Correlation and Regression Analyses using Sudoku Grids

Carlos Ml. Rodríguez-Peña[1], José Ramón Martínez Batlle[2], Willy Marcelo Maurer[3]

[1]Instituto de Investigaciones Botánicas y Zoológicas (IIBZ), Universidad Autónoma de Santo Domingo (UASD), Santo Domingo 10105, Dominican Republic

[2]Universidad Autónoma de Santo Domingo (UASD), Santo Domingo 10105, Dominican Republic

[3]Instituto Especializado de Estudios Superiores Loyola (IEESL), San Cristóbal, Dominican Republic

Abstract: ***In this paper, we analysed selected statistical properties of 27, 402 Sudoku grids, which we generated either by using software packages or by consulting sources on the Internet. We classified the Sudoku grids in four different groups according to their provenance as A (10k grids), B (10k grids), C (6.4k grids) and D (~1k grids). We calculated the Pearson product-moment correlation coefficient (r), as well as the corresponding correlation tests, to the 36 maximum possible column pairs of each Sudoku grid. We determined that a maximum of 18 significantly correlated column pairs (SCCP) can be obtained in a single Sudoku grid. In addition, we obtained a total of 42, 826 SCCP (8.68%) out of the 493, 236 possible in our sample. We determined that the number of SCCP with negative r are more common than those with a positive one. We generated linear regression models using SCCP, 32 models resulted for all matrices, 20 with negative correlation values and 12 with positive correlation values. The ratios of negative: positive SCCP for each group yielded 1.00: 0.18 in group A, 1.00: 0.15 in group B, 1.00: 0.16 for group C, 1.00: 0.16 for group D, and an overall ratio of 1.00: 0.16. We found that the number of Sudoku with at least one SCCP was smaller in groups A and B (37.58% and 14.03% respectively) than in groups C and D (89.43% and 89.02% respectively). We hypothesise that the total probability of models can be obtained if an algorithm can be found to build a group of Sudoku in which all SCCP can be found. We transposed each SCCP, so we turned them into position vectors or points. We conveniently assumed that the points belonged to a nine-dimensional real numbers space. We computed squared distance between points pairwise, and formed the Euclidean distance matrices, which we used to classify the SCCP in groups with a hierarchical cluster analysis. We conclude that Sudoku grids are ideal matrices for simulations and modelling with, at least 5.47 billion matrices showing the same characteristics, with different arrangement of numbers.***

Keywords: correlation, negative positive ratios, coefficient of determination, cluster analysis, modelling

1. Introduction

1.1 Origin of Sudoku

Sudoku is a logic (Pillay, 2012) and mathematical (Louis Lee et al., 2008) puzzle used by millions of people around the world (Felgenhauer and Jarvis, 2005; Newton and DeSalvo, 2010; Becerra Tomé et al., 2016). Several scientific and popular papers have been published highlighting the applications of Sudoku grids for several analysis related to significance, complexity, origin, reappearance, evolution, and goals achievement.

Sudoku may have started in the 19th century when Le Siècle, a newspaper from Paris, published a 9×9 magic square, subdivided in a 3×3 cells. However, Sabrin (2009) realised that this was not a Sudoku because it contained double digits, as well as it required arithmetic without logic to solve the puzzle. The newspaper La France (July 6, 1895), a competitor of Le Siècle, introduced changes to this puzzle to position it as it is today. However, differences can be found with today puzzle, because the former allows having more than one solution (Sabrin, 2009). The modern version of Sudoku is considered to be designed anonymously by the retired architect Howard Garns and published by Dell Magazines as Number Places (Grossman, 2013). In 1984, Nikoli Corporation introduced the puzzle to Japan, thanks to the initiatives of its president Mr. Maki Kaji. Because of this introduction, in Japan the puzzle acquired by Maki Kaji changes in 1986 the name to Sudoku (Sabrin, 2009). In Japanese the sound “Su” (soo) means number and “Doku” (doe koo) just one place in the puzzle but the original name in Japanese is “Sujii wa dokushini ni kagiru” (Intelm, 2005). The first computer program to generate Sudoku may have been built by Gould (Cornell University Department of Mathematics, 2009; Gould, 2007).

1.2 Sudoku in Mathematics

Several authors (e.g. Pfaffmann and Collins, 2007; Kwan, 2010; Tengah, 2011; Williams, 2011; Liao and Shih, 2013; Brophy and Hahn, 2014) remark the uses of Sudoku grids as educational models designed for both to aid intellectual steering of the brain and to foster the critical thinking for better understanding of complex problems in mathematics. The solution of Sudoku puzzles is not a trivial problem (Pillay, 2012), so most of the papers found in scientific journals deal with the development of computer algorithms for solving Sudoku puzzles (Maji et al., 2013; Mandal and Sadhu, 2013; McGerty and Moisiadis, 2014).

Pillay (2012) used a genetic programming approach (GPA) algorithms based in Darwin’s paradigm (sensu Kuhn 1963) of evolution. However, evolution was helpful only to reduce the run-time related to the production of solutions for more difficult Sudoku problems.

Newton and DeSalvo (2010) focuses on entropy to build the Sudoku matrices up. Based on their comparison, they conclude that the ensemble averaged Shannon entropy of the collection of Sudoku matrices is slightly lower than a collection of Latin squares, but higher than a collection of appropriately chosen random matrices (Newton and DeSalvo (2010): 1974).

Felgenhauer and Jarvis (2005) calculated the maximum number of Sudoku as 6.67 × 1021 matrices, but in a deeper analysis Russell and Jarvis (2006) considered that, given the symmetries, the numbers of Sudoku matrices are a lot fewer, i.e. 5.47 × 109.

From the mathematical point of view, Williams (2011) wrote a book to learn algebra using Sudoku as a resource for helping both students and teachers to understand more math and to boost creativity. For a philosophy of mathematics, Floyd (2011) analyzed Sudoku under the Wittgenstein philosophy of mathematics perspective. From another mathematical prospective Weyland (2015) discussed how not to solve Sudoku, challenging Geem (2007) harmony search algorithm for solving Sudoku. Weyland’s main argument is that Geem’s harmony search algorithm offers no novelties. Sudoku also have been use in Psychology (Johnson-Laird, 2010; Louis Lee et al., 2008) as mental models and its logic and mathematics nature.

Becerra Tomé et al. (2016) did an important historical review of Sudoku as a puzzle as well as the diversity of forms in it. They also mentioned the mathematical properties of the puzzle. However, they do not pretend to write a scientific paper, because their goal was to make Sudoku understandable widespread.

Bailey et al. (2008) related Sudoku puzzle solution with “gerechte design”, a kind of specialization of Latin squares of the mathematician Leonard Euler (1707-1783) introduced by Behrens (1956). Newton and DeSalvo (2010) and Rouse Ball and Coxeter (1987) tried to establish the relationships of Sudoku with Latin Square and this have been used for four experimental design models by Hui-Dong and Ru-Gen (2008), Danbaba (2016), Danbaba and Dauran (2016) and Shehu and Danbaba (2018).

Pelánek (2011) designed a computational model to test the difficult rating of Sudoku under the critical question “What determine which problems are difficult for humans?”. His main interest was to contribute to human problem-solving difficulty. Related with the problem-solving approach, Williams (2011) used the erasure correcting codes of MacWilliams and Sloane (1977) in Sudoku, which are techniques used for enabling reliable recovery of digital data. On the same topic, Louis Lee et al. (2008) tried to estimate the ability of individuals in pure deductive reasoning through Sudoku puzzles, under the assumptions that the solution cannot be done with pragmatic schemes or innate modules through specific contents.

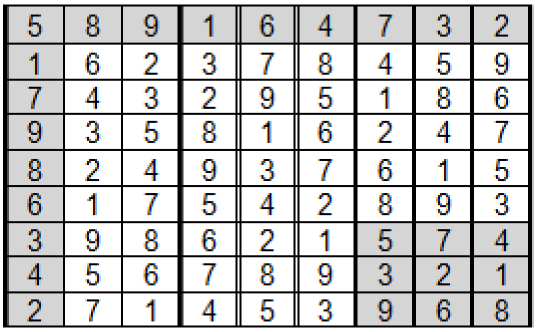
Sudoku’s math was analysed by Farris (2011) who realised that Sudoku is a logical game that has nothing to do with mathematics. However “The sort of reasoning that goes in to solving a Sudoku puzzle, on the other hand, is at the heart of what mathematics is all about” (Rosenhouse and Taalman, 2011) and “When one hears that no math is required to solved Sudoku, what it really means is that no arithmetic is required. In fact, mathematical thinking in the form of logical deduction is very useful in solving Sudoku” (Cornell University Department of Mathematics, 2009). Farris (2011) see Sudoku as a very simple concept using numbers, therefore the game by itself only has relevance to write digits in cells but do not need any sum or subtraction. He remarks that nine different letters, forms or colours instead of the nine digits might be use, but the logic and concept will remain without change, i.e. filling up Sudoku grids with qualitative or quantitative characters, does not have effect on the total number of Sudoku calculated by either Felgenhauer and Jarvis (2005) or Russell and Jarvis (2006). However, the use of Sudoku as a matrix for mathematical analysis requires only numbers, because symbols and letters are less relevant.

Sabrin (2009) developed a multimedia application for Sudoku solution. De Ruiter (2010) analysed several memorizing and optimization techniques, as well as the amount of dissections in a matrix (n) (n) with n numbers of polyomino of a given size n computed from almost all possible digits use in a Sudoku. He found that when n ≥ 4 in a cover polyomino does not allows the solution for any puzzle Sudoku. Regarding the properties of matrix Sudoku see Dahl (2009).

Finally, Sudoku puzzles have also been suggested for applications in Coding Theory (Johnson-Laird, 2010) specifically in the erasure correcting codes basically in the recovery of erasers (see Williams, 2011).

1.3 Sudoku terminology and rules

Each cell in the Sudoku puzzle is a unique box that accepts one single natural number less than or equal to 9. Based on music terminology, each critical name becomes a nonet. Therefore, in the Sudoku puzzle, a nonet row is a horizontal line of cells, a no net column is a vertical line of cells, and a nonet box is a 3×3 block of cells (Fig.1). Therefore, the random Sudoku matrix is redundant



5	8	9	1	6	4	7	3	2
1	6	2	3	7	8	4	5	9
7	4	3	2	9	5	1	8	6
9	3	5	8	1	6	2	4	7
8	2	4	9	3	7	6	1	5
6	1	7	5	4	2	8	9	3
3	9	8	6	2	1	5	7	4
4	5	6	7	8	9	3	2	1
2	7	1	4	5	3	9	6	8

Figure 1: Sudoku matrix showing integer nonet (column, row, box)

(Newton and DeSalvo, 2010) in such a way that all columns are filled up with natural numbers less than or equal to 9, but without repeating any of those numbers in a nonet.

1.4 Rationale of this paper

The objective of this research is to use Sudoku as a matrix for testing models of regression analysis, in order to understand the nature of these relationships within a redundant theoretical system restricted to a number of dimensions and observations. For this, we performed correlation and regression analyses with pairs of columns extracted from Sudoku grids. We used Pearson's r to establish the relationships between the variables (see Lee Rodgers and Nicewander, 1988). We faced a big challenge to perform this task, because we performed thousands of correlation analysis to our sample data. In addition, we performed regression analysis with data extracted from Sudoku grids. Regression analysis is a very common statistical tool used for building models that, under proper conditions, may be suitable for predictive purposes Chambers and Hastie (1992). However, in this research, we used regression analysis for building as many models as possible with pairs of columns from the Sudoku.

2. Materials and Methods

We analysed 27, 402 Sudoku grids, which we generated either by using algorithms included in both open-source and commercial software packages, such as R Core Team (2015), or by consulting sources on the Internet. Hence, we classified the Sudoku grids in four different groups according to their provenance, which are described as follows.

- Group A. We generated 10, 000 grids using the function generate sudoku from sudokuplus R package (Gan et al., 2012). The algorithm chooses a random number of cells to leave blank in a "primordial" grid, by using the uniform distribution. Afterwards, the algorithm itself solves the grid with an array of logical conditions and discarding the erroneous solutions one by one, until all rules of Sudoku are satisfied. If an inconsistency arises, it throws an error. We used a for loop (control flow statement for repeating code) to generate the 10k grids.
- Group B. We generated 10, 000 grids using the function generate Sudoku from the Sudoku R package (Brahm et al., 2009, 2014). The algorithm starts from a "primordial" Sudoku grid, which contains a predefined number of blank cells. Afterwards, the algorithm randomly swaps around rows and columns to fill the blank cells, ensuring compliance of basic rules of Sudoku. We generated 10k grids using a for loop and using a different random number generator (seed) in each iteration.
- Group C. We downloaded and "scrapped" an anonymous PDF file obtained from www.b-ok.org, from which we retrieved 6, 400 grids. We confirmed that all the grids from this collection fulfilled the Sudoku rules.
- Group D. We generated 1, 002 grids using LibreOffice Calc. For this task, we established the Sudoku rules in an empty sheet. Then we filled successive arrays of 9 × 9 cells, either by hand or with the aid of a random number generator. Simultaneously, while typing the numbers in the sheet, we checked for the compliance of the rules.

We calculated that 36 column pairs can be combined in a Sudoku grid without repetition and regardless the order of selection, by using the typical equation of the binomial coefficient:

$$\binom{n}{k} = {}^{n}C_{k} = \frac{n!}{k!(n-k)!} = {}^{9}C_{2} = \frac{9!}{2!(9-2)!} = 36$$

where n is the number of columns in the Sudoku grid (9), and k is the number of columns to combine (2).

In addition, we calculated the Pearson product moment correlation coefficient (r) of the 36 column pairs of each matrix. We performed correlation tests to determine the statistically significant correlated column pairs (here-after SCCP), for which we arbitrarily set the significance level (α) at 0.05. We classified the SCCP in 32 subgroups according to their corresponding r values.

It is worth noting that we conducted no exploratory data analysis, since our objective was initially set to use the Pearson coefficient without considering the mandatory assumptions of the correlation analysis.

For each group of Sudoku grids (A, B, C and D), we analysed the subgroups of SCCP with similar r value to simultaneously detect scatter plot redundancy and reduce the number of SCCP. To accomplish this task, we transposed the N columns forming SCCP with similar r value, so we turned them into N position vectors or points $\{x_l, l = 1, 2, \ldots, N\}$. We were aware that the numbers of the Sudoku grids belong to N, which is not a vector space. Therefore, we conveniently assumed that our points belong to R^m space, where m = 9, for which we computed distance-square d_{ij} between points pairwise. After addressing all possible pairs of points, where i or j = {1, 2, . . ., N}, we formed the Euclidean distance matrices Δ in $R_+^{N\times N}$Dattorro (2005, v2011.04.25):

$$\begin{aligned}
x_i &= (x_{i1}, x_{i2}, \ldots, x_{i9})\\
x_j &= (x_{j1}, x_{j2}, \ldots, x_{j9})\\
d_{ij} &= \|x_i - x_j\|_2^2 \triangleq \langle x_i - x_j, x_i - x_j\rangle\\
&= \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \cdots + (x_{i9}-x_{j9})^2}\\
&= \sqrt{\sum_{m=1}^{9}(x_{im}-x_{jm})^2}\\
\Delta = [d_{ij}] &= \begin{bmatrix} 0 & d_{12} & \ldots & d_{1N}\\ d_{21} & 0 & \ldots & d_{2N}\\ \vdots & \vdots & \ddots & \vdots\\ d_{N1} & d_{N2} & \ldots & 0 \end{bmatrix}
\end{aligned}$$

Afterwards, we performed regression analyses to fit into lineal models each column of the SCCP subset, using least-squares as a fitting criterion (Chambers and Hastie, 1992), according to the following lineal equation model:

$$Y = a + bX$$

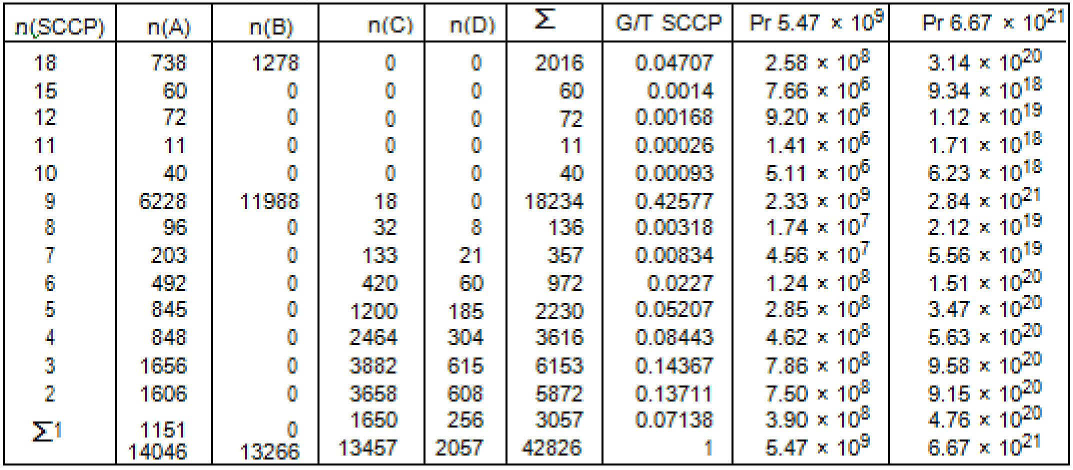
where Y is the dependent variable, X is the independent variable, and a and b are the coefficients.

Finally, we performed hierarchical cluster analyses applying the Ward criterion and using the distance matrices as input, with the goal of producing tree structures (Ward Jr, 1963; Murtagh and Legendre, 2014). We plotted the trees into circular layouts, which we consider an efficient way to arrange our results.

3. Results

We found that the maximum number of SCCP per Sudoku was 18 out of 36 which happens Within the two larger groups (A and B). A summary of the number of SCCP by group is seen in Table 1:

Table 1: Number of SCCP by group (A, B, C, D) and the resulting probabilities, based on the maximum number of Sudoku estimated by Felgenhauer and Jarvis (2005) and by Russell and Jarvis (2006)



n(SCCP)	n(A)	n(B)	n(C)	n(D)	Σ	G/T SCCP	Pr 5.47 × 10^9	Pr 6.67 × 10^{21}
18	738	1278	0	0	2016	0.04707	2.58 × 10^8	3.14 × 10^{20}
15	60	0	0	0	60	0.0014	7.66 × 10^6	9.34 × 10^{18}
12	72	0	0	0	72	0.00168	9.20 × 10^6	1.12 × 10^{19}
11	11	0	0	0	11	0.00026	1.41 × 10^6	1.71 × 10^{18}
10	40	0	0	0	40	0.00093	5.11 × 10^6	6.23 × 10^{18}
9	6228	11988	18	0	18234	0.42577	2.33 × 10^9	2.84 × 10^{21}
8	96	0	32	8	136	0.00318	1.74 × 10^7	2.12 × 10^{19}
7	203	0	133	21	357	0.00834	4.56 × 10^7	5.56 × 10^{19}
6	492	0	420	60	972	0.0227	1.24 × 10^8	1.51 × 10^{20}
5	845	0	1200	185	2230	0.05207	2.85 × 10^8	3.47 × 10^{20}
4	848	0	2464	304	3616	0.08443	4.62 × 10^8	5.63 × 10^{20}
3	1656	0	3882	615	6153	0.14367	7.86 × 10^8	9.58 × 10^{20}
2	1606	0	3658	608	5872	0.13711	7.50 × 10^8	9.15 × 10^{20}
Σ1	1151	0	1650	256	3057	0.07138	3.90 × 10^8	4.76 × 10^{20}
	14046	13266	13457	2057	42826	1	5.47 × 10^9	6.67 × 10^{21}

Legend: n (SCCP), number of statistically significant correlated column pairs (SCCP) per Sudoku; n (A), n (B), n (C) and n (D), number of SCCP in Sudoku grids with n (SCCP) in groups A, B, C and D, respectively; G/T SCCP, proportion of n (SCCP) relative to the total number of SCCP along the four groups; Pr 5.47 × 10^9 expected number of Sudoku grids with n (SCCP) based on the maximum number of grids estimated by Felgenhauer and Jarvis (2005); Pr 6.67 × 10^{21} expected number of Sudoku grids with n (SCCP) based on the maximum number of grids estimated by Russell and Jarvis (2006)

Group A lack Sudoku grids with 17-16 and 14-13 SCCP, therefore the 14 SCCP found were 18, with relative frequency of 0.0041; 15 with relative frequency of 0.0004. For 12-9 SCCP, relative frequency were 0.0006, 0.0001, 0.0004 and 0.0692 respectively. For 8-1 SCCP with relative frequency of 0.0012, 0.0029, 0.0082, 0.0169, 0.0212, 0.0552, 0.0803 and 0.1151 respectively, the remaining 0.6242 did not show SCCP (Table 1). Group B lack probabilities 17-10 and 8-1, therefore the only two SCCP were 18, with relative frequency of 0.0071, and 9, with relative frequency of 0.1332; the remaining 0.8597 did not show SCCP (Table 1).

Within the medium size group C (6.4k) Sudoku, SCCP were 9-1, with 9-7 SCCP at a very low relative frequency, i.e. 0.0003, 0.0006 and 0.0030 respectively; relative frequency from 6-1, are 0.0109, 0.0375, 0.0963, 0.2022, 0.2857, 0.2578 respectively. A similar pattern shows by the small D (1.0k) size group Sudoku with number of SCCP from 8-1, with 8-7 SCCP 0.0010 and 0.0030 respectively; relative frequency of 6-1 were 0.0100, 0.0349, 0.0768, 0.2046, 0.3044, 0.2555 respectively (Table 1).

The correlation coefficients were between -0.983≤r≤0.900, with the correspondent 0.001≤ p ≤0.04 -where p is the "p-value" in the previous hypothesis test carried out previously- and coefficient of determination ranging 0.445≤ r^2 ≤0.967 (Table 2). SCCP 42826 represent 8.68% of 493236. Segregated by group, non-segregated by Sudoku: a) 14046 SCCP out of 180000 (7.80%) SCCP for A; b) and 13266 SCCP out of 180000 (7.37%) SCCP for B;

Table 2: Pearson correlation coefficient and coefficient of determination in each group Sudoku grids (A, B, C, D), frequency (f) and relative frequency (rf)

r	r^2	SCCP				Proportion of SCCP				Prop. relative to total	
		A	B	C	D	rf A	rf B	rf C	rf D	f42826	rf42826
-0.983	0.966	0	0	5	1	0.00000	0.00000	0.00037	0.00049	6	0.00014
-0.967	0.935	5	0	29	3	0.00036	0.00000	0.00216	0.00146	37	0.00086
-0.95	0.903	17	0	36	9	0.00121	0.00000	0.00268	0.00438	62	0.00145
-0.933	0.870	61	9	43	13	0.00434	0.00068	0.0032	0.00632	126	0.00294
-0.917	0.841	142	72	113	18	0.01011	0.00543	0.0084	0.00875	345	0.00806
-0.900	0.810	118	126	116	20	0.00840	0.00950	0.00862	0.00972	380	0.00887
-0.883	0.780	155	99	197	21	0.01104	0.00746	0.01464	0.01021	472	0.01102
-0.867	0.752	224	234	261	40	0.01595	0.01764	0.01940	0.01945	759	0.01772
-0.850	0.723	293	261	303	47	0.02086	0.01967	0.02252	0.02285	904	0.02111
-0.833	0.694	541	486	407	70	0.03852	0.03664	0.03024	0.03403	1504	0.03512
-0.817	0.667	492	243	526	69	0.03503	0.01832	0.03909	0.03354	1330	0.03106
-0.800	0.640	540	603	626	106	0.03845	0.04545	0.04652	0.05153	1875	0.04378
-0.783	0.613	580	594	641	108	0.04129	0.04478	0.04763	0.05250	1923	0.04490
-0.767	0.588	840	648	909	127	0.05980	0.04885	0.06755	0.06174	2524	0.05894
-0.750	0.563	1016	1206	873	103	0.07233	0.09091	0.06487	0.05007	3198	0.07467
-0.733	0.537	1197	1134	994	152	0.08522	0.08548	0.07386	0.07389	3477	0.08119
-0.717	0.514	1292	1413	1162	160	0.09198	0.10651	0.08635	0.07778	4027	0.09403
-0.700	0.490	1312	1269	1261	205	0.09341	0.09566	0.09371	0.09966	4047	0.09450
-0.683	0.466	1400	1341	1355	245	0.09967	0.10109	0.10069	0.11911	4341	0.10136
-0.667	0.445	1682	1845	1739	254	0.11975	0.13908	0.12923	0.12348	5520	0.12889
0.667	0.445	110	54	210	28	0.00783	0.00407	0.01561	0.01361	402	0.00939
0.683	0.466	267	216	320	54	0.01901	0.01628	0.02378	0.02625	857	0.02001
0.700	0.490	228	180	280	39	0.01623	0.01357	0.02081	0.01896	727	0.01698
0.717	0.514	837	729	265	34	0.05959	0.05495	0.01969	0.01653	1865	0.04355
0.733	0.537	143	72	154	29	0.01018	0.00543	0.01144	0.01410	398	0.00929
0.750	0.563	80	9	78	14	0.00570	0.00068	0.00580	0.00681	181	0.00423
0.767	0.588	56	0	112	15	0.00399	0.00000	0.00832	0.00729	183	0.00427
0.783	0.613	108	0	188	30	0.00769	0.00000	0.01397	0.01458	326	0.00761
0.800	0.640	37	0	75	14	0.00263	0.00000	0.00557	0.00681	126	0.00294
0.833	0.694	18	0	71	10	0.00128	0.00000	0.00528	0.00486	99	0.00231
0.850	0.723	250	423	87	18	0.01780	0.03189	0.00647	0.00875	778	0.01817
0.900	0.810	5	0	21	1	0.00036	0.00000	0.00156	0.00049	27	0.00063
Total		14046	13266	13457	2057					42826	

c) 13457 SCCP out of 115200 (11.80%) SCCP for C, and d) for 2057 SCCP out of 18036 (11.40%) SCCP for D (Fig. 3).

The output of the analysis, by Sudoku, within each group (A, B, C, D) shows that the distribution of SCCP, proportion of r, the ratios of negative: positive correlation as well as the general ratios for the whole groups and the total ratios for the four groups together, are as the descriptions below.

Distribution of SCCP by Sudoku shows that 9 SCCP were the common ones, 42.58%, followed by 3 SCCP (14.37%), 2 SCCP (13.71%), 4 SCCP (8.44%), 1 SCCP (7.14%), 5 SCCP (5.21%), 18 SCCP (4.71%), 6 SCCP (2.37%), 7 SCCP (0.83%), 8 SCCP (0.32%), 12 SCCP (0.17%), 15 SCCP (0.14%), 10 SCCP (0.09%), and 11 SCCP with 0.03% (Fig. 4).

For Group A (Tables 1, 2, 3, 4). - The 18 SCCP were the higher amount, 41 cases (41x18 = 738 SCCP); the variability and relative frequency of r

Table 3: Ratios of negative and positive (Neg: Pos) correlation, segregated by group (A, B, C, D) of SCCP Sudoku

n (SCCP)	A		B		C		D	
	n	Neg: Pos	n	Neg: Pos	n	Neg: Pos	n	Neg: Pos
1	1151	1.00: 0.15	0	0	1650	1.00: 0.11	256	1.00: 0.12
2	1606	1.00: 0.16	0	0	3658	1.00: 0.13	608	1.00: 0.14
3	1656	1.00: 0.16	0	0	3882	1.00: 0.16	615	1.00: 0.14
4	848	1.00: 0.20	0	0	2464	1.00: 0.20	304	1.00: 0.22
5	845	1.00: 0.18	0	0	1200	1.00: 0.22	185	1.00: 0.28
6	492	1.00: 0.17	0	0	420	1.00: 0.25	60	1.00: 0.20
7	203	1.00: 0.42	0	0	133	1.00: 0.32	21	1.00: 0.24
8	96	1.00: 0.45	0	0	32	1.00: 0.39	8	1.00: 0.14
9	6228	1.00: 0.12	11988	1.00: 0.10	18	1.00: 0.50	0	0
10	40	1.00: 0.67	0	0	0	0	0	0
11	11	0.83: 1.00	0	0	0	0	0	0
12	72	1.00: 0.38	0	0	0	0	0	0
15	60	1.00: 0.67	0	0	0	0	0	0
18	738	1.00: 1.00	1278	1.00: 1.00	0	0	0	0

Legend: Neg: Pos, negative: positive correlation ratio within each group; n (SCCP), number of statistically significant correlated column pairs (SCCP) per Sudoku; n, number of SCCP in Sudoku grids with n (SCCP) in each group.

were between -0.933 (0.02439) and 0.750 (0.02439), here 0.717 (0.18292), 0.683 (0.12195), and -0.917 (0.10976) were the common ones. Negative (369): positive (369) correlation ratio is 1.00: 1.00. For 15 SCCP a total of 60 cases, where the variability and relative frequency of r were between -0.950 (0.05000) and 0.783 (0.15000), here -0.767, 0.717, and 0.783 were the common ones with a relative frequency of 0.15000 each (0.45000). Negative (36): positive (24) correlation ratio is 1.00: 0.67. For 12 SCCP a total of 72 cases; the variability and relative frequency of r were between -0.917 (0.106944) and 0.800 (0.08333), here -0.767 (0.16667) and -0.912 (0.10694) were the common ones. Negative (52): positive (20) correlation ratio is 1.00: 0.38. For 11 SCCP a total of 11 cases; the variability and relative frequency of r were between -0.883 (0.45454) and 0.733 (0.45454), also the common ones. Negative (5): positive (6) correlation ratio is 1.00: 0.833. For 10 SCCP a total of 40 cases; the variability and relative frequency of r were between -0.933 (0.12500) and 0.767 (0.12500), here 0.733 and 0.750 with 0.12500 each (0.2500) were the common ones. Negative (24): positive (16) correlation ratio is 1.00: 0.67. For 9 SCCP, a total of 6228 cases; the variability and relative frequency of r were between -0.883 (0.00145) and 0.850 (0.02890), here -0.667 (0.14355) and -0.733, -0.717 and -0.700 were the common ones with 0.10308 each. Negative (5550): positive (678) correlation ratio is 1.00: 0.12. For 8 SCCP a total of 96 cases; the variability and relative frequency of r were between -0.900 (0.01042) and 0.850 (0.04167), here -0.667 (0.14583) and -0.700 (0.11458) were the common ones. Negative (66): positive (30) correlation ratio is 1.00: 0.45. For 7 SCCP a total of 203 cases; the variability and relative frequency of r were between -0.950 (0.01970) and 0.850 (0.05419), here -0.667 (0.10345) was the common one. Negative (143): positive (60) correlation ratio is 1.00: 0.42. For 6 SCCP, total of 492 cases; the variability and relative frequency of r were between -0.933 (0.00407) and 0.900 (0.00209), here -0.683 (0.12805) was the common one. Negative (422): positive (70) correlations ratio is 1.00: 0.17.

Table 4: Absolute number of SCCP per value of r and number of SCCP per Sudoku in group A

r	1	2	3	4	5	6	7	8	9	10	11	12	15	18
-0.983	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.967	1	3	1	0	0	0	0	0	0	0	0	0	0	0
-0.950	3	2	1	4	0	0	4	0	0	0	0	0	3	0
-0.933	3	13	6	10	1	2	0	0	0	5	0	0	3	18
-0.917	7	11	13	10	3	2	2	0	0	5	0	5	3	81
-0.900	11	7	7	9	4	3	1	1	0	0	0	3	0	72
-0.883	16	21	20	18	6	3	3	0	9	0	5	0	0	54
-0.867	20	26	31	24	12	15	4	1	36	1	0	6	3	45
-0.850	22	37	32	18	11	11	3	2	153	0	0	4	0	0
-0.833	41	44	45	22	29	21	1	1	274	0	0	3	6	54
-0.817	50	64	56	22	29	22	8	5	229	4	0	0	3	0
-0.800	43	61	67	39	23	36	7	3	237	0	0	6	0	18
-0.783	66	88	68	30	42	30	18	6	228	1	0	0	3	0
-0.767	69	101	115	50	56	36	9	8	373	2	0	12	9	0
-0.750	78	102	101	43	57	28	12	5	561	1	0	1	0	27
-0.733	85	113	115	57	80	43	18	4	675	4	0	3	0	0
-0.717	127	136	175	79	81	43	6	3	642	0	0	0	0	0
-0.700	111	173	154	78	92	28	16	11	642	1	0	6	0	0
-0.683	135	182	215	106	89	63	10	3	597	0	0	0	0	0
-0.667	120	197	204	89	103	36	21	14	894	1	0	0	3	0
0.667	10	15	18	20	18	1	6	1	0	0	0	3	0	18
0.683	21	31	43	20	10	12	16	7	5	3	0	6	3	90
0.700	17	49	41	17	22	10	4	5	6	0	0	3	0	54
0.717	41	40	47	23	28	15	3	4	487	1	1	3	9	135
0.733	10	19	10	15	21	1	2	1	0	5	5	0	0	54
0.750	4	16	10	6	3	10	4	1	0	5	0	0	3	18
0.767	11	9	14	7	4	6	3	1	0	1	0	0	0	0
0.783	11	17	23	15	7	9	10	5	0	0	0	2	9	0
0.800	0	12	6	9	3	1	0	0	0	0	0	6	0	0
0.833	2	6	3	1	4	1	1	0	0	0	0	0	0	0
0.850	15	10	14	7	6	3	11	4	180	0	0	0	0	0
0.900	1	1	1	0	1	1	0	0	0	0	0	0	0	0
Total	1151	1606	1656	848	845	492	203	96	6228	40	11	72	60	738

Legend: r, Pearson correlation coefficient for each SCCP; 1-18, number of SCCP per Sudoku.

For 5 SCCP, a total of 845 cases; the variability and relative frequency of r were between -0.933 (0.00118) and 0.900 (0.00118), here -0.667 (0.12189) was the common one. Negative (718): positive (127) correlation ratio is 1.00: 0.18.

For 4 SCCP, a total of 848 cases; the variability and relative frequency of r were between -0.950 (0.00472) and 0.850 (0.00825), here -0.683 (0.12500) and -0.667 (0.10495) were the common ones. Negative (708): positive (140) correlation ratio is 1.00: 0.20. For 3 SCCP, a total of 1656 cases; the variability and relative frequency of r were between -0.967 (0.00060) and 0.900 (0.00060), where -0.683 (0.12983) and -0.667 (0.12319) are the common ones. Negative (1426): positive (230) correlation ratio is 1.00: 0.16.

Table 5: Absolute number of SCCP per value of r and number of SCCP per Sudoku in group B

r	9	18
-0.983	0	0
-0.967	0	0
-0.950	0	0
-0.933	0	9
-0.917	0	72
-0.900	0	126
-0.883	18	81
-0.867	90	144
-0.850	234	27
-0.833	450	36
-0.817	243	0
-0.800	540	63
-0.783	567	27
-0.767	648	0
-0.750	1152	54
-0.733	1134	0
-0.717	1413	0
-0.700	1269	0
-0.683	1341	0
-0.667	1845	0
0.667	0	54
0.683	0	216
0.700	0	180
0.717	621	108
0.733	0	72
0.750	0	9
0.767	0	0
0.783	0	0
0.800	0	0
0.833	0	0
0.850	423	0
0.900	0	0
Total	11988	1278

Legend: r, Pearson correlation coefficient for each SCCP; 9 and 18, number of SCCP per Sudoku.

Table 6: Absolute number of SCCP per value of r and number of SCCP per Sudoku in group C

r	1	2	3	4	5	6	7	8	9
-0.983	2	2	0	1	0	0	0	0	0
-0.967	6	4	14	5	0	0	0	0	0
-0.950	8	4	10	8	3	3	0	0	0
-0.933	5	12	12	8	3	1	2	0	0
-0.917	15	36	34	17	9	2	0	0	0
-0.900	19	23	37	22	13	1	1	0	0
-0.883	22	52	63	38	18	2	1	1	0
-0.867	31	72	73	51	17	12	4	1	0
-0.850	33	89	78	58	18	24	3	0	0
-0.833	48	116	116	81	31	12	2	1	0
-0.817	71	159	137	93	42	18	4	0	2
-0.800	97	189	168	101	51	14	5	1	0
-0.783	59	187	197	110	66	11	7	2	2
-0.767	135	262	256	162	72	15	7	0	0
-0.750	100	275	256	120	76	26	16	2	2
-0.733	98	267	321	168	107	22	8	3	0
-0.717	178	281	353	212	91	35	6	4	2
-0.700	147	369	343	247	104	35	9	3	4
-0.683	158	379	398	252	107	47	12	2	0
-0.667	252	468	484	307	159	52	14	3	0
0.667	25	45	65	48	15	7	3	2	0
0.683	26	76	106	63	29	14	2	2	2
0.700	26	56	75	75	34	10	4	0	0
0.717	23	47	87	47	34	14	10	1	2
0.733	15	38	42	30	23	6	0	0	0
0.750	9	22	17	18	9	2	0	1	0
0.767	8	32	27	18	13	9	4	1	0
0.783	15	48	44	40	24	11	3	1	2
0.800	7	14	20	17	8	4	5	0	0
0.833	5	17	18	19	9	2	1	0	0
0.850	5	12	26	22	14	7	0	1	0
0.900	2	5	5	6	1	2	0	0	0
Total	1650	3658	3882	2464	1200	420	133	32	18

Legend: r, Pearson correlation coefficient for each SCCP; 1-9, number of SCCP per Sudoku.

For 2 SCCP, a total of 1606 cases; the variability and relative frequency of r were between-0.967 (0.00187) and 0.900 (0.00062), here -0.667 (0.12267), -0.683 (0.11333), and -0.7000 (0.10772) are the common ones. Negative (1381): positive (225) correlation ratio is 1.00: 0.16. For 1 SCCP, a total of 1151 cases; the variability and relative frequency of r were between -0.967 (0.00087) and 0.900 (0.00087), here -0.683 (0.11729), -0.717 (11034), and -0.667 (0.10426) were the common ones. Negative (1008): positive (153) correlation ratio is 1.00: 0.15. General negative (11907): positive (2139) ratio for this group (A) is 1.00: 0.18 (Table 3).

For Group B (Tables 1, 3, 2, 5).- The 18 SCCP was the higher amount with 71 cases (71×18=1278 SCCP), a total of 1278 cases; the variability and relative frequency of r were between -0.933 (0.00704) and 0.750 (0.00704), here 0.683 (0.16901), 0.700 (0.14085), and -0.867 (0.11268) were the common ones. Negative (639): positive (639) correlation ratio is 1.00: 1.00. For 9 SCCP, a total of 11988 cases; the variability and relative frequency of r were between -0.883 (0.00150) and 0.850 (0.03529), here -0.667 (0.15390), -0.717 (0.11787), and -0.683 (0.11186) were the common ones. Negative (10944): positive (1044) correlation ratio is 1.00: 0.10. General negative (11583): positive (1683) correlation ratio for this group (B) is 1.00: 0.15. (Table 3).

For Group C (Tables 1, 3, 2, 6).- The 9 SCCP was the higher amount with 2 cases (2x9=18 SCCP), a total of 18 cases; the variability and relative frequency of r were between -0.817 and 0.783 with 0.11111 each, here -0.700 (0.22222) was the common one, but it is important to remarks that the rest of them were 0.11111. Negative (12): positive (6) correlation ratio is 1.00: 0.50. For 8 SCCP, a total of 32 cases; the variability and relative frequency of r were between -0.883 (0.03125) and 0.850 (0.03125), here commonalities are 0.06250 and 0.03125. Negative (23): positive (9) correlation ratio is 1.00: 0.39. For 7 SCCP, a total of 133 cases; the variability and relative frequency of r were between -0.933 (0.01504) and 0.833 (0.00752), here -0.750 (0.12030) and -0.667 (0.10526) were common ones. Negative (101): positive (32) correlation ratio is 1.00: 0.32. For 6 SCCP, a total of 420 cases; the variability and relative frequency of r were between -0.950 (0.00714) and 0.900 (0.00472), here -0.667 (0.12381) and -0.683 (0.11190) were the common ones. Negative (332): positive (88) correlation ratio is 1.00: 0.25.

For 5 SCCP, a total of 1200 cases; the variability and relative frequency of r were between -0.950 (0.00250) and 0.900 (0.00083), here -0.667 (0.13250) is the common one. Negative (987): positive (213) correlation ratio is 1.00: 0.22. For 4 SCCP, a total of 2464 cases; the variability and relative frequency of r were between -0.983 (0.00041) and 0.900 (0.00244), here -0.667 (0.12459), -0.683 (0.10227), and-0.700 (0.10024) were the common ones. Negative (2061): positive (403) correlation ratio is 1.00: 0.20. For 3 SCCP, a total of 3882 cases; the variability and relative frequency of r were between -0.967 (0.00361) and 0.900 (0.00129), here -0.667 (0.12468) and -0.683 (0.10252) were the common ones. Negative (3350): positive (532) correlation ratio is 1.00: 0.16.

For 2 SCCP, a total of 3658 cases; the variability and relative frequency of r were between -0.983 (0.00055) and 0.900 (0.00137), here -0.667 (0.12794), -0.683 (0.10361), and -0.700 (0.10087) were the common ones. Negative (3256): positive (412) correlation ratio is 1.00: 0.13. For 1 SCCP, a total of 1650 cases; the variability and relative frequency of r were between -0.983 (0.00121) and 0.900 (0.00121), here -0.667 (0.15273) was the common one. Negative (1484): positive (166) correlation ratio is 1.00: 0.11. General negative (11596): positive (1861) correlation ratio for this group (C) is 1.00: 0.16 (Table 3).

For Group D (Tables 1, 2, 3, 7). - The 8 SCCP was the higher amount with 1 case; the variability and relative frequency of r were between -0.833 (0.12500) and 0.683 (0.12500), here -0.700 (0.25000) was the common ones, all others were 0.12500. Negative (7): positive (1) correlation ratio is 1.00: 0.14. For 7 SCCP, a total of 21 cases; the variability and relative frequency of r were between -0.917 (0.04762) and 800 (0.04762), here -0.717 and -0.667 were 0.14286 each. Negative (17): positive (4) correlation ratio is 1.00: 0.24. For 6 SCCP, a total of 60 cases; the variability and relative frequency of r were between -0.933 (0.01667) and 0.833 (0.01667), here -0.683 and -0.667 with 0.13333 each, and -0.783 (0.10000) were the common ones. Negative (50): positive (10) correlation ratio is 1.00: 0.20. For 5 SCCP, a total of 185 cases; the variability and relative frequency of r were between -0.967 (0.00541) and 0.850 (0.02162), here -0.683 (0.12432) was the common one. Negative (144): positive (41) correlation ratio is 1.00: 0.28. For 4 SCCP, a total of 304 cases; the variability and relative frequency of r were between -0.967 (0.00329) and 0.900 (0.00329), here -0.667 (0.12829), -0.700 (0.11184), and -0.683 (0.10526) were the common ones. Negative (249): positive (55) correlation ratio is 1.00: 0.22. For 3 SCCP, a total of 615 cases; the variability and relative frequency of r were between -0.950 (0.00163) and 0.850 (0.00650), here -0.667 (0.13496), -0.683 (0.12683), and -0.700 (0.10081) were the common ones. Negative (540): positive (75) correlation ratio is 1.00: 0.14.

Table 7: Absolute number of SCCP per value of r and number of SCCP per Sudoku in group D

r	1	2	3	4	5	6	7	8
-0.983	0	1	0	0	0	0	0	0
-0.967	0	1	0	1	1	0	0	0
-0.950	0	5	1	3	0	0	0	0
-0.933	5	3	2	1	1	1	0	0
-0.917	1	7	5	1	3	0	1	0
-0.900	3	7	6	3	0	1	0	0
-0.883	2	10	3	3	2	1	0	0
-0.867	6	14	12	4	1	3	0	0
-0.850	6	15	11	9	5	1	0	0
-0.833	13	22	19	7	8	0	0	1
-0.817	14	18	19	7	7	2	2	0
-0.800	16	30	34	14	9	2	1	0
-0.783	14	39	35	9	3	6	1	1
-0.767	21	27	47	17	9	5	1	0
-0.750	7	33	32	16	12	2	0	1
-0.733	17	47	47	24	12	2	2	1
-0.717	20	47	44	25	16	4	3	1
-0.700	28	60	62	34	14	4	1	2
-0.683	27	75	78	32	23	8	2	0
-0.667	29	74	83	39	18	8	3	0
0.667	3	8	7	7	3	0	0	0
0.683	6	13	8	12	10	3	1	1
0.700	5	8	15	8	1	2	0	0
0.717	1	9	16	5	2	0	1	0
0.733	2	10	5	4	7	0	1	0
0.750	2	3	2	4	3	0	0	0
0.767	1	2	5	5	1	1	0	0
0.783	2	10	6	5	6	1	0	0
0.800	2	3	5	0	1	2	1	0
0.833	0	1	2	3	3	1	0	0
0.850	3	6	4	1	4	0	0	0
0.900	0	0	0	1	0	0	0	0
Total	256	608	615	304	185	60	21	8

Legend: r Pearson correlation coefficient for each SCCP; 1-8, number of SCCP per Sudoku.

For 2 SCCP, a total of 608 cases; the variability and relative frequency of r were between -0.983 (0.00164) and 0.850 (0.00987), here -0.683 (0.12336) and -0.667 (0.12171) were the common ones. Negative (535): positive (73) correlation ratio is 1.00: 0.14. For 1 SCCP, a total of 256 cases; the variability and relative frequency of r were between -0.933 (0.01953) and 0.850 (0.01172), here -0.667 (0.11328), -0.700 (0.10938), and -0.683 (0.10547) were the common ones. Negative (229): positive (27) correlation ratio is 1.00: 0.12. General negative (1771): positive (286) ratio for this group (D) is 1.00: 0.16 (Table 3).

In general, of the 42826 SCCP out of 493, 236 total probabilities of SCCP, segregated by Sudoku without caring about to which one it belongs, negative and positive correlations were 36857 (86.49%) and 5969 (13.51%) respectively (Fig. 3) with a ratio (1.00: 0.16). Within the negative correlation were r = - 0.667 (12.89%), r = -0.683 (10.14%), r = -0.700 (9.44%), r = -0.717 (9.40%), r = -

Table 8: Negative: Positive ratios, frequencies, proportions and summary

Neg: Pos	Fr	Pr	%
1.00: 0.10	11988	0.2799	27.99
1.00: 0.11	1650	0.0385	3.85
1.00: 0.12	7878	0.1514	15.14
1.00: 0.13	3658	0.0854	8.54
1.00: 0.14	1231	0.0287	2.87
1.00: 0.15	1151	0.0269	2.69
1.00: 0.16	7144	0.1668	16.68
1.00: 0.17	492	0.0115	1.15
1.00: 0.18	845	0.0197	1.97
1.00: 0.20	3372	0.0787	7.87
1.00: 0.22	1504	0.0351	3.51
1.00: 0.24	21	0.0005	0.05
1.00: 0.25	420	0.0098	0.98
1.00: 0.28	185	0.0043	0.43
1.00: 0.32	133	0.0031	0.31
1.00: 0.38	72	0.0017	0.17
1.00: 0.39	32	0.0007	0.07
1.00: 0.42	203	0.0047	0.47
1.00: 0.45	96	0.0022	0.22
1.00: 0.50	18	0.0004	0.04
1.00: 0.67	100	0.0023	0.23
1.00: 1.00	2016	0.0471	4.71
0.83: 1.00	11	0.0003	0.03
Overall Neg: Pos		Sume	
1.00: 0.16	42826	1	100

Legend: Neg: Pos ratios; Fr, frequency of ratios; Pr and %, proportions and % of the total.

0.733 (8.12%), r = -0.750 (7.47%) and r = -0.767 (5.89%), representing 63.35% of all 32 correlations (negative and positive) and 35.00% (7 out of 20) negative correlation, therefore in 21.88% (7 out of 32) of the total. For positive correlation the larger percentages were r = 0.717 (4.35%), r = 0.683 (2.00%), r = 0.850 (1.82%), and r = 0.700 (1.70%), representing 9.87% of all 32 correlations, and 33.33% (4 out of 12) positive correlations, therefore 12.50% (4 out of 32) of the total. The smallest percentages corresponded to the extreme (lowest and highest) correlations, -0.983 (0.01401%) and 0.900 (0.06305%) (Table 2; Fig. 4).

The most frequent negative: positive ratios are: 1.00: 0.10 (27.99%), 1.00: 0.16 (16.68%), 1.00: 0.12 (15.14%), 1.00: 0.13 (8.54%), 1.00: 0.20 (7.87%), 1.00: 1.00 (4.71%), 1.00: 0.11 (3.85%), 1.00: 0.12 (3.51%), 1.00: 0.14 (2.87%), 1.00: 0.15 (2.69%), 1.00: 0.18 (1.97%) and 1.00: 0.17 (1.58%); overall ratio is 1.00: 0.16 (100.00%). Five (5) ratios out of 23, i.e. 0.10, 0.16, 0.12, 0.13, and 0.20, represent 76.22% (Table 8). The three ratios with two entire digits % before the dot represent 59.81% 27.99+16.68+15.14), the nine ratios with one digit represent 37.16% (8.54+ 7.87+4.71+3.85+ 3.51+2.87+ 2.69+1.97+1.15) and the eleven ratios with 0 digit represent 3.00% (0.98+0.47+0.43+0.31+0.23+0.22+0.17+0.07+0.05 +0.04+0.03) (Table 8).

Number of correlated models, based on determination coefficient (r^2), to understand SCCP of combinations. Assumptions: All Sudoku groups are the same no matter what the algorithms were used to build them, therefore they have the same diversity of r.

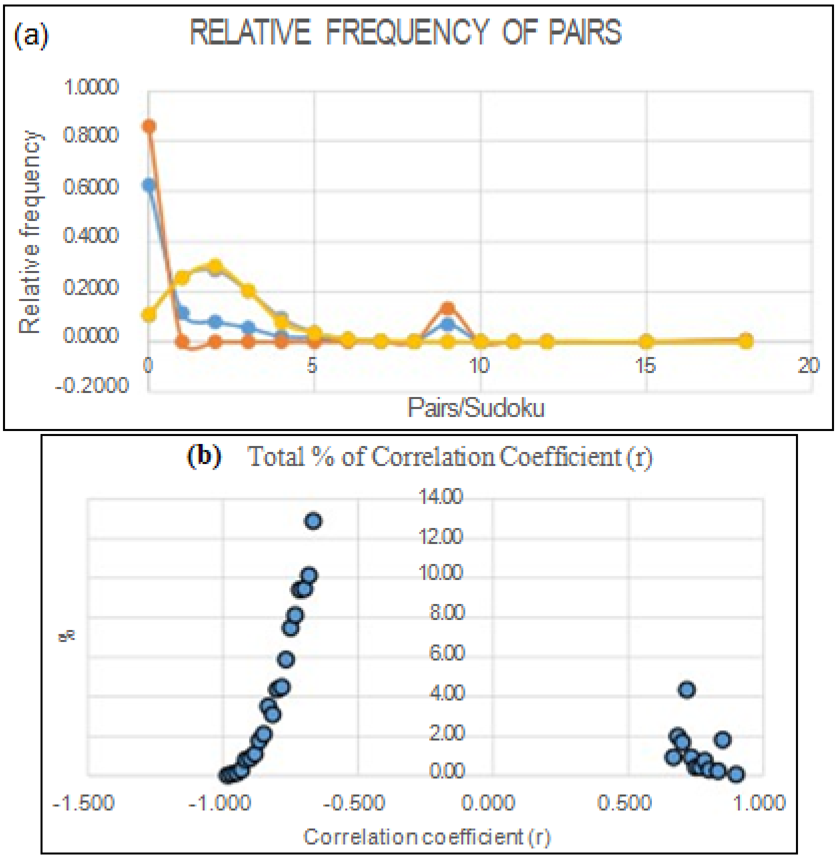


Figure 2: (a) Relative frequency of pairs of statistically significant variables per Sudoku. Red (B=10.0 k), blue (A=10.0 kp), gray (C=6.4k), and orange (D=1.0 k). (b) Total percentage of r where the highest percentages are 12.89% for r=-0.667,

10.14% for r=-0.683, 9.44% for r=-0.700, 9.40% for r=-0.717, 8.12% for r=-0.733, 7.47% for r=-0.750, and 5.89% for r=-0.767. All this represent 63.35%.

In all four groups (A, B, C, D) 32 Correlation were -0.9833 ≤ r ≤ 0.9000 (Table 2; Fig. 5), 20 for negative correlation (-.9833 to -0.6667) correspondent r^2 between 0.445 to 0.967 and 12 for positive correlation (0.667 ≤ r ≤ 0.900) with correspondent r^2 between 0.444 to 0.810 (Figs. 6 and 7). Group A shows 31 correlation, -0.967 to 0.900 but lack -0.983; group B shows 24 correlation models, -0.933 to 0.767 but lack -0.983 to -0.950, 0.767 to 0.833 and 0.900; group C shows all the 32 correlation models, -0.983 to 0.900 and group D shows 30 correlation models, -0.967 to 0.850 but lack -0.983 and 0.900 (Fig. 5; Table 9).

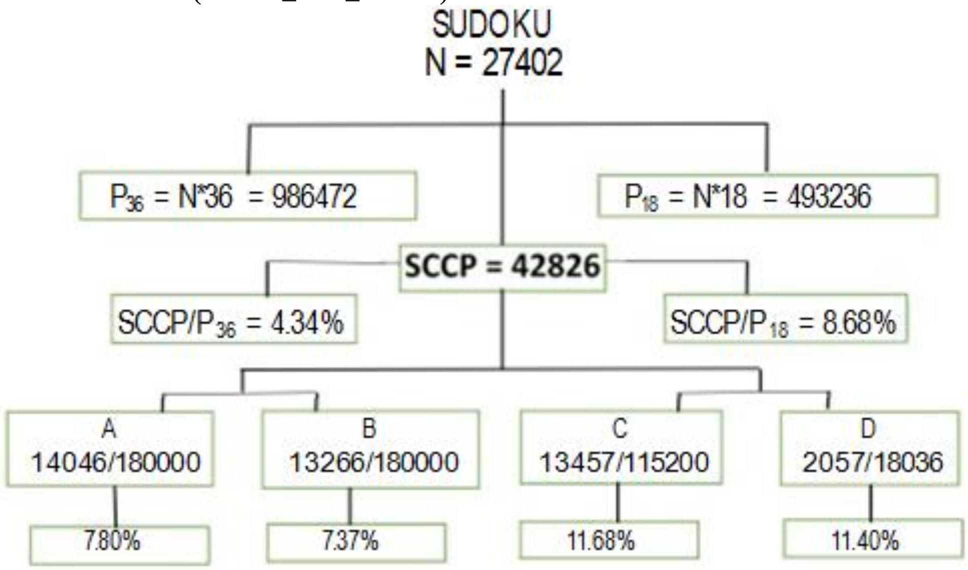
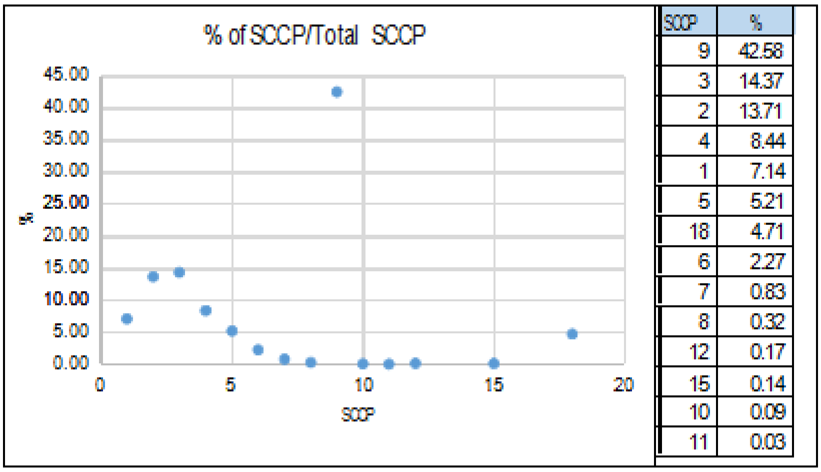


Figure 3: Output of the analysis of regression for 27402 Sudoku where the 42826 SCCP (statistically significant correlation pairs) resulted, representing 4.34% of the total theoretical probabilities (986472) and 8.68% of the calculated probabilities (493236). In the four groups of matrices values representing each group (A, B, C, D) times 18, resulted in the following percentages, 7.80% of 180000, 7.37% of 180000, 11.68% of 115200, and 11.40% of 18036 respectively

Number of groups, within each r^2, through cluster analysis, to understand the nature of the relationships of clusters within the same r^2. Seems like the inclusion of maximum number of correlation pairs of variables within a matrix will depend of the number of Sudoku. The algorithms to build Sudoku may play a relevant role in the number of SCCP correlation between variables. Although A and B (10.0k) were the two groups of matrices where 18 SCCP were found. It most be remarked that A lack 4 out of 18 outcomes, therefore correlated ones represent 37.58% versus the non-correlated which are 62.42%. On the other hand, in the group B, 16 out of 18 outcomes were absent, so the correlated ones are 13.21% versus the non-correlated which are 86.79% (Tables 1, 3, 4, 5). The two smaller matrices, i.e. C and D, with at least one SCCP are 89.44% and 89.02% respectively; only 10.46% and 10.98% lacked SCCP (Tables 1, 3, 6, 7). The number of Sudoku with at least one SCCP were smaller in large groups (37.58% for A and 14.03% for B) than in middle (89.43% for r shows a high level of redundancy (identical scatterplot) because Euclidean distance within the same r = 0. This trimming to redundancy reduces the number of terminal elements between 0.00 and 98.11% (Table 9). Variability within the same r were tested using Spearman correlation. Analysis of distance within the same correlation coefficient



SCCP	%
9	42.58
3	14.37
2	13.71
4	8.44
1	7.14
5	5.21
18	4.71
6	2.27
7	0.83
8	0.32
12	0.17
15	0.14
10	0.09
11	0.03

Figure 4 (g): Percentages of SCCPP out of t the o total SCCP.t Note S that 9 SCCP is w the 4 largest percentage with 42.58%

resulted in: a) an output of 31 dendrograms of group A, where redundancies were removed, lacking between 0.00 and 92.00 elements, and reducing the number of branches between 8.00 and 100.00%, with X±S = 36.81±18.79 and quartiles Q1 Q3 of 32.69 and 43.24 respectively. b) an output of 24 deprogram of group B, where redundancies were removed lacking between 77.78 and 98.11% of elements, reducing number of branches between 1.89 and 22.22%, with X±S = 12.05±4.47, and Q1 and Q3 of 10.44 and 12.67 respectively. c) An output of 32 dendrograms of group C, where redundancies were removed, lacking between 53.49 and 72.41% of elements, and reducing number of branches between 27.59 and 46.51%, with x̶±S = 38.87±3.72, and Q1 and Q3 of 36.78 and 40.95 respectively, and d). An output of 30 dendrograms of group D, where redundancies were removed, lacking between 0.00 and 66.67%, reducing the number of branches between3.33 and 100.00%, with X±S = 74.30±12.32 and Q1 and Q3 of 66.67 and 83.33 (Table 10).

Group B variability was the smallest one among the four groups and it is correlating with the small variability in number of SCCP per Sudoku. In this case, group B only two kinds of SCCP found were 18 and 9.

Comparing shapes of dendrograms resulted from the cluster analysis in 5 correlation coefficients, i.e. -0.933, -0.850, -0.667, 0.667, and 0, 900 to understand trimming of trees, show no correlation between number of cases and

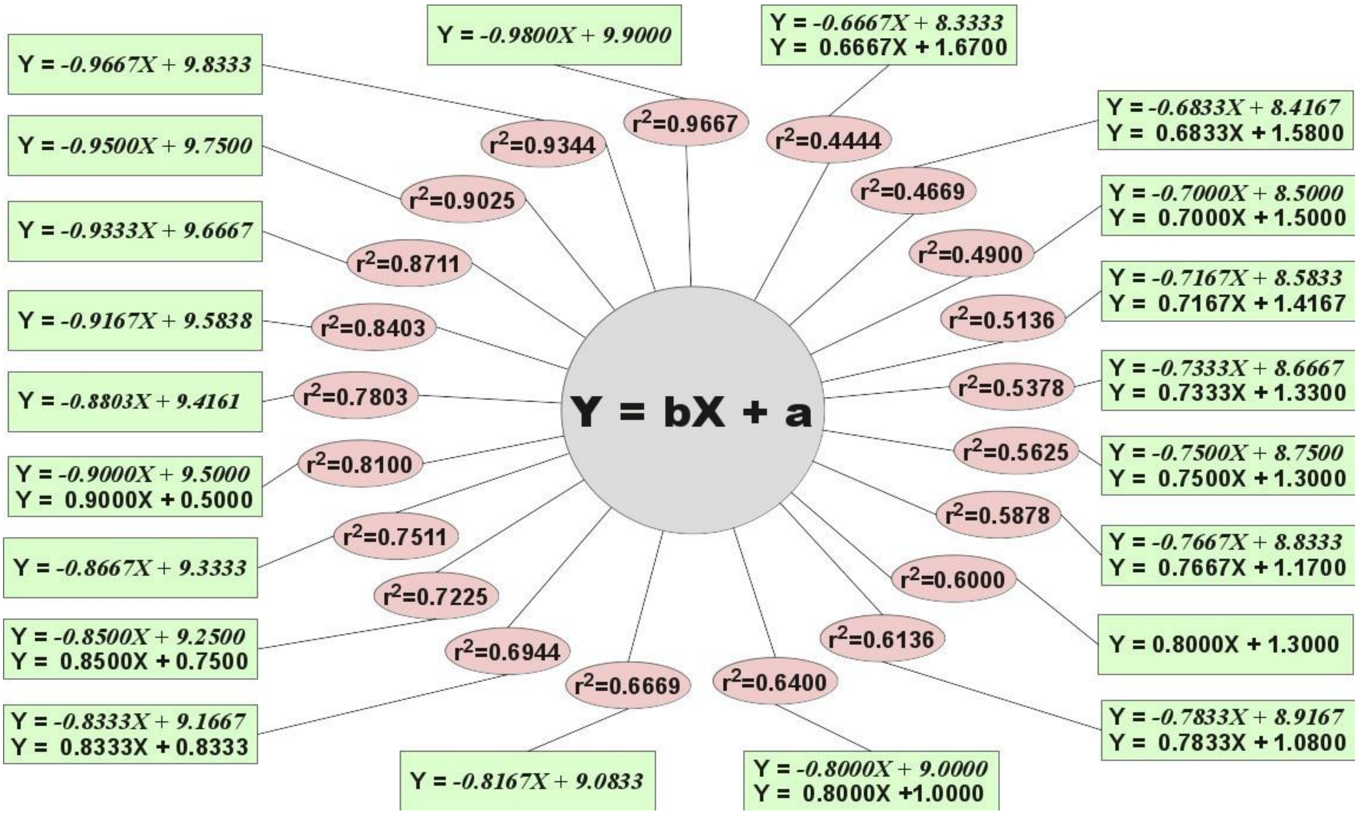


Figure 5: The 32 lineal regression models obtained from the 27402 (A, B, C, D) Sudoku matrices. In the center, general regression equation, with coefficient of determination (r^2) oval

In square boxes, bold equations are positive correlations and italics are negative correlation strimming of the branching diagram. These r were taking base on the occurrence in all four groups of matrices (A, B, C, D).

4. Discussion

Why correlation and regression analysis and why Significance of SCCP per Sudoku? Correlation and regression analyses might help to understand small samples with low value variables, which are unrelated to any specific real situation, that can be associated and how this association might be transformed in a predictive variable f (x) = y, and each variable can be either independent or dependent variable. This situation is considered as special because the rules to build the 9 × 9 matrix, constitute a condition for the relationships between each pair of the 36 pairs of combinations. According to Glass and Hopkins (1996), under the assumptions of equal variance ($\sigma^2_x = \sigma^2_y$) where values distribution

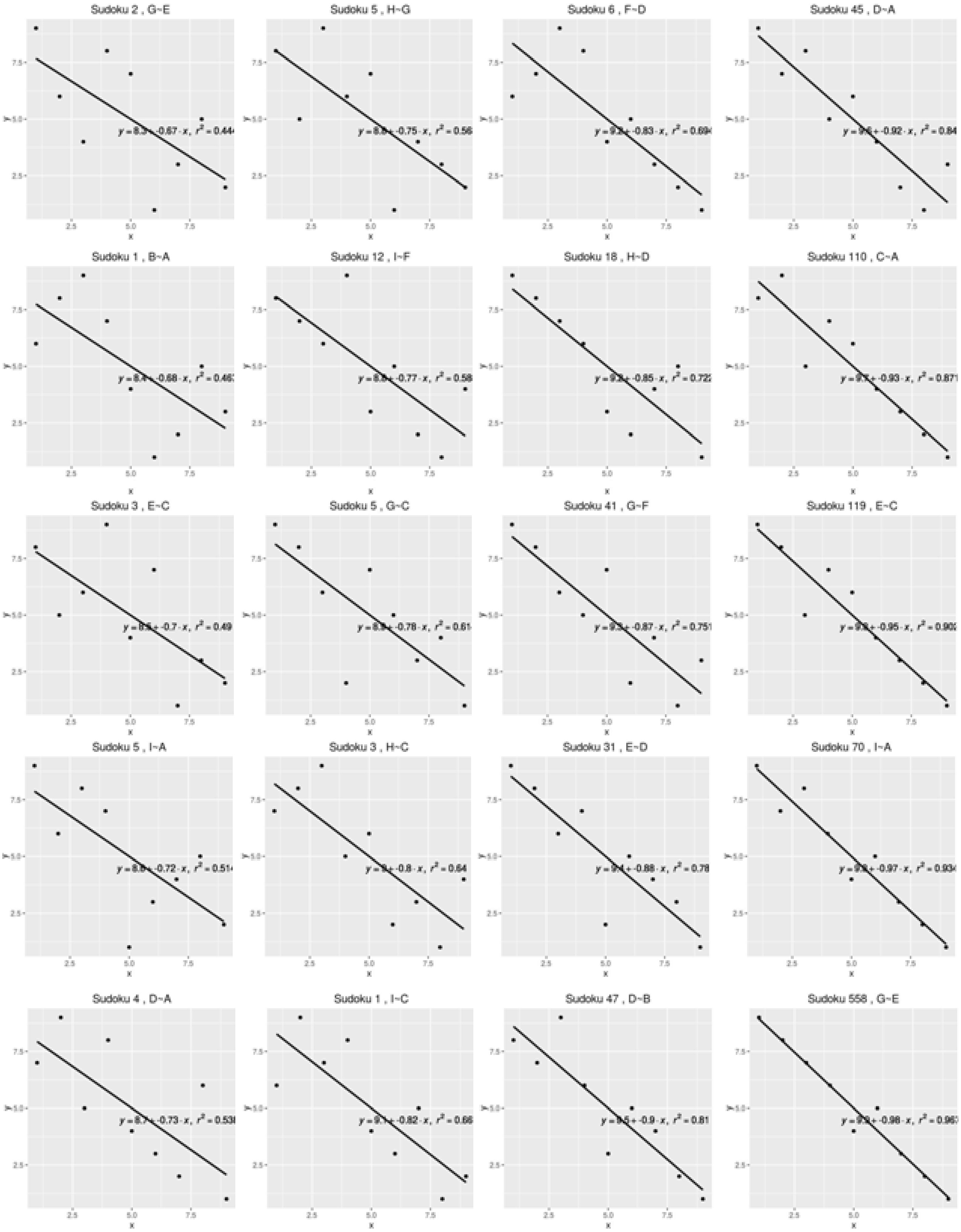


Figure 6: Negative correlation, an example for each r^2

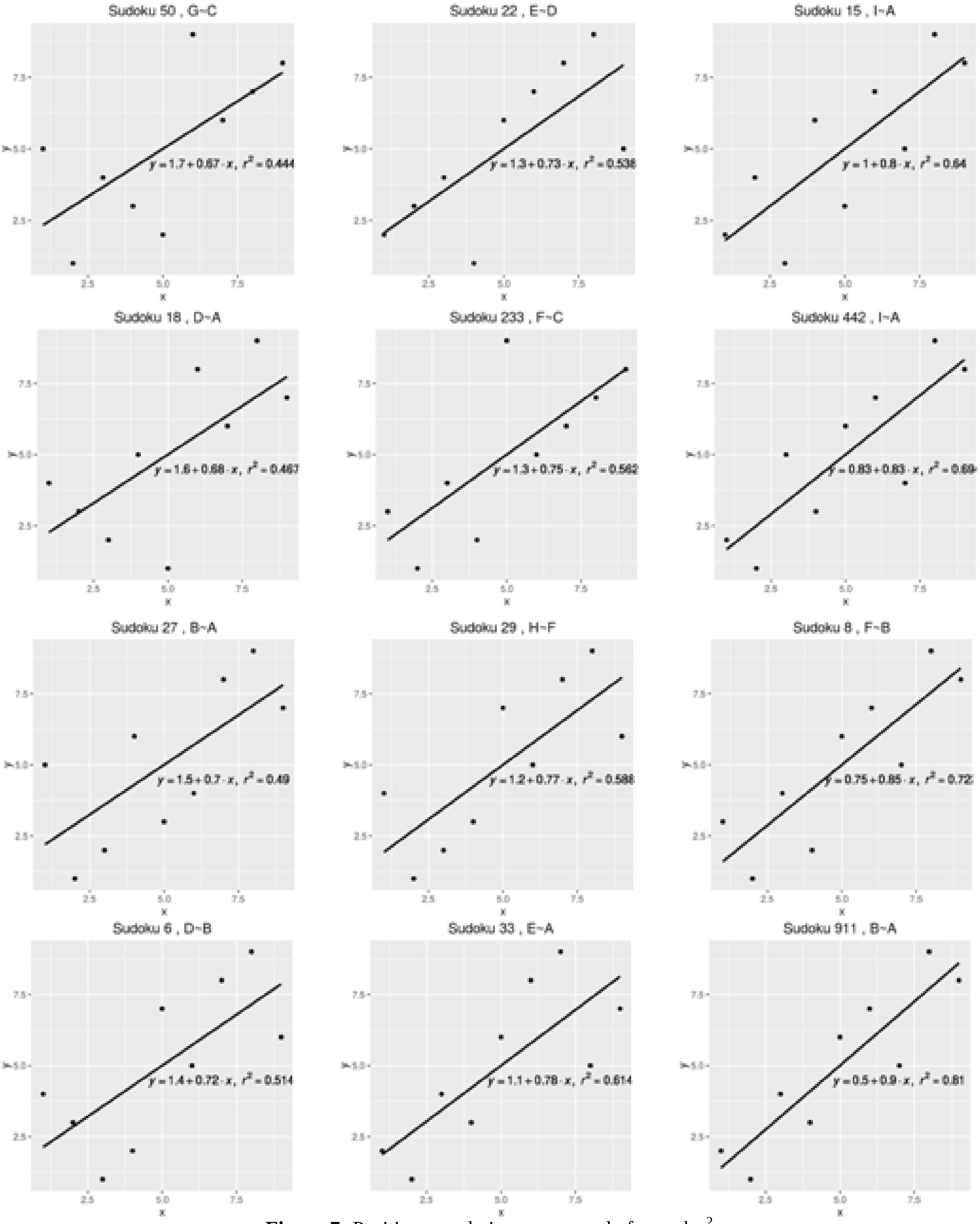


Figure 7: Positive correlation, an example for each r^2

Table 9: Diversity based on the remaining cases (rem%) of correlation coefficient (r), number (orig) of terminal correlations in dendrograms and percentages r remained (rem%) of r after deleting (del%) redundant cases dendrograms.

r	r^2	A			B			C			D		
		orig	del%	rem%	orig	del%	rem%	orig	del%	rem%	orig	del%	rem%
-0.983	0.966	0	0.00	0.00	0	0.00	0.00	5	60.00	40.00	0	0.00	0.00
-0.967	0.935	5	0.00	100	0	0.00	0.00	29	72.41	27.59	3	0.00	100.00
-0.950	0.903	17	47.06	52.94	0	0.00	0.00	36	61.11	38.89	9	66.67	33.33
-0.933	0.870	61	63.93	36.07	9	77.78	22.22	43	53.49	46.51	13	15.38	84.62
-0.917	0.841	142	73.24	26.76	72	88.89	11.11	113	56.64	43.36	18	11.11	88.89
-0.900	0.810	118	69.49	30.51	126	87.3	12.7	116	57.76	42.24	20	10.00	90.00
-0.883	0.780	155	70.32	29.68	99	89.9	10.1	197	61.43	38.58	21	14.29	85.71
-0.867	0.752	224	66.52	33.48	234	89.74	10.26	261	59.39	40.61	40	22.50	77.50
-0.850	0.723	293	68.94	31.06	261	87.74	12.26	303	60	39.93	47	14.89	85.11
-0.833	0.694	541	78.93	21.07	486	87.24	12.67	407	62.41	37.59	70	31.43	68.57
-0.817	0.667	492	71.34	28.66	243	84.36	15.64	526	61.41	38.59	69	20.29	79.71
-0.800	0.640	540	68.89	31.11	603	88.06	11.94	626	63.1	36.9	106	31.13	68.87
-0.783	0.613	580	68.97	31.03	594	88.2	11.78	641	62.4	37.6	108	33.33	66.67
-0.767	0.588	840	71.43	28.57	648	89.51	10.49	909	62.82	37.18	127	29.92	70.08
-0.750	0.563	1016	76.28	23.72	1206	88.89	11.11	873	64.26	35.74	103	33.98	66.02
-0.733	0.537	1197	76.36	23.64	1134	88.54	11.46	994	65.49	34.51	152	36.84	63.16
-0.717	0.514	1292	75.54	24.46	1413	88.96	11.04	1162	65.4	34.6	160	31.87	68.12
-0.700	0.490	1312	74.31	25.69	1269	88.34	11.66	1261	64.55	35.45	205	33.66	66.34
-0.683	0.466	1400	73.14	26.86	1341	89.56	10.44	1355	63.91	36.09	245	39.18	60.68
-0.667	0.445	1682	75.74	24.26	1845	88.29	11.71	1739	66.24	33.76	254	38.98	61.02
0.667	0.445	110	43.64	56.36	54	81.48	18.52	210	59.05	40.95	28	21.43	78.57
0.683	0.466	267	63.3	36.7	216	88.89	11	320	57.5	42.50	54	27.78	72.22
0.700	0.490	228	60.09	39.91	180	91.11	8.89	280	59.29	40.71	39	20.51	79.49
0.717	0.514	837	89.37	10.63	729	95.88	4.12	265	60.00	40.00	34	26.47	73.53
0.733	0.537	143	60.84	39.16	72	86.11	13.89	154	55.19	44.81	29	34.48	65.52
0.750	0.563	80	67.5	32.5	9	77.78	22.22	78	56.41	43.89	14	14.29	86.67
0.767	0.588	56	42.86	57.14	0	0.00	0.00	112	58.93	41.07	15	33.33	73.33
0.783	0.613	108	50	50	0	0.00	0.00	188	61.70	38.30	30	26.67	73.33
0.800	0.640	37	56.76	43.24	0	0.00	0.00	75	61.33	38.67	14	21.43	78.57
0.833	0.694	18	22.22	77.78	0	0.00	0.00	71	57.75	42.25	10	20.00	80.00
0.850	0.723	250	92	8	423	98.11	1.89	87	63.22	36.78	18	16.67	83.33
0.900	0.810	5	40	60	0	0.00	0.00	21	61.9	38.10	0	0.00	0.00

Table 10: Descriptive statistics of reduction (%) of terminal elements (SCCP) in dendrograms by elimination of redundancy

Group	n	Min-Max (%)	X±S	Q1-Q3
A	31	8.00-100.00	36.81±18.79	32.69-43.24
B	24	1.89-22.22	12.05±4.47	10.44-12.67
C	32	27.59-46.51	38.87±3.72	36.78-40.95
D	30	33.33-100.00	74.30±12.32	66.67-83.33

is quite similar, the maximum value of r (= 1.00) is achievable if the shape of the variables x and y are the same (Goodwin and Leech, 2006). However, we learned that variables base in Sudoku's matrix does not reach a perfect r, i.e. -1.00 or 1.00. The maximum r value found for SCCP within a matrix Sudoku was r = −0.983 (Fig. 6; Table 9) and this might be explained because for negative correlation, the smallest integers can perfectly match with the larger ones but the two in the centre do not match for perfect correlation, r = −1.00 e.g. (9, 1), (8, 2), (7, 3), (6, 4), (5, 6), (4, 5), (3, 7), (2, 8), (1, 9) as can be seen in scatter plot Fig 6 (r^2 = 0.963); the meaning of the term centre is the distribution of each pair. For the positive ones, it is more complex to allow a perfect r = 1.00 because, given that in Sudoku rules, no pair integer can be the same in any of the nonet columns, e.g. (9, 8), (8, 9), (7, 6), (6, 7), (5, 4), (4, 5), (3, 1), (2, 3), (1, 2) as can be seen in scatter plot of Fig 7 (r^2 = 0.81) for maximum value found. This situation helps to explain why the negative correlation is a lot more frequent. Also, correlation near the maximum found in both, negative and positive, is almost impossible in closely neighbour variables, i.e. members of three nonet columns that belong to three nonet boxes, is not allowed because Sudoku rules. However other kinds of regression such as polynomial model might be performed. In this paper, general negative: positive ratios was 1.00: 0.16, very similar values as in algorithms A, B, C and D, with 1.00: 0.18; 1.00: 0.15; 1.00: 0.16 and 1.00: 0.16 respectively (Table 3).

Regarding SCCP per Sudoku, 70.66% belong to three SCCP, i.e. 9 (42.58%), 3 (14.37%) and 2 (13.71%); the following five (SCCP 4, 1, 5, 18 and 6) represents 27.98%, and the rest of them represents 1.58% (Fig.4). The smallest group, D, lack SCCP 18 to 9; Group C, lacks SCCP 18 to 10, however group B lack almost all of them but SCCP 9 and 18. Group A, which only lack SCCP 17-16 and 14-13, might represent the closure algorithm to the reality, but performs made until now do not support this hypothesis. We might interpret that in any algorithm to build Sudoku, regression analysis performed to the 9 × 9 matrix, the set of matrices should contain 9 SCCP in high proportion.

Regarding the probability of SCCP per Sudoku, no output is small enough when a huge number of systems are considered, assuming each Sudoku as a system with 81 grids (9x9). As we learned, the maximum number of Sudoku are

6.67 x10^{21} (sensu Felgenhauer and Jarvis, 2005) but symmetries analysis reduces this number to 5.47x10^{9} (Russell and Jarvis, 2006). In any way, finding 71 (0.0071) out of B (10.0k) and 41 (0.0041) out of A (10.0k), with 18 significant correlated SCCP, because 7.1x10^{-3} x 5.47x10^{9} = 38.837x10^{6}, and analysing the original amount ofFelgenhauer and Jarvis (2005) it will be 7.1x10^{-3} x 6.67x10^{21} = 47.357x10^{18}. None of these numbers is small, so the relative proportions calculated in our sample are statistically relevant in order to estimate the population parameters. As physicists do similar things with their calculus about number of galaxies in the universe, e.g. Christopher Conselice from University of Nottingham estimated the number of galaxies in the universe as 2.0x10^{12}, based on data from the Deep Sky Hubble Space Telescope.

Significance of SCCP per group of algorithm and Clustering group of algorithms. The minimum number and less diverse SCCP found here might be related to the algorithm to build Sudoku grids. Cluster analysis to establish how the four groups cluster using a ratio of negative and positive correlation, shows that the more distantly related group was B, i.e. (((D C) A) B), as in the dendrogram shown on Fig. 8.

Application of cluster analysis to the same correlation coefficient to detect redundancy shows that overall patterns of redundancy of terminal elements in the dendrogram differ among each group. The larger redundancy belongs to group B where the remaining terminal groups, after trimming dendrograms, are between 1.89 and 22.00%; 17 out of the 24 reductions between 11.11 and 13.89% with mode of 11.00-11.94% and (X) S of 12.05 4.47. The second larger redundancy belongs to Group C, where the remaining terminal groups after trimming dendrograms are between 27.59 and 46.51%; 18 out of the 32 reductions between 33.63 and 39.93%, this is also the mode and (X) S of 38.87, 3.72% ; and Q1 to Q3 of 36.78 to 40.95% respectively (Fig. 10). These results raise the questions: Is this a consequence of the algorithm? It is because only two kinds of cases were found? (Table 10).

As might be expected, A is the most widespread and B is the narrowest in terms of the % diversity because the redundancy deleted. Is the sample size good enough to represent the universe of both scenarios, i.e. 5.47x$10^{9}$and 6.67x10^{21}? In terms of the sample, yes, but the algorithm might introduce some bias for data analysis therefore other studies will go deeper into this problem to use the results for probabilities and to test hypotheses of chi square. Are the algorithms to build groups good enough to put together and then apply chi-square?

One of the factors affecting correlation is variability of X and Y variables (sensu Goodwin and Leech, 2006) and here the amount of variability has been test using cluster analysis that might help to understand the diversity within the same group of r. A test like this is unusual because not often modelling correlation allows the amount of r therefore scatterplot help to understand how different they are.

Diversity within the same r and relevance of scatter plot. Once the relationships is known, i.e. 20 negatives and 12 positives correlation coefficient, several of the 42826 SCCP has the same r, with the following distribution: a) group A, in 14046 SCCP, 1682 (11.97%) belong to r = -0.667; 1400 (9.97%) belong to r = −0.683; 1312 (9.34%) belong to r = −0.700, (Table 8, Fig.5).

b) group B, in 13266 SCCP, 1845 (13.91%) correspond to r = −0.667; 1413 (10.65%) belong to r = −0.717; 1341 (10.11%) belong to r = −0.683; 1269 (9.57%) belong to r = −0.700; 1206 (9.09%) belong to r = −0.750; 1134 (8.45%) with r = −0.7333. The remaining ones with less than one thousand, six of them with less than 100 or 0.75%. c) group C, in 13457 SCCP, 1739 (12.92%) belong to r = −0.667; 1355 (10.07%) belong to r = −0.683; 1261 (9.37%) correspond to r = −0.700; 1162 (8.63%) belong to r = −0.717. The remaining ones with less than 1000 (7.43%), 9 of them with less than 100 (0.74%). d) group D, in 2057 SCCP, 254 (12.35%) belong to r = −0.667; 245 (11.91%) correspond to r = −0.683; 205 (9.96%) belong to r = −0.700; 160 (7.78%) belong to r = −0.717; 152 (7.39%) belong to r = −0.733; 127 (6.17%) belong to r = −0.767; 108 (5.25%) belong to r = −0.783; 106 (5.15%) belong to r = −0.800 (Table 9).

Sometimes, it is also difficult to judge whether a correlation measure is “high” or “low” sensu Cohen (1977 and 1988). For behavioural science, Cohen (1977, 1988) classified correlation coefficient in three categories, small (r=0.10), medium (r =0.30) and large (r=0.50) (Cohen, 1992). on the other hand, Hebel and McCarter (2012) considered five categories, i.e. negligible (0.0 to 0.2), weak (0.2 to 0.5), moderate (0.5 to 0.8) and strong (0.8 to 1.0) (Looney, 2018). However, there are certain situations where a correlation measure of 0.3, for example, may be consider negligible. Looney (2018) paid attention to the sample size where his minimum value of r to yield p smaller of equal to 0.05 in a sample n=10 is r =0.632; he uses a range of n between 10 and 200 (his table 3). Our sample size, n=9, is constant without any choice. However, we did have a choice in the number of matrices, each of which allows 36 correlation pairs where the minimum r =0.667 (SCCP).

This area of research is critical to decide whether or not r is negligible. The diversity of criteria to decide whether or not an index is significant for new knowledge to support a hypothesis, the correlation categories discussed in here suggest that further examination is needed. As with all data analysis, context of the data must be understood in order to evaluate any results (Stockwell, 2008). King (2013) realised to test the hypotheses of significance if the value of r is low and certain uncertainty exist, to find whether or not r arising by chance alone or if the relationship can apply to the whole population. Repeating, thousand of times, the analysis with exactly the same sample size per variable support Goodwin and Goodwin (1999) affirmation about the common misconception assuming a direct relationship between the size of N and the size of r (Goodwin and Leech, 2006). However, Hinkle et al. (1988) realise that the size of the samples might affect the stability and accuracy of the results. Looney (2018) main argument is based on size of the sample for a strong regression model. Because non null hypothesis might be appropriate when you have CI so wide to provide very little useful information regarding magnitude of the population regression; as a consequence of this, Looney (2018) proposed two alternative approach: a) testing null values other than po = 0 and b) determining the sample size so as to achieve a certain level of precision of the estimate of p, as measured by the width of the resulting C.I. (Looney, 2018). In this case, further analysis like this should have more in depth discussion about the effect of the rules of the Sudoku and the size of the sample, considering that each variable always will have 9 integers, restricted by Sudoku rules. The reasoning above makes matrix Sudoku ideal for simulation and modelling, to use in bivariate and multivariate analysis such as regression, cluster, principal component and functional discriminant analyses. Testing the relationships of the same r in a group of matrices where simultaneous regression analysis was performed, a multivariate analysis was applied to bi-variate analysis results as the basis for simulations and models.

It is a big opportunity to learn whether it is possible to feel confident with the same r as the same model. Given the pattern to build the matrices, all matrices are different but keep the same properties and rules; however, it does not prevent exactly the same shape in several SCCP from different matrices. Scatter plot accuracy to compare models were tested using cluster analysis.

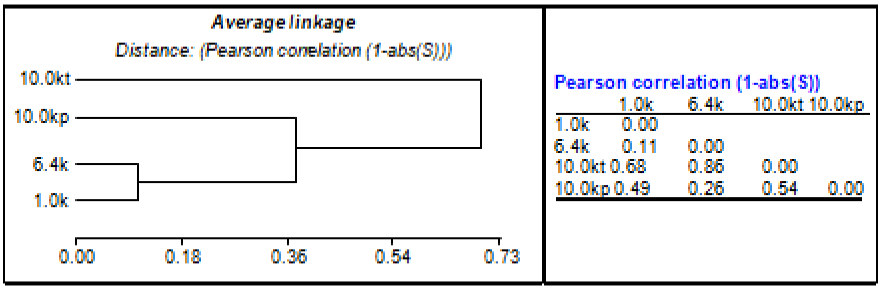
In 1842 resulted with same r the ones in the same dendrogram that present nodes with zero distance, interpreted as no differences in the model.

However, the percentage that remains with distances different than 0 is interpreted as relevant for diversity of scatter plot because, after trimming the ones with 0 distance, between 0.00 and 98.2% of the values or r remains, this is interpreted as a lack of redundancy shows a wide range of variation. Trimming was applied to eliminate redundancy, to understand the diversity of branches for a given group. Using thousands of matrices as the sample, where the maximum number of r values without repetition is 36, the same r can be found in several scatter plots across the sample. The challenge in this case is to establish an association through cluster analysis, using Pearson's r of a group of elements to understand variability of r as well as the level of redundancy. To show the distances equal zero, which reduces considerably the diversity within the same simulations through the elimination of the terminal elements in the same branch of a dendrograms with distance equal zero.

Modelling is a common tool in areas such as neuroscience (Nombela et al., 2011), chemistry education (Pérez and Lamoureux, 2007), climate (), several biological disciplines (Pianka, 1973;Pielou, 1981;Sneath and Sokal, 1962) such as Ecology and population genetic and phenetics, economy (), health (). A tendency to organize information to understand systems, either natural or artificial, is a human attribute used to benefit education (Crute and Myers, 2007), health (Bhattacharyya et al., 2014;Nombela et al., 2011), agriculture (Dan-baba and Dauran, 2016;Danbaba, 2016; Hui-Dong and Ru-Gen, 2008; Shehu and Danbaba, 2018), technology (), and economy () among others. Puzzles, e.g. Sudoku, are models created for entertainment to develop skills for logic and mathematics.

This paper method does not respond to any of the four models (I-IV) discussed by Hui-Dong and Ru-Gen (2008), Danbaba (2016a, b) and Shehu and Danbaba (2018) who used Sudoku as matrix for experimental design for Latin square. They used 3x3 block column row for the different treatments in agriculture experiments and did ANOVA, ANCOVA and regression multiple regression) that is different to the models used in this paper.Cluster analysis of the scatter plot found, considered by Asuero et al. (2006) the first step in all data analysis, suggest that fusion of the terminal branches within a dendrogram might be between 0.00 and 100.00%, after redundancy are deleted (Fig. 9). Cook and Weisberg (2009) realised that scatter plot give a lot more information than the correlation coefficient (Asuero et al., 2006). To understand patterns cluster analysis using Euclidean distance is more accurate than scatter plot.

Asuero et al. (2006) realised that, as a summary of data, scatter plot matrix can be better than a correlation matrix. However, having several (thousands) identical correlation coefficient, the addition of cluster analysis to build dendrograms, where terminal elements within the same terminal branch having zero distance, i.e. one hundred % similarity, are fused to obtain the actual diversity of scatter plot within that r, helps a lot to understand the variability within the same r, e.g. the 5520 cases of r = -0.667, representing an overall frequency of 12.89%, allow to test level of redundancy to answer the question whether or not the same r represent the same scatter plot. We call this trimming the branching diagram tree, trimming the dendrogram. The branches of the dendrogram with distance equal zero is interpreted as variables X and Y having exactly the same shape, as has



Pearson correlation (1-abs(S))

	1.0k	6.4k	10.0kt	10.0kp
1.0k	0.00			
6.4k	0.11	0.00		
10.0kt	0.68	0.86	0.00	
10.0kp	0.49	0.26	0.54	0.00

Figure 8: Dendrogram with Euclidean distances, based on the negative: positive correlation ratio for all groups of Sudoku (A, B, C, D) from data in Table 10)

been proposed by Glass and Hopkins (1996). This might be test of identical distribution of each pair of data in variables X and Y. Redundancy in nature is relatively common, e.g. DNAs nucleotide, with four nitrogen bases on three of this aminoacids coding an aminoacid represented in the RNA in three nucleotide, one of them different, with maximum probability of 64 combinations (4^3); intron (Gilbert, 1978) is also functionally monotonous. This has been use in molecular biology in cases of gene code, also to compare sequences for phylogenetic analysis. To build branching diagrams for overall similarity to establish distance relationships. Although phylogenetic analysis, where it is necessary to polarise data to establish as character, the analogy here is to interpret the branching diagrams.

As far as we know, this is the first paper in which correlation and regression analysis are apply to Sudoku as matrices. The advantage of this is the uniformity of the matrices, all of them with the same structure, a kind of fractal, shows the same site to establish relationships among elements of each matrix and very relevant for modelling. Polynomial analysis performed in the nonet column within the same nonet box might show SCCP higher than any lineal model but is not part of the objective of this paper. This analysis might does not represent anything because any Sudoku performed can be transformed in easy, middle or very hard to solve. The only thing that is needed to do a transformation from easy to

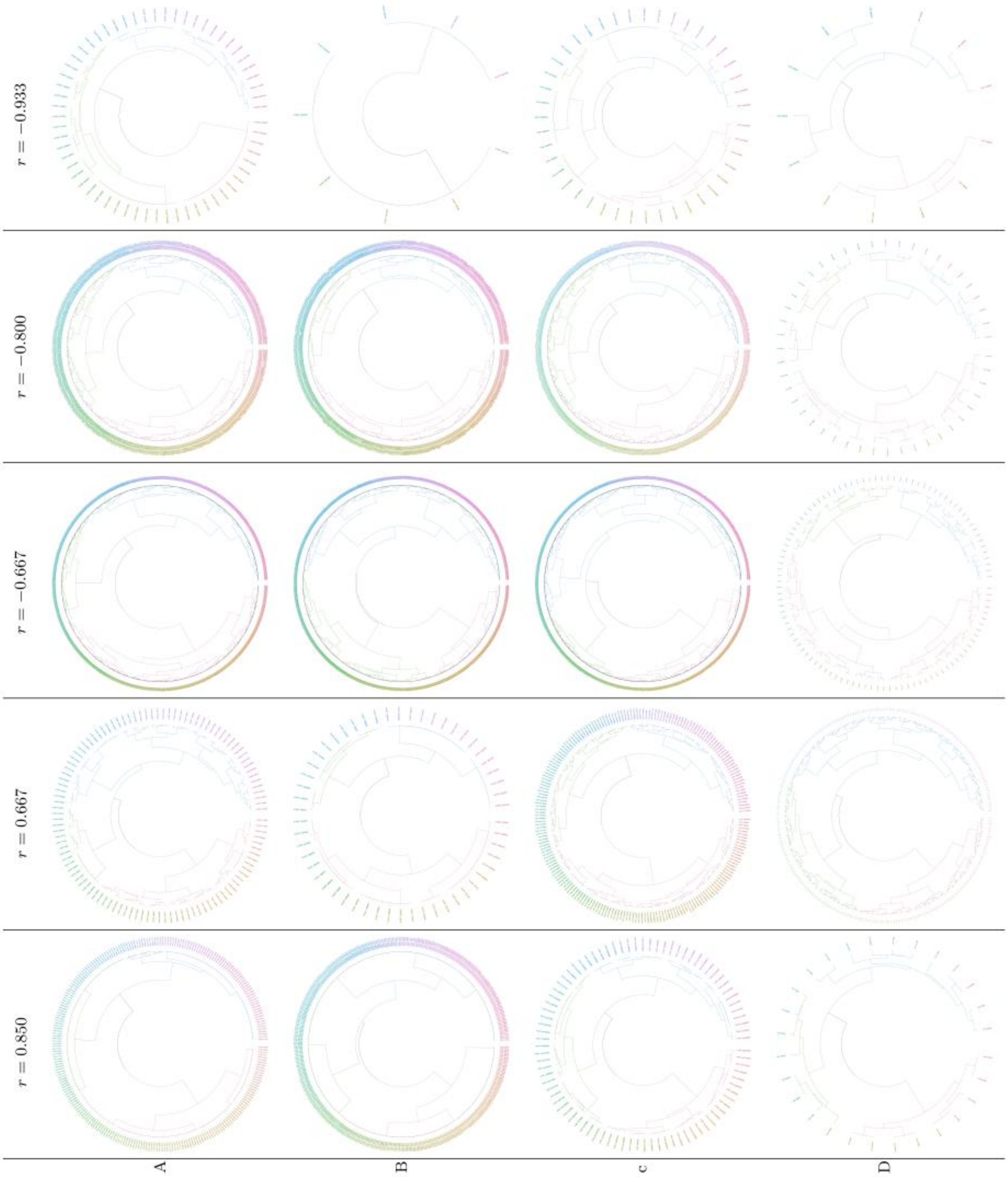


Figure 9: Selected dendrograms to compare redundancy (zero distance within terminal elements) by Group (A, B, C, D). For -0.933 terminal lines: A) south east and south terminal lines show high level of redundancy; B) the two branches shows high level of redundancy, therefore the nine elements will fuse in two; C) all branches present high level of redundancy but lower

in proportion to previous two, i.e. A and B, and D) low level of redundancy. Column -0.800 values are mentions as -0.800A to -0.800D, and so forth.

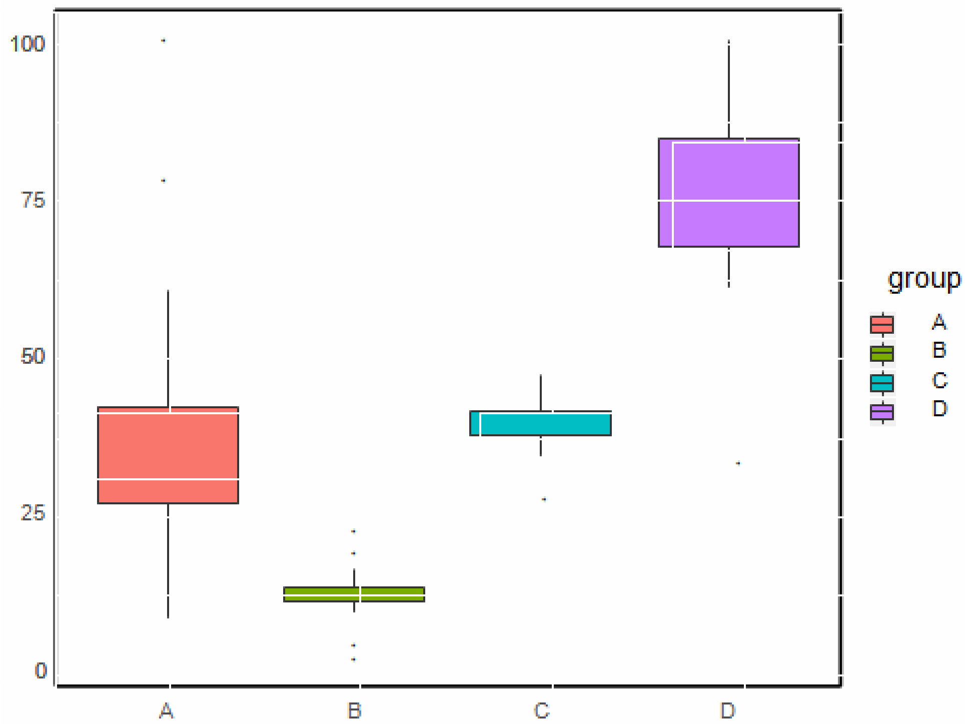


Figure 10: Percentage of remaining's SCCP by group after pruning the dendrograms

hard to solve is the distribution of the filled grid in the puzzle, therefore, what is relevant is the matrix to build up.

5. Conclusions

In conclusion, the 42826 SCCP belong to 32 regression models from the 27402 Sudoku grids. Here we hypothesise that 32 models of regression is the maximum number of models (or very close to it) to be found in Felgenhauer and Jarvis (2005) and Russell and Jarvis (2006) maximum number of Sudoku. Another hypothesis is that negative: positive correlation ratio reported here is very close to the real in all possible Sudoku. Overall negative: positive correlation ratios were 1.00: 0.16 and the explanation is that Sudoku favoured negative correlation because larger and smaller pairs integer can perfectly match but it is impossible to perfectly match larger integers. None of the algorithms satisfy having all SCCP per Sudoku, the most inclusive one is the one use for Group A. However, the one for Group B is the less inclusive which lack all the SCCP but 18 and 9. Level of association of r for each of the two variables and scatter plot help when measuring limited number correlation and regression but not if there are huge amount in a complex system like Sudoku matrices where cluster analysis provide wider overview because redundancy within the same r can be reduce trimming the dendrogram and then reflex the real diversity of r and show a reduction between 0.00 and 98.20%.

6. Acknowledgements

Authors wish to thank the late Andrés Guerrero (1952-2014), the first scientist with whom the project was discussed and with whom a podium talk was presented in VIII International Scientific Research Congress (2012) in the Dominican Republic. Diógenes Aybar, for the vivid discussions and advise at the beginning of the study; John Henry Morales, help with the algorithms to build Sudoku matrices D; Ángel Pimentel, for helping the first author testing several times the kinds of analysis to do. Special thank to América Sánchez, Katiusca Rodríguez, Evelyn Gavot, and Candy Ramírez the students with whom the first author start testing hypotheses using Sudoku as matrices. Thank to Edward Blanco, Santiago W. Bueno López, Teresita González, Kiero Guerra Peña, and Leandra Tapia for the review and advises to strengthen the paper and the two anonymous reviewers for their critical review and advice to improve this paper.

References

[1] Asuero A, Sayago A, Gonzalez A (2006) The correlation coefficient: An overview. Critical reviews in analytical chemistry 36 (1): 41-59
[2] Bailey R, Cameron PJ, Connelly R (2008) Sudoku, gerechte designs, resolutions, affine space, spreads, reguli, and hamming codes. The American MathematicalMonthly 115 (5): 383-404
[3] Becerra Tomé A, Núñez Valdés J, ¿Perea González JM (2016) Cuánta Matemática hay en los sudokus? Pensamiento Matemático 6 (1): 113-136
[4] Behrens W (1956) Feldversuchsanordnungen mit verbessertem ausgleich der bodenunterschiede. Zeitschriftfür Landwirtschaftliches Versuchsund Untersuchungswesen 2: 176-193.
[5] Bhattacharyya S, Cai X, Klein J (2014) Dyscalculia, dysgraphia, and left-right confusion from a left posterior peri-insular infarct. Behavioural Neurology 2014

[6] Brahm D, Snow G, Brahm MD (2009) Package 'sudoku'

[7] Brahm D, Snow G, with contributions from Curt Seeliger, Bengtsson H (2014) sudoku: Sudoku Puzzle Generator and Solver. URLhttps: //CRAN. R-project.org/package=sudoku, r package version 2.6

[8] Brophy C, Hahn L (2014) Engaging students in a large lecture: An experiment using sudoku puzzles. Journal of Statistics Education 22 (1)

[9] Chambers J, Hastie T (1992) Statistical Models in S. Wadsworth & Brooks/Cole computer science series, Wadsworth & Brooks/Cole Advanced Books & Software

[10] Cook RD, Weisberg S (2009) An introduction to regression graphics, vol 405. John Wiley & Sons

[11] Cornell University Department of Mathematics (2009) The math behind sudoku. URLhttp: //pi.math.cornell.edu/~mec/Summer2009/Mahmood/ Intro.html, [Online; accessed 30-July-2018]

[12] Crute TD, Myers SA (2007) Sudoku puzzles as chemistry learning tools. Journal of Chemical Education 84 (4): 612

[13] Dahl G (2009) Permutation matrices related to sudoku. Linear Algebra and its Applications 430 (8-9): 2457-2463

[14] Danbaba A (2016) Construction and analysis of samurai sudoku. World Academy of Science, Engineering and Technology, International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering 10 (4): 165-170

[15] Danbaba A, Dauran N (2016) Construction and analysis of partially balanced sudoku design of prime order. International Journal of Statistics and Applications 6 (5): 325-327

[16] Dattorro J (2005, v2011.04.25) Convex Optimization & Euclidean Distance Geometry. Meboo Publishing

[17] De Ruiter J (2010) On jigsaw sudoku puzzles and related topics. Tech. rep., Universiteit Leiden, Opleiding Informatica

[18] Farris J (2011) Sudoku and math-are they related? URLhttp: //www.associatedcontent.com/article/43100/sudoku_and_math_ are_they_related.html?cat=2, [Online; accessed 1-November-2011]

[19] Felgenhauer B, Jarvis F (2005) Enumerating possible sudoku grids. Preprint available at http: //wwwafjarvis staff shef ac uk/sudoku/sudoku pdf

[20] Floyd J (2011) Wang and wittgenstein. Hao Wang, logician and philosopher, Texts in Philosophy pp 145-194

[21] Gan Z, Hua Y, Zhang M, Lai C, Gan MZ (2012) Package 'sudokuplus'

[22] Geem ZW (2007) Harmony search algorithm for solving sudoku. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, pp 371-378

[23] Gilbert W (1978) Why genes in pieces? Nature 271 (5645): 501

[24] Glass GV, Hopkins KD (1996) Statistical methods in education and psychology. Psyccritiques 41 (12): 1224

[25] Goodwin LD, Goodwin WL (1999) Measurement myths and misconceptions. School Psychology Quarterly 14 (4): 408

[26] Goodwin LD, Leech NL (2006) Understanding correlation: Factors that affect the size of r. The Journal of Experimental Education 74 (3): 249-266

[27] Gould W (2007) Wayne gould puzzles. URLhttp: //waynegouldpuzzles.com/sudoku/features/, [Online; accessed 30-July-2018]

[28] Grossman L (2013) "the answer men". URLhttp: //www.time.com/time/ magazine/article/0, 9171, 2137423, 00.html

[29] Hinkle DE, Wiersma W, Jurs SG, et al. (1988) Applied statistics for the behavioral sciences. Houghton Mifflin Boston

[30] Hui-Dong M, Ru-Gen X (2008) Sudoku squarea new design in field. Acta AgronomicaSinica 34 (9): 1489-1493

[31] Intelm H (2005) How to solve every sudoku puzzle, vol. 2

[32] Johnson-Laird PN (2010) Mental models and human reasoning. Proceedings of the National Academy of Sciences 107 (43): 18243-18250

[33] King M (2013) Fisheries biology, assessment and management. John Wiley& sons

[34] Kuhn TS (1963) The structure of scientific revolutions, vol 2. University of Chicago press Chicago

[35] Kwan CC (2010) Spreadsheet-based sudoku as a tool for teaching logical deduction. Spreadsheets in Education (eJSiE) 4 (1): 3

[36] Lee Rodgers J, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. The American Statistician 42 (1): 59-66

[37] Liao Gz, Shih Yj (2013) Between sudoku rules and labyrinthine paths-a study on design for creative sudoku learning. Designs for Learning 6

[38] Louis Lee N, Goodwin GP, Johnson-Laird P (2008) The psychological puzzle of sudoku. Thinking & Reasoning 14 (4): 342-364

[39] MacWilliams FJ, Sloane NJA (1977) The theory of error-correcting codes. Elsevier

[40] Maji AK, Jana S, Pal RK (2013) An algorithm for generating only desired permutations for solving sudoku puzzle. Procedia Technology 10: 392-399

[41] Mandal SN, Sadhu S (2013) Solution and level identification of sudoku using harmony search. International Journal of Modern Education and Computer Science 5 (3): 49

[42] McGerty S, Moisiadis F (2014) Are evolutionary algorithms required to solve sudoku problems? Computer Science & Information Technology

[43] Murtagh F, Legendre P (2014) Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? Journal of classification 31 (3): 274-295

[44] Newton PK, DeSalvo SA (2010) The shannon entropy of sudoku matrices. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, The Royal Society, vol 466, pp 1957-1975

[45] Nombela C, Bustillo PJ, Castell P, Medina V, Herrero MT (2011) Cognitive rehabilitation in parkinsons disease: evidence from neuroimaging. Frontiers in neurology 2: 82

[46] Pelánek R (2011) Difficulty rating of sudoku puzzles by a computational model. In: FLAIRS Conference

[47] Pérez AL, Lamoureux G (2007) Sudoku puzzles for first-year organic chemistry students. Journal of Chemical Education 84 (4): 614

[48] Pfaffmann JO, Collins WJ (2007) Teaching artificial intelligence across the computer science curriculum using sudoku as a problem domain. In: FLAIRS Conference, pp 327-332

[49] Pianka ER (1973) The structure of lizard communities. Annual review of ecology and systematics 4 (1): 53-74

[50] Pielou EC (1981) The usefulness of ecological models: a stock-taking. The Quarterly Review of Biology 56 (1): 17-31

[51] Pillay N (2012) Finding solutions to sudoku puzzles using human intuitive heuristics. South African Computer Journal 49 (1): 25-34

[52] R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https: //www.R-project.org/

[53] Rosenhouse J, Taalman L (2011) Taking sudoku seriously: The math behind the world’s most popular pencil puzzle. OUP USA

[54] Rouse Ball W, Coxeter H (1987) Mathematics recreation and essay

[55] Russell E, Jarvis F (2006) Mathematics of sudoku ii. Mathematical Spectrum 39 (2): 54-58

[56] Sabrin AP (2009) Multimedia application for solving a sudoku game. arXivpreprint arXiv: 09054203

[57] Shehu A, Danbaba A (2018) Variance components of models of sudoku square design. Annals Computer Science Series 16 (1)

[58] Sneath PH, Sokal RR (1962) Numerical taxonomy. Nature 193 (4818): 855-860 Stockwell I (2008) Introduction to correlation and regression analysis. In: Statistics and Data Analysis. SAS Global Forum

[59] Tengah KA (2011) Using simplified sudoku to promote and improve pattern discovery skills among school children. Journal of Mathematics Education at Teachers College 2 (1)

[60] Ward Jr JH (1963) Hierarchical grouping to optimize an objective function. Journal of the American statistical association 58 (301): 236-244

[61] Weyland D (2015) A critical analysis of the harmony search algorithmhow not to solve sudoku. Operations Research Perspectives 2: 97-105

[62] Williams TG (2011) Algebraic Sudoku Bk 2: A Fun Way to Develop, Enhance, and Review Students Algebraic Skills. Milliken Publishing Company