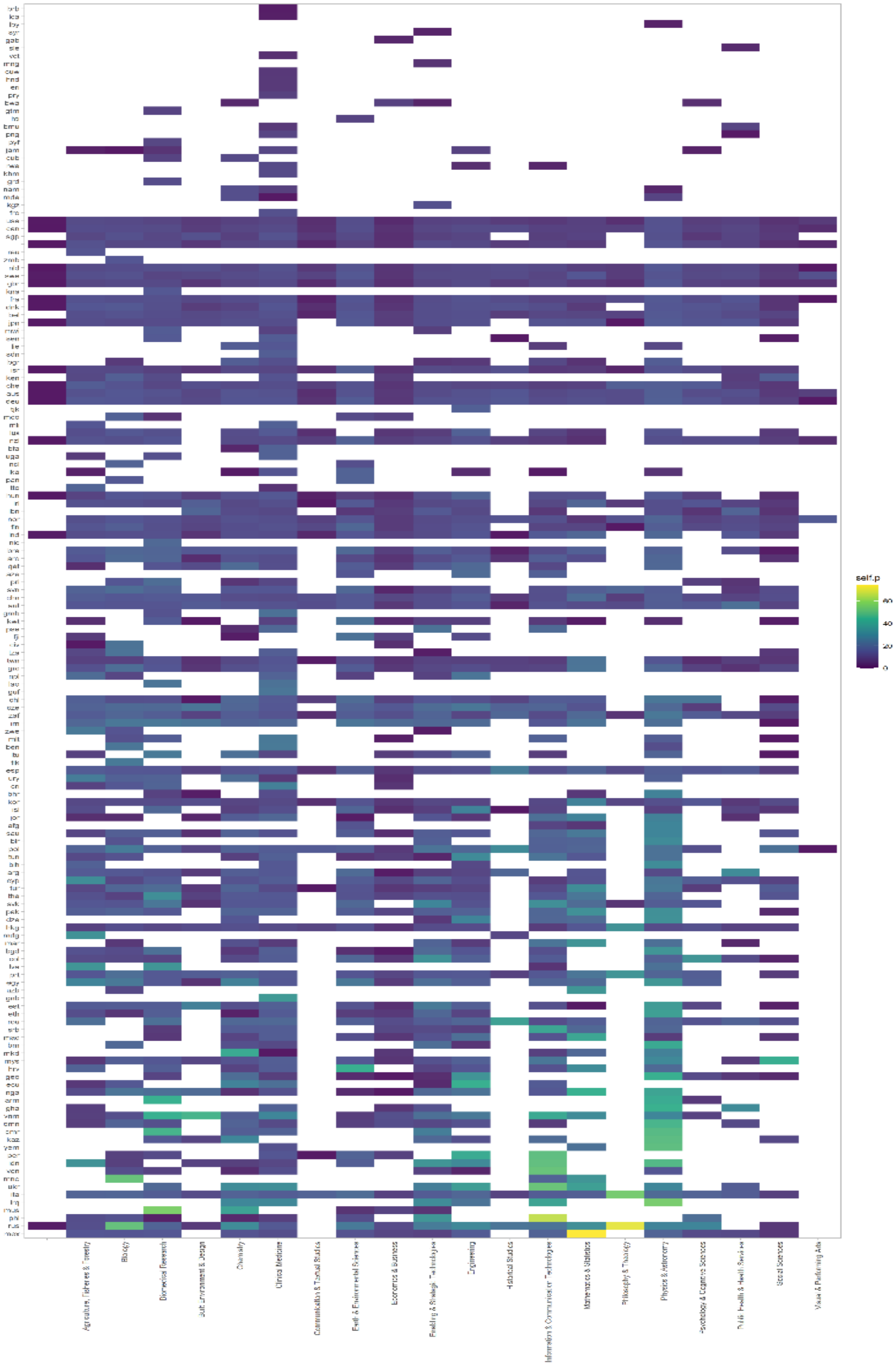


Figure 23: Global histogram

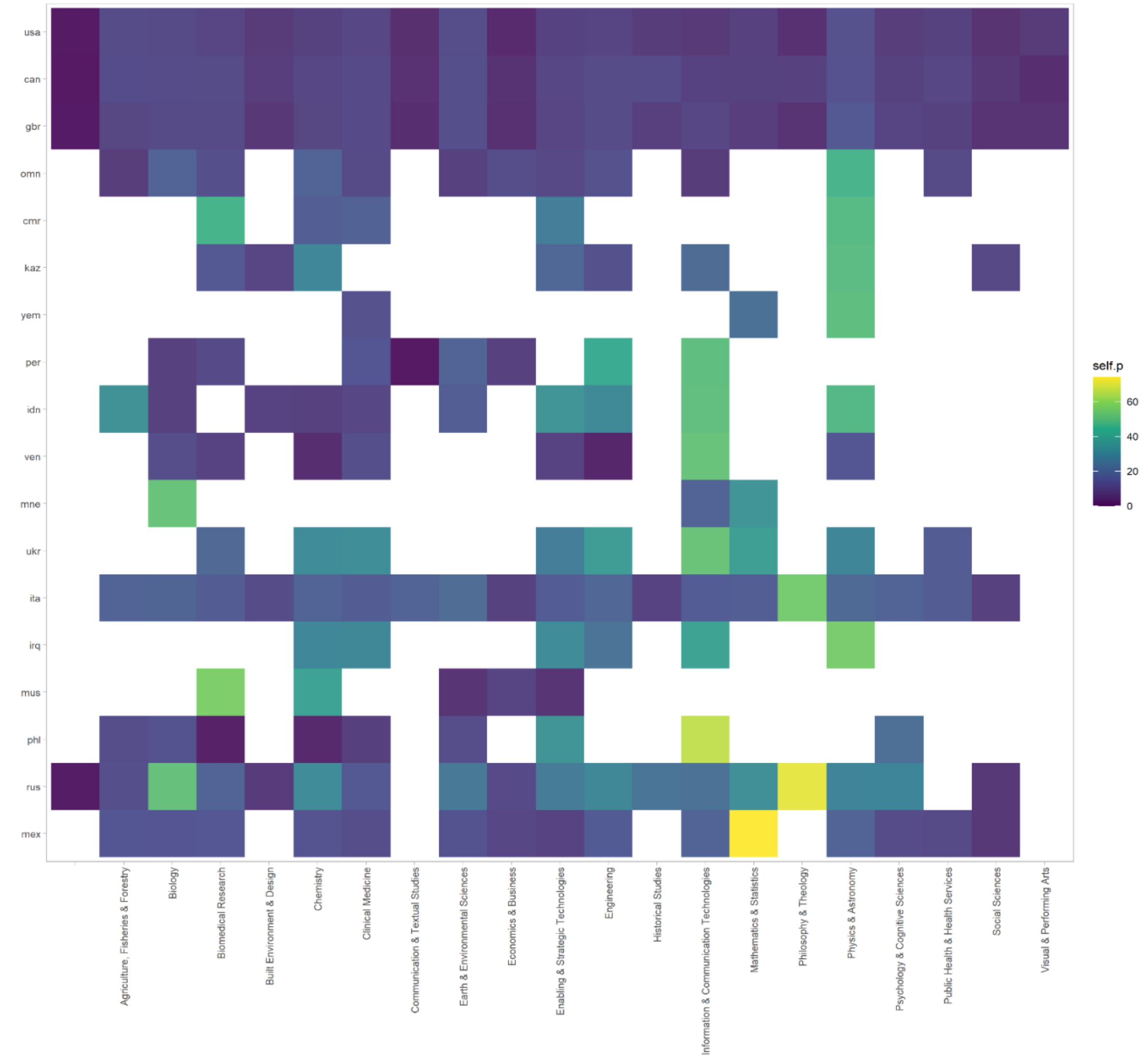
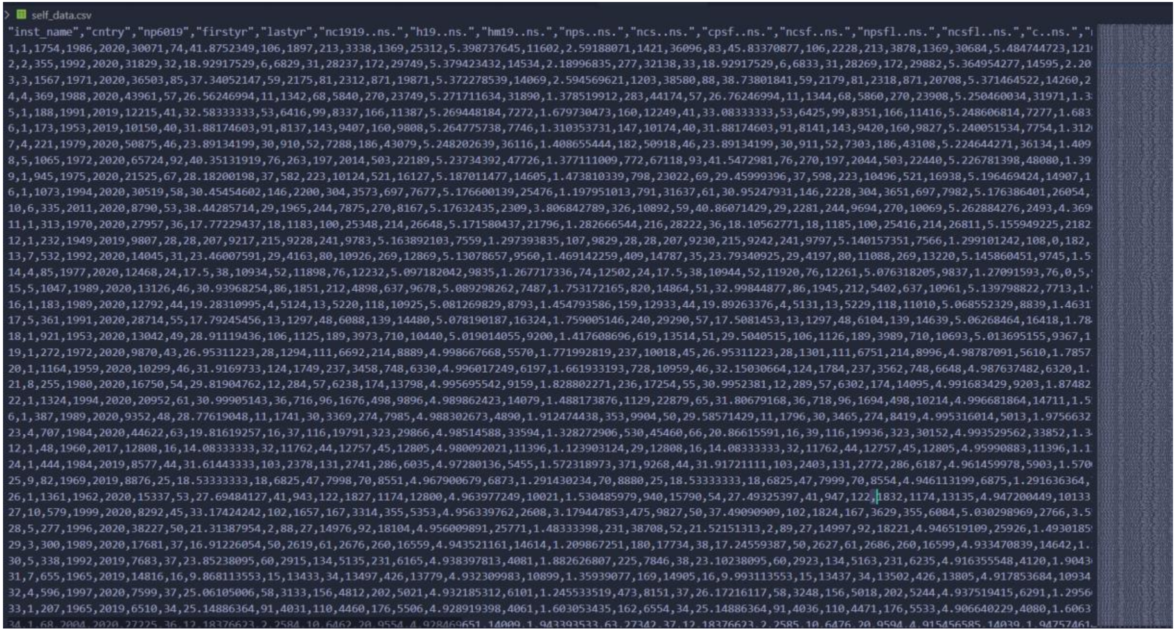


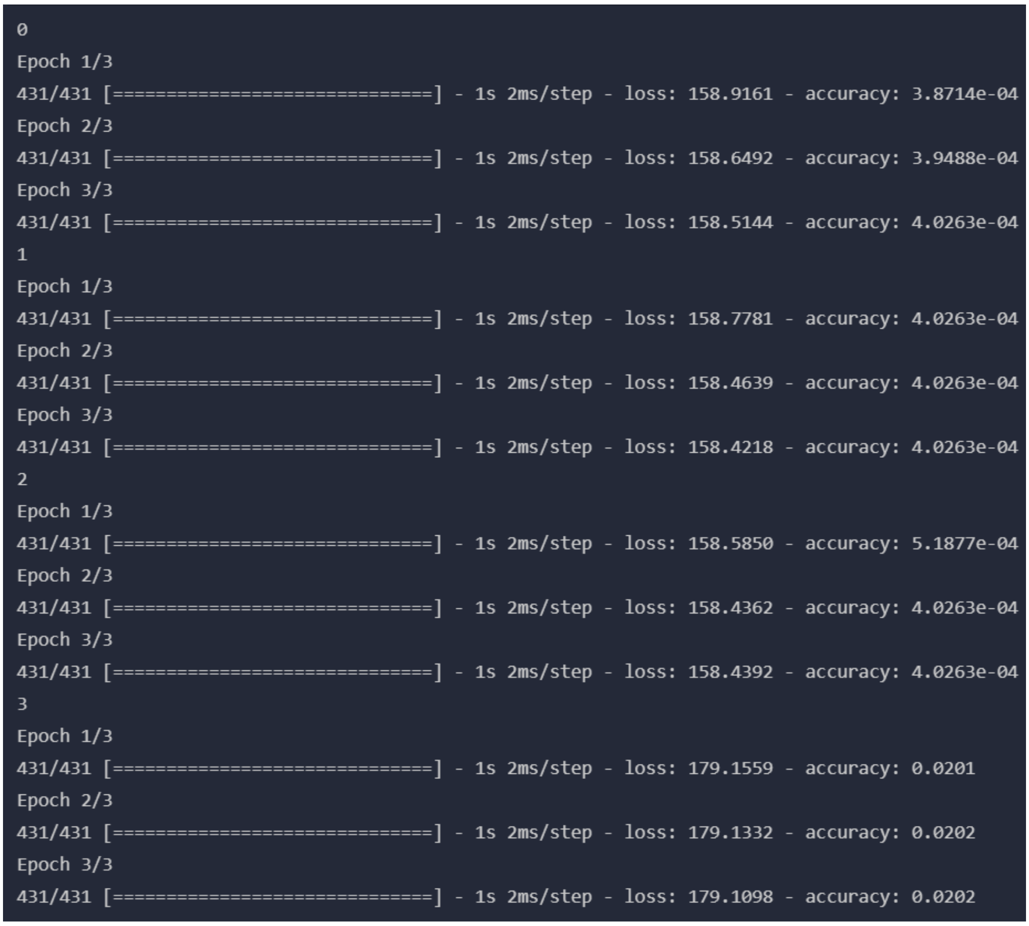
Figure 24: Partial version of the histogram



```
"inst_name","cntry","np6019","firstyr","lastyr","nc1919..ns.","h19..ns.","hm19..ns.","nps..ns.","ncs..ns.","cpsf..ns.","ncsf..ns.","npsfl..ns.","ncsfl..ns.","c..ns.","
1,1,1754,1986,2020,30071,74,41.8752349,106,1897,213,3338,1369,25312,5.398737645,11602,2.59188071,1421,36096,83,45.83370877,106,2228,213,3878,1369,30684,5.484744723,121
2,2,355,1992,2020,31829,32,18.92917529,6,6829,31,28237,172,29749,5.379423432,14534,2.18996835,277,32138,33,18.92917529,6,6833,31,28269,172,29882,5.364954277,14595,2.20
3,3,1567,1971,2020,36503,85,37.34052147,59,2175,81,2312,871,19871,5.372278539,14069,2.594569621,1203,38580,88,38.73801841,59,2179,81,2318,871,20708,5.371464522,14260,2
4,4,369,1988,2020,43961,57,26.56246994,11,1342,68,5840,270,23749,5.271711634,31890,1.378519912,283,44174,57,26.76246994,11,1344,68,5860,270,23908,5.250460034,31971,1.3
5,1,188,1991,2019,12215,41,32.58333333,53,6416,99,8337,166,11387,5.269448184,7272,1.679730473,160,12249,41,33.08333333,53,6425,99,8351,166,11416,5.248606814,7277,1.683
6,1,173,1953,2019,10150,40,31.88174603,91,8137,143,9407,160,9808,5.264775738,7746,1.310353731,147,10174,40,31.88174603,91,8141,143,9420,160,9827,5.240051534,7754,1.312
7,4,221,1979,2020,50875,46,23.89134199,30,910,52,7288,186,43079,5.248202639,36116,1.408655444,182,50918,46,23.89134199,30,911,52,7303,186,43108,5.224644271,36134,1.409
8,5,1065,1972,2020,65724,92,40.35131919,76,263,197,2014,503,22189,5.23734392,47726,1.377111009,772,67118,93,41.5472981,76,270,197,2044,503,22440,5.226781398,48080,1.39
9,1,945,1975,2020,21525,67,28.18200198,37,582,223,10124,521,16127,5.187011477,14605,1.473810339,798,23022,69,29.45999396,37,598,223,10496,521,16938,5.196469424,14907,1
6,1,1073,1994,2020,30519,58,30.45454602,146,2200,304,3573,697,7677,5.176600139,25476,1.197951013,791,31637,61,30.95247931,146,2228,304,3651,697,7982,5.176386401,26054,
10,6,335,2011,2020,8790,53,38.44285714,29,1965,244,7875,270,8167,5.17632435,2309,3.806842789,326,10892,59,40.86071429,29,2281,244,9694,270,10069,5.262884276,2493,4.369
11,1,313,1970,2020,27957,36,17.77229437,18,1183,100,25348,214,26648,5.171580437,21796,1.282666544,216,28222,36,18.10562771,18,1185,100,25416,214,26811,5.155949225,2182
12,1,232,1949,2019,9807,28,28,207,9217,215,9228,241,9783,5.163892103,7559,1.297393835,107,9829,28,28,207,9230,215,9242,241,9797,5.140157351,7566,1.299101242,108,0,182,
13,7,532,1992,2020,14045,31,23.46007591,29,4163,80,10926,269,12869,5.13078657,9560,1.469142259,409,14787,35,23.79340925,29,4197,80,11088,269,13220,5.145860451,9745,1.5
14,4,85,1977,2020,12468,24,17.5,38,10934,52,11898,76,12232,5.097182042,9835,1.267717336,74,12502,24,17.5,38,10944,52,11920,76,12261,5.076318205,9837,1.27091593,76,0,5,
15,5,1047,1989,2020,13126,46,30.93968254,86,1851,212,4898,637,9678,5.089298262,7487,1.753172165,820,14864,51,32.99844877,86,1945,212,5402,637,10961,5.139798822,7713,1.
16,1,183,1989,2020,12792,44,19.28310995,4,5124,13,5220,118,10925,5.081269829,8793,1.454793586,159,12933,44,19.89263376,4,5131,13,5229,118,11010,5.068552329,8839,1.4631
17,5,361,1991,2020,28714,55,17.79245456,13,1297,48,6088,139,14480,5.078190187,16324,1.759005146,240,29290,57,17.5081453,13,1297,48,6104,139,14639,5.06268464,16418,1.78
18,1,921,1953,2020,13042,49,28.91119436,106,1125,189,3973,710,10440,5.019014055,9200,1.417608696,619,13514,51,29.5040515,106,1126,189,3989,710,10693,5.013695155,9367,1
19,1,272,1972,2020,9870,43,26.95311223,28,1294,111,6692,214,8889,4.998667668,5570,1.771992819,237,10018,45,26.95311223,28,1301,111,6751,214,8996,4.98787091,5610,1.7857
20,1,1164,1959,2020,10299,46,31.9169733,124,1749,237,3458,748,6330,4.996017249,6197,1.661933193,728,10959,46,32.15030664,124,1784,237,3562,748,6648,4.987637482,6320,1.
21,8,255,1980,2020,16750,54,29.81904762,12,284,57,6238,174,13798,4.995695542,9159,1.828802271,236,17254,55,30.9952381,12,289,57,6302,174,14095,4.991683429,9203,1.87482
22,1,1324,1994,2020,20952,61,30.99905143,36,716,96,1676,498,9896,4.989862423,14079,1.488173876,1129,22879,65,31.80679168,36,718,96,1694,498,10214,4.996681864,14711,1.5
6,1,387,1989,2020,9352,48,28.77619048,11,1741,30,3369,274,7985,4.988302673,4890,1.912474438,353,9904,50,29.58571429,11,1796,30,3465,274,8419,4.995316014,5013,1.9756632
23,4,707,1984,2020,44622,63,19.81619257,16,37,116,19791,323,29866,4.98514588,33594,1.328272906,530,45460,66,20.86615591,16,39,116,19936,323,30152,4.993529562,33852,1.3
12,1,48,1960,2017,12808,16,14.08333333,32,11762,44,12757,45,12805,4.980092021,11396,1.123903124,29,12808,16,14.08333333,32,11762,44,12757,45,12805,4.95990883,11396,1.1
24,1,444,1984,2019,8577,44,31.61443333,103,2378,131,2741,286,6035,4.97280136,5455,1.572318973,371,9268,44,31.91721111,103,2403,131,2772,286,6187,4.961459978,5903,1.570
25,9,82,1969,2019,8876,25,18.53333333,18,6825,47,7998,70,8551,4.967900679,6873,1.291430234,70,8880,25,18.53333333,18,6825,47,7999,70,8554,4.946113199,6875,1.291636364,
26,1,1361,1962,2020,15337,53,27.69484127,41,943,122,1827,1174,12800,4.963977249,10021,1.530485979,940,15790,54,27.49325397,41,947,122,1832,1174,13135,4.947200449,10133
27,10,579,1999,2020,8292,45,33.17424242,102,1657,167,3314,355,5353,4.956339762,2608,3.179447853,475,9827,50,37.49090909,102,1824,167,3629,355,6084,5.030298969,2766,3.5
28,5,277,1996,2020,38227,50,21.31387954,2,88,27,14976,92,18104,4.956009891,25771,1.48333398,231,38708,52,21.52151313,2,89,27,14997,92,18221,4.946519109,25926,1.493018
29,3,300,1989,2020,17681,37,16.91226054,50,2619,61,2676,260,16559,4.943521161,14614,1.209867251,180,17734,38,17.24559387,50,2627,61,2686,260,16599,4.933470839,14642,1.
30,5,338,1992,2019,7683,37,23.85238095,60,2915,134,5135,231,6165,4.938397813,4081,1.882626807,225,7846,38,23.10238095,60,2923,134,5163,231,6235,4.916355548,4120,1.9043
31,7,655,1965,2019,14816,16,9.868113553,15,13433,34,13497,426,13779,4.932309983,10899,1.35939077,169,14905,16,9.993113553,15,13437,34,13502,426,13805,4.917853684,10934
32,4,596,1997,2020,7599,37,25.06105006,58,3133,156,4812,202,5021,4.932185312,6101,1.245533519,473,8151,37,26.17216117,58,3248,156,5018,202,5244,4.937519415,6291,1.2956
33,1,207,1965,2019,6510,34,25.14886364,91,4031,110,4460,176,5506,4.928919398,4061,1.603053435,162,6554,34,25.14886364,91,4036,110,4471,176,5533,4.906640229,4080,1.6063
34,1,68,2004,2020,27225,36,12.18376623,2,2584,10,6462,20,9554,4.928469651,14009,1.943393533,63,27342,37,12.18376623,2,2585,10,6476,20,9594,4.915456585,14039,1.94757461
```

Figure 25: Fragment of input data

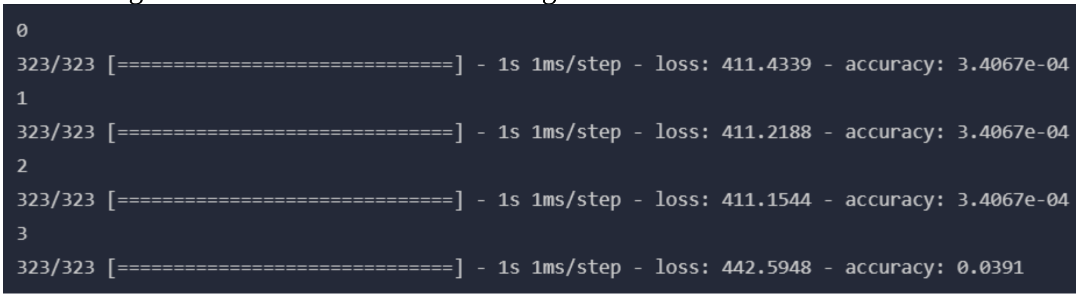
During training, the number of repetitions according to the data is equal to 3, and only accuracy is selected as a metric. After starting the training the following results have been obtained:



```
0
Epoch 1/3
431/431 [==============================] - 1s 2ms/step - loss: 158.9161 - accuracy: 3.8714e-04
Epoch 2/3
431/431 [==============================] - 1s 2ms/step - loss: 158.6492 - accuracy: 3.9488e-04
Epoch 3/3
431/431 [==============================] - 1s 2ms/step - loss: 158.5144 - accuracy: 4.0263e-04
1
Epoch 1/3
431/431 [==============================] - 1s 2ms/step - loss: 158.7781 - accuracy: 4.0263e-04
Epoch 2/3
431/431 [==============================] - 1s 2ms/step - loss: 158.4639 - accuracy: 4.0263e-04
Epoch 3/3
431/431 [==============================] - 1s 2ms/step - loss: 158.4218 - accuracy: 4.0263e-04
2
Epoch 1/3
431/431 [==============================] - 1s 2ms/step - loss: 158.5850 - accuracy: 5.1877e-04
Epoch 2/3
431/431 [==============================] - 1s 2ms/step - loss: 158.4362 - accuracy: 4.0263e-04
Epoch 3/3
431/431 [==============================] - 1s 2ms/step - loss: 158.4392 - accuracy: 4.0263e-04
3
Epoch 1/3
431/431 [==============================] - 1s 2ms/step - loss: 179.1559 - accuracy: 0.0201
Epoch 2/3
431/431 [==============================] - 1s 2ms/step - loss: 179.1332 - accuracy: 0.0202
Epoch 3/3
431/431 [==============================] - 1s 2ms/step - loss: 179.1098 - accuracy: 0.0202
```

Figure 26: History of perceptron training

The following results were obtained when testing the networks:



```
0
323/323 [==============================] - 1s 1ms/step - loss: 411.4339 - accuracy: 3.4067e-04
1
323/323 [==============================] - 1s 1ms/step - loss: 411.2188 - accuracy: 3.4067e-04
2
323/323 [==============================] - 1s 1ms/step - loss: 411.1544 - accuracy: 3.4067e-04
3
323/323 [==============================] - 1s 1ms/step - loss: 442.5948 - accuracy: 0.0391
```

Figure 27: Testing results

Figures 26-27 show that no network has any level of accuracy in determining the level of the author self-citation on the available parameters, so it can be argued that these properties have no significant relationship with the level of self-citation.

6. Discussions

The comprehensive research, which was aimed to test a number of assumptions about the relationship between the numerical characteristics of the world authors activity on the level of their self-citation has been made. The study did not reveal any patterns, relationships between the investigated properties. Such results lead to certain assumptions about the individualization of characteristics, motives, level of intellectual and moral development of authors, as well as general trends in science, existing problems of society, etc. to the level of self-citation. Psychological analysis may show a link between self-citation abuse and its use as it is a need.

Further research requires to expand the subject area by finding other possible sources of data.

On the other hand, the other part of the study that is the geo-economic distribution of certain parameters has more positive results. The analysis has confirmed a number of hypotheses, including the connection between the economic development of the country and its ability to create scientific papers, articles with sufficient quality as there is a link between authors citation and the country GDP.

The following analysis can be carried out with in-depth division of the components of GDP, possibly with the expansion of geographical areas, which authors are grouped by.

There may be a number of other indicators that have an impact on the characteristic of the state in terms of scientific articles authors views. The paper considers the level of corruption in countries as an example of possible indices / parameters. In particular, data on the level of education, research costs and the availability of private research and art centers are suitable for analysis.

So, the analysis of the geo - economic distribution of parameters in a specific sample of countries - Ukraine, Poland, Georgia, Slovakia and Romania has been made.

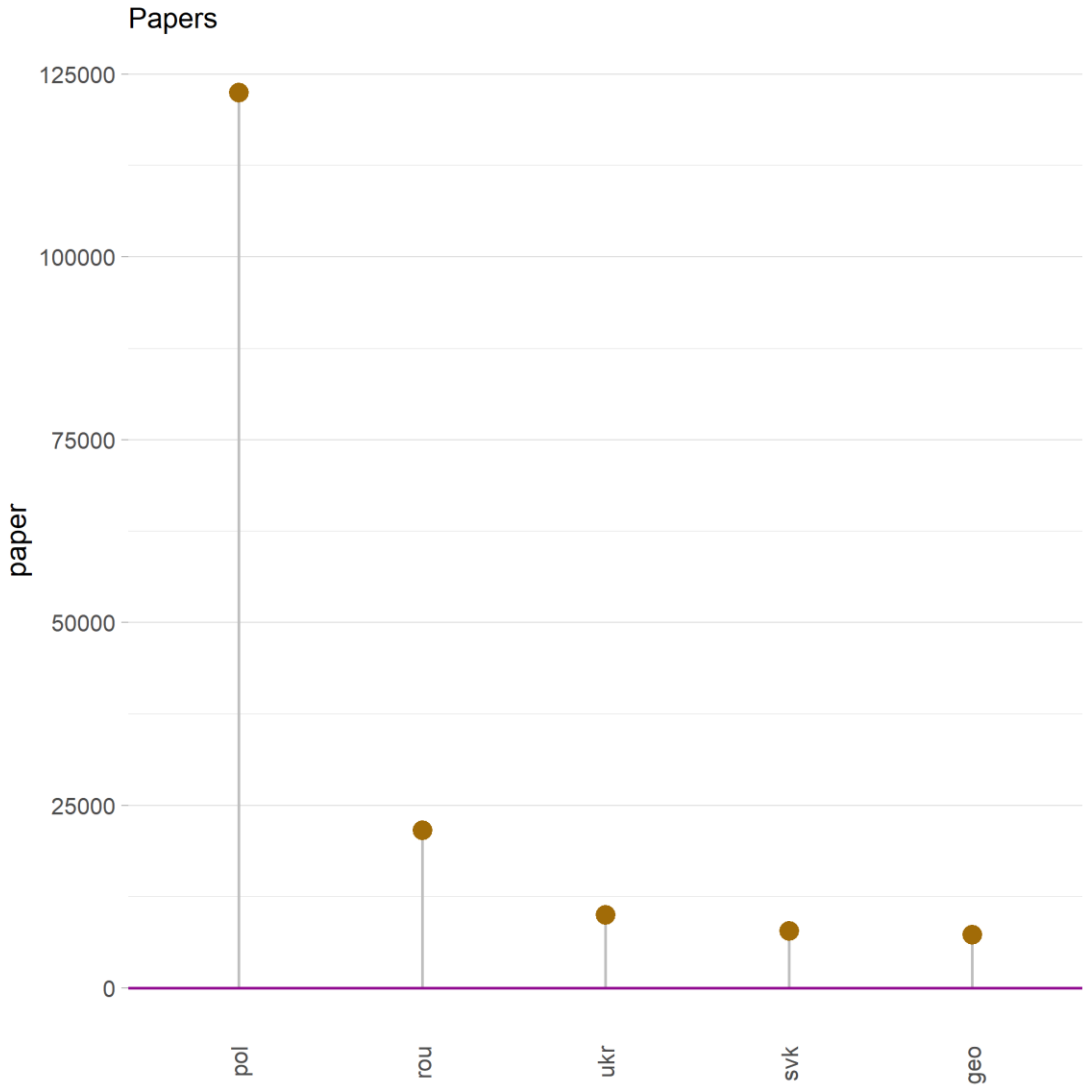


Figure28: Lolliplot distribution of articles

The graph in Fig. 28 shows that the largest number of articles was published in Poland, with Ukraine in third place. The following graphs from Fig. 29 show that Poland ranks first in all rankings, except for self-citation, where the championship, unfortunately, belongs to Ukraine.

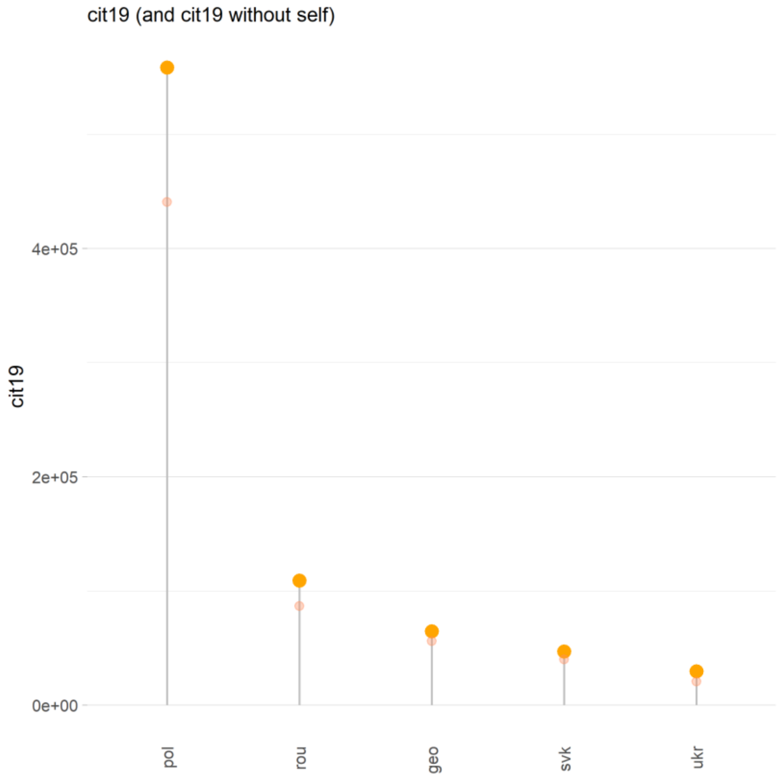


Figure 28: Distribution of author citations numbers for 2019

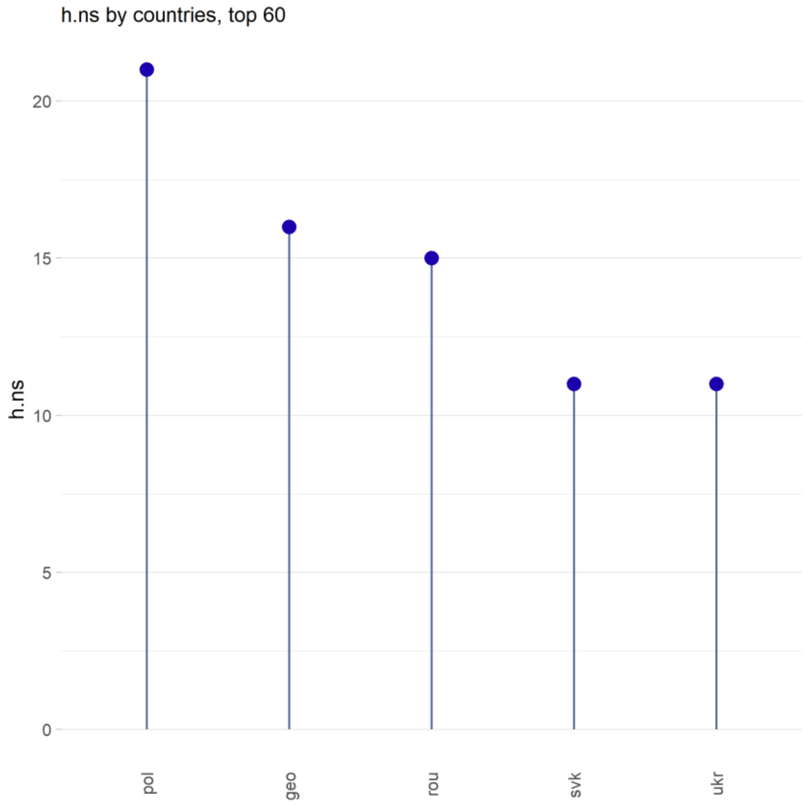


Figure 29: Rating of the countries h-index

From the last graph in Fig. 30 it can be concluded that the existing dataset has ten Ukrainian authors with a minimum h-index equal to 10. The generalized picture is shown in Fig. 31. Additionally, graph for the distribution of these countries self-citations percentage by categories of authors has been created. The corresponding histogram is shown in Fig. 32.

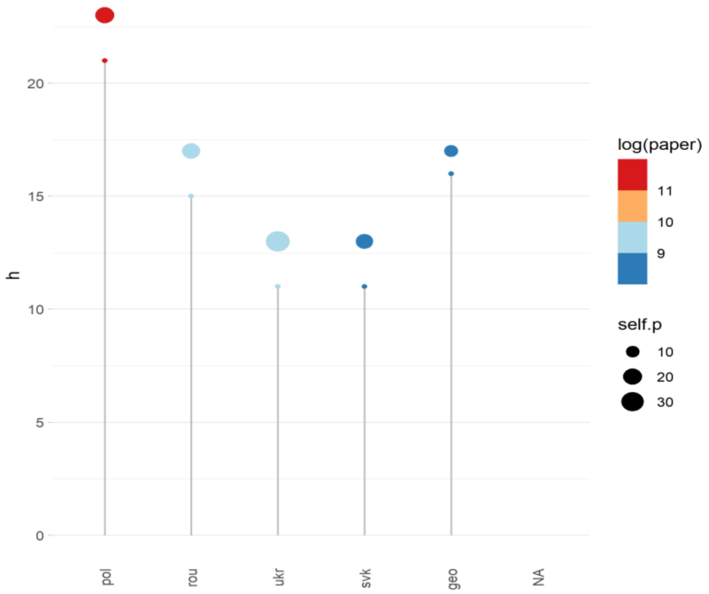


Figure 30: General graph for country characteristics

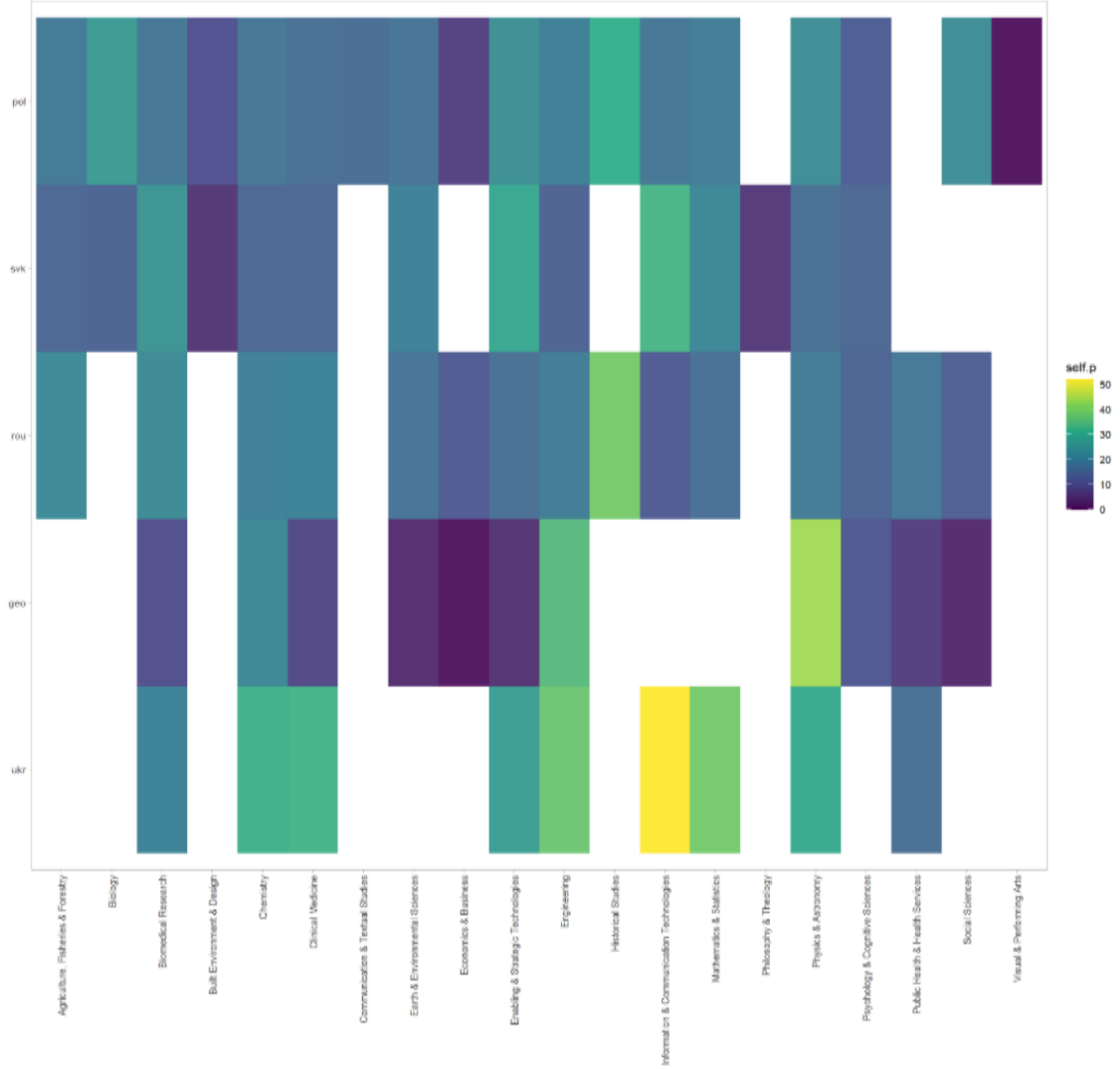


Figure 31: Three-dimensional histogram of self-citation distribution by countries and categories

7. Conclusions

In the result of the research, the study on the presence, analytical view or lack of links between the geo-economic distribution of human resources that produce certain scientific papers and articles has been conducted. The paper presents and uses possible approaches to estimate the country, for example, with a double overlap of the h-index. In a comprehensive study of the parameters and generalized characteristics of the world the authors give examples of using classical and modern mathematical information apparatus to find and prove the presence of patterns in data including the use of graphical data representation and small neural networks combination.

8. References

[1] A. Kacem, J. W. Flatt, P. Mayr, Tracking self-citations in academic publishing, Scientometrics 123 (2020) 1157–1165. doi: 10.1007/s11192-020-03413-9.

[2] Corruption Perceptions Index. 2019. URL: https://www.transparency.org/en/cpi/2019/index/nzl

[3] D. W. Aksnes, A macro study of self-citation, Scientometrics 56 (2003) 235–246. doi: 10.1023/A:1021919228368.

[4] G. Kaptay, The k-index is introduced to replace the h-index to evaluate better the scientific excellence of individuals, Heliyon 6(7) (2020) e04415. doi: 10.1016/j.heliyon.2020.e04415.

[5] V. A. Bloomfield, Using R for Numerical Analysis in Science and Engineering. Minneapolis, USA: University of Minnesota, 2014. URL: http://hsrm-mathematik.de/SS2020/semester4/Datenanalyse-und-ScientificComputing-mit-R/book.pdf.

[6] World Bank GD Pranking, 2019. URL: https://www.kaggle.com/theworldbank/world-bank-gdp-ranking.

[7] J. Ioannidis, J. Baas, R. Klavans, K. Boyack, Supplementary data tables for "A standardized citation metrics author database annotated for scientific field" (PLoS Biology 2019), 2019. doi: 10.17632/btchxktzyw.1. URL: https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/1.

[8] J. Baas, K. Boyack, J. P. A. Ioannidis, Data for "Updated science-wide author databases of standardized citation indicators, 2020. doi: 10.17632/btchxktzyw.2. URL: https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/2.

[9] J. Baas, K. Boyack, J. P. A. Ioannidis, August 2021 data-update for "Updated science-wide author databases of standardized citation indicators", 2021. doi: 10.17632/btchxktzyw.3. URL: https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/3.

[10] B. Rusyn, O. Lutsyk, R. Kosarevych, Y. Obukh, Application Peculiarities of Deep Learning Methods in the Problem of Big Datasets Classification, Lecture Notes in Electrical Engineering 831 (2022) 493–506. doi: 10.1007/978-3-030-92435-5_28.

[11] M. Emmerich, V. Lytvyn, V. Vysotska, V. B. Fernandes, V. Lytvynenko, Preface: 3rd International Workshop on Modern Machine Learning Technologies and Data Science (MoMLeT &DS 2021), CEUR Workshop Proceedings Vol-2917 (2021).

[12] B. Polishchuk, A. Berko, L. Chyrun, M. Bublyk, V. Schuchmann, The rain prediction in Australia based Big Data analysis and machine learning technology, in: Proceedings of IEEE 16th International conference on computer science and information technologies 2021, 97–100.

[13] A. Demchuk, B. Rusyn, L. Pohreliuk, A. Gozhyj, I. Kalinina, L. Chyrun, N. Antonyuk, Commercial content distribution system based on neural network and machine learning, CEUR Workshop Proceedings 2516 (2019) 40–57.

[14] V. Lytvynenko, W. Wojcik, A. Fefelov, I. Lurie, N. Savina, M. Voronenko, O. Boskin, S. Smailova, Hybrid Methods of GMDH-Neural Networks Synthesis and Training for Solving Problems of Time Series Forecasting, Lecture Notes in Computational Intelligence and Decision Making 1020 (2020) 513–531.

[15] A. Safonyk, M. Mishchanchuk, V. Lytvynenko, Intelligent information system for the determination of iron in coagulants based on a neural network, CEUR Workshop Proceedings 2853 (2021) 142–150.

[16] O., Ivanov, L. Koretska, V. Lytvynenko, Intelligent modeling of unified communications systems using artificial neural networks, CEUR Workshop Proceedings 2623 (2020) 77–84.

[17] S. Babichev, B. Durnyak, O. Sharko, A. Sharko, Technique of metals strength properties diagnostics based on the complex use of fuzzy inference system and hybrid neural network, Communications in Computer and Information Science 1158 (2020) 114–126.

[18] P. Mukalov, O. Zelinskyi, R. Levkovych, P. Tarnavskyi, A. Pylyp, N. Shakhovska, Development of System for Auto-Tagging Articles, Based on Neural Network, CEUR Workshop Proceedings Vol-2362 (2019) 106–115.

[19] S. Leoshchenko, A. Oliinyk, S. Skrupsky, S. Subbotin, T. Zaiko, Parallel Method of Neural Network Synthesis Based on a Modified Genetic Algorithm Application, CEUR Workshop Proceedings Vol-2386 (2019) 11–23.

[20] I. Tsmots, M. Medykovskyy, O. Skorokhoda, Synthesis of hardware components for vertical-group parallel neural networks, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2015, 1–4.

[21] M. O. Medykovskyi, I. G. Tsmots, O. V. Skorokhoda, Spectrum neural network filtration technology for improving the forecast accuracy of dynamic processes in economics, Actual Problems of Economics 162(12) (2014) 410–416.

[22] N. B. Shakhovska, R. Yu. Noha, Methods and tools for text analysis of publications to study the functioning of scientific schools, J. of Automation and Information Sciences 47 (2015) 29–43.

[23] A. Berko, V. Andrunyk, L. Chyrun, M. Sorokovskyy, O. Oborska, O. Oryshchyn, M. Luchkevych, O. Brodovska, The Content Analysis Method for the Information Resources Formation in Electronic Content Commerce Systems, CEUR Workshop Proceedings 2870 (2021) 1632–1651.

[24] V. Kuchkovskiy, V. Andrunyk, M. Krylyshyn, L. Chyrun, A. Vysotskyi, S. Chyrun, N. Sokulska, I. Brodovska, Application of Online Marketing Methods and SEO Technologies for Web Resources Analysis within the Region, CEUR Workshop Proceedings 2870 (2021) 1652–1693.

[25] V. Lytvyn, V. Danylyk, M. Bublyk, L. Chyrun, V. Panasyuk, O. Korolenko, The lexical innovations identification in English-languagee eurointegration discourse for the goods analysis by comments in e-commerce resources, in: Proceedings of IEEE 16th International conference on Computer science and information technologies, Lviv, Ukraine, 2021, 85–97.

[26] A. Gozhyj, L. Chyrun, A. Kowalska-Styczen, O. Lozynska, Uniform method of operative content management in web systems, CEUR Workshop Proceedings 2136 (2018) 62–77.

[27] L. Chyrun, Y. Burov, B. Rusyn, L. Pohreliuk, O. Oleshek, A. Gozhyj, I. Bobyk, Web resource changes monitoring system development, CEUR Workshop Proceedings 2386 (2019) 255–273.

[28] L. Chyrun, A. Kowalska-Styczen, Y. Burov, A. Berko, A. Vasevych, I. Pelekh, Y. Ryshkovets, Heterogeneous data with agreed content aggregation system development, CEUR Workshop Proceedings 2386 (2019) 35–54.

[29] B. Rusyn, L. Pohreliuk, A. Rzheuskyi, R. Kubik, Y. Ryshkovets, L. Chyrun, S. Chyrun, A. Vysotskyi, V.B. Fernandes, The mobile application development based on online music library for socializing in the world of bard songs and scouts' bonfires, Advances in Intelligent Systems and Computing 1080 (2020) 734–756. doi: 10.1007/978-3-030-33695-0_49.

[30] N. Antonyuk, L. Chyrun, V. Andrunyk, A. Vasevych, S. Chyrun, A. Gozhyj, I. Kalinina, Y. Borzov, Medical news aggregation and ranking of taking into account the user needs, CEUR Workshop Proceedings 2488 (2019) 369–382.

[31] T. Yu, G. Yu, M. Y. Wang, Classification method for detecting coercive self-citation in journals, Journal of Informetrics 8(1) (2014) 123–135.

[32] T. Yu, G. Yu, Y. Song, M. Y. Wang, Toward the more effective identification of journals with anomalous self-citation, Malaysian Journal of Library & Information Science 23(2) (2018) 25–46.

[33] M. Szomszor, D. A. Pendlebury, J. Adams, How much is too much? The difference between research influence and self-citation excess, Scientometrics 123(2) (2020) 1119–1147.

[34] R. H. Gálvez, Assessing author self-citation as a mechanism of relevant knowledge diffusion, Scientometrics 111(3) (2017) 1801–1812.

[35] G. Abramo, C. A. D'Angelo, L. Grilli, The effects of citation-based research evaluation schemes on self-citation behavior, Journal of Informetrics 15(4) (2021) 101204.

[36] Y. Liu, M. Chen, Applying text similarity algorithm to analyze the triangular citation behavior of scientists, Applied Soft Computing 107 (2021) 107362.