Visual Variations between Pairs of SARS-CoV-2 Genomes on Integrated Density Matrix

Minghan Zhu
Yunnan University
Jeffrey Zheng (conjugatelogic@yahoo.com)
Yunnan University

Research Article

Keywords: SARS-CoV-2, Genomic Sequence, COVID-19, Genetic Variation, Bose Einstein distribution

Posted Date: January 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-72020/v2

Visual Variations between Pairs of SARS-CoV-2 Genomes on Integrated Density Matrix

Minghan Zhu, Jeffrey Zheng

Abstract This paper is the B2 module of the MAS. The quantification matrix is formed according to the four-base arrangement in the genome sequence. The differences in new coronavirus genome sequencing sequences in different samples were demonstrated by using the most concise methods. Using 4 primitive variable value measures, changes in the virus genome sequence base order conditions were determined. When two relatively large genomic sequences are slightly different, the integrated distribution of the difference calculation is subtly similar to the Bose-einstein distribution, while the sum calculation shows a powerful distribution complexity. It can be formed under the macroscopic angle and can distinguish 16 combinations of supersymmetric structures. In view of the abundant transformation structure in this kind of transformation system, the detailed exploration remains to be followed by the systematic expansion of theory and medical application.
Keyword SARS-CoV-2, Genomic Sequence, COVID-19, Genetic Variation, Bose Einstein distribution

Minghan Zhu
School of Software, Yunnan University
e-mail: crystalj000@163.com

Jeffrey Zheng
Key Laboratory of Quantum Information of Yunnnan, School of Software, Yunnan University
e-mail: conjugatelogic@yahoo.com

Funding Supported by the NSFC (62041213), the Key Project of Quantum Communication Technology (2018ZI002)

Introduction

Since the end of 2009, SARS-cov-2 has rapidly broken out, and the international epidemic situation has gradually become severe. The cases of Japan, South Korea, the United States and Italy are growing sharply (rapidly), and European countries are at a high risk of infection. The World Health Organization (WHO) director-general Tan Desai pointed out that now is the time for all countries, communities, families and individuals to concentrate on controlling the epidemic and preparing for a possible "pandemic". The WHO will continue to conduct risk assessments and constantly monitor the development and changes of the epidemic. Until the beginning of 2020, the International Virus Classification Committee declared that the new coronavirus would be named "SARS-CoV-2" (Severe Acute Respiratory Syndrome Coronavirus 2). SARS-CoV-2 is a type of RNA virus with an envelope and a linear single-stranded genome. The particles are round or oval, with a diameter of approximately 60 to 140 nm. Positive-strand RNA means that the virus can enter the cell directly via direct protein synthesis and self-replicate by generating negative strands with RNA polymerase. People with SARS-CoV-2 infection will have fever, blood clotting symptoms, whitish lungs and other symptoms, and severe cases may be life-threatening. Recently, asymptomatic new coronary carriers of viruses may have mutated their genes.

To study the possible genomic sequence variation, this article randomly selects 8 countries and uses the visualization method under the variant construction to add or quantify the four bases of the viral genomic sequence. Based on vector logic, modern matrix theory, geometric measure theory, combinatorial algebra and discrete mathematics, variant construction starts from n 0-1 variables to form 2^n states and 2^{2^n} functions via vector permutation and complement operations on state space to establish a variant logic framework to contain $2^n! \times 2^{2^n}$ configurations as a variation space. Variant measurement acts as a core of quantitative measurement, starting from m 0-1 variables to explore relevant clustering conditions on 2^m states. Many sample applications have been developed for 40 years using variant construction [1]-[7], such as content-based image retrieval, medical image processing, bat echo identifications, DNA maps, hierarchical organization, phase space classification, feature extraction, filtering, combinations, projections and conjugate transformations [8]. From the perspective of overall invariance, comparing the statistical distribution characteristics of the 2D and 3D diagrams, the possibility of genomic sequence differences between countries is explored macroscopically, which lays the foundation for the study of the difference between SARS-CoV-2 and the typical coronavirus genome.

Data Sources

The genomic sequence data used in this article are downloaded from the open source databases NCBI (National Center for Biotechnology Information) and GI-

SAID (Global Shared Influenza Data Initiative)[11]-[12]. The description of the data used is shown in Table 1.

samples	NO.	*Locality*
SARS-COV-2	$(2019-nCoV)$	
	$NC-045512$	China
	$LC-528233$	Japan
	$EPI-ISL-412974$	Italy
		America
	$EPI-ISL-413016$	Brazil
	$EPI-ISL-410720$	France
	$EPI-ISL-4089771$	Australia
	$EPI-ISL-413014$	Canada

Fig. 1 Datasets of SARS-CoV-2 and other cases worldwide

Distribution Characteristics and Method Description

Download representative new coronavirus SARS-CoV-2 and various influenza virus-related data at NCBI (National Center for Biotechnology Information) and GISAID (Global Influenza Data Initiative). First, the genome sequence was screened and cleaned and processed in sections to calculate the number of ATCG, four bases in the corresponding sections. Second, substitution and combination operations were performed on the calculation results of the same genome sequence, and the numbers were counted according to the same counting information contained in different segments. At the same time, different genome sequences of each virus are used for difference or sum operations, and the results are recorded to form 256 quantization matrices. Finally, combined with the visual analysis method in the variant logic system, the characteristics of the possible variation differences of the base pairs of the genome sequence are displayed from a macro perspective to form a distinguishable classification diagram with supersymmetric reflection characteristics. For the specific formula derivation process, please refer to the paper [10]. The flow of the method used is shown in Fig. 1.

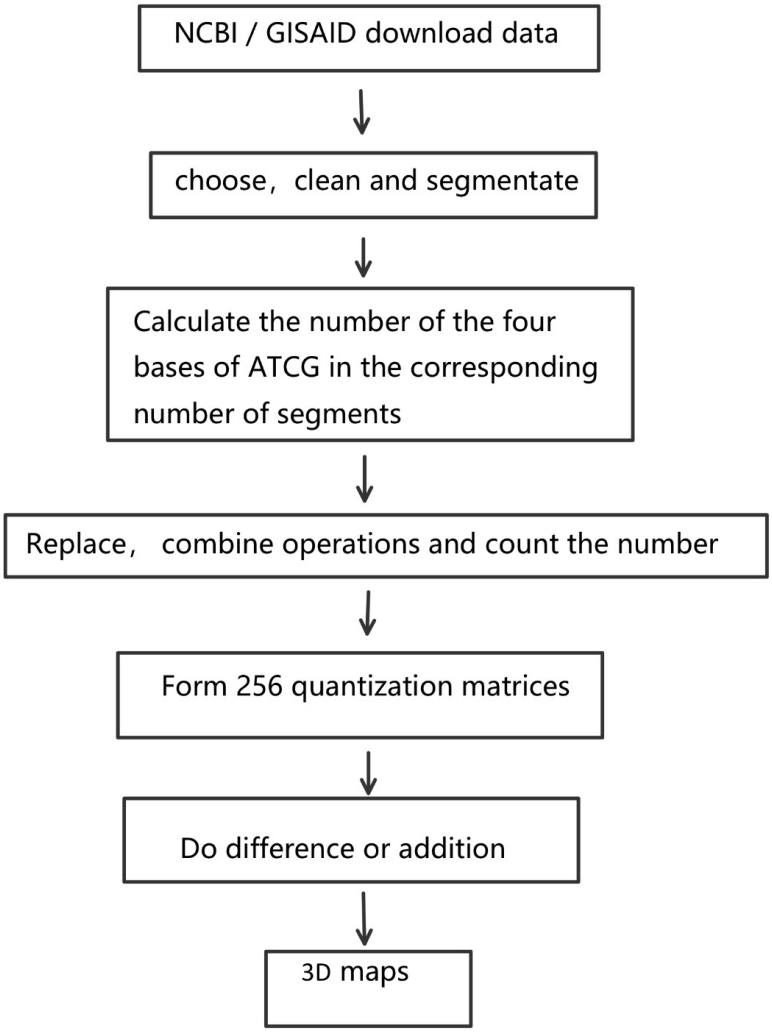


Fig. 2 The flow of the method maps

Results and Discussion

Difference Comparison Results

Randomly select the sequence of the SARS-cov-2 genome belonging to eight countries, calculate the number of bases in each segment, and perform the difference operation between the two sequences. The data size used is approximately 30k, the

number of segments is 18, the image color is strengthened from blue to red, blue represents the least data distribution, and red represents the largest number of scatter values. Taking the data from Australia and Canada as differences, Brazil and Japan, Italy and the United States, and China and France as examples, the distribution characteristics of the four images are not the same, and the differences are obvious. Among them, the difference between Brazil and Japan showed the most abundant image color and showed a similar Bose Einstein distribution. The difference between Australia and Canada is clearly distinguished by color, with yellow areas as the main color and blue as the paving. The difference between Italy and the United States is distinctive, and the data are dense. Almost all are distributed in the yellow area. However, the difference between China and France is mainly in the dark blue area, with little color fluctuations, almost no difference, and a large degree of similarity. The specific overall distribution of the difference comparison maps is shown in Fig. 2.

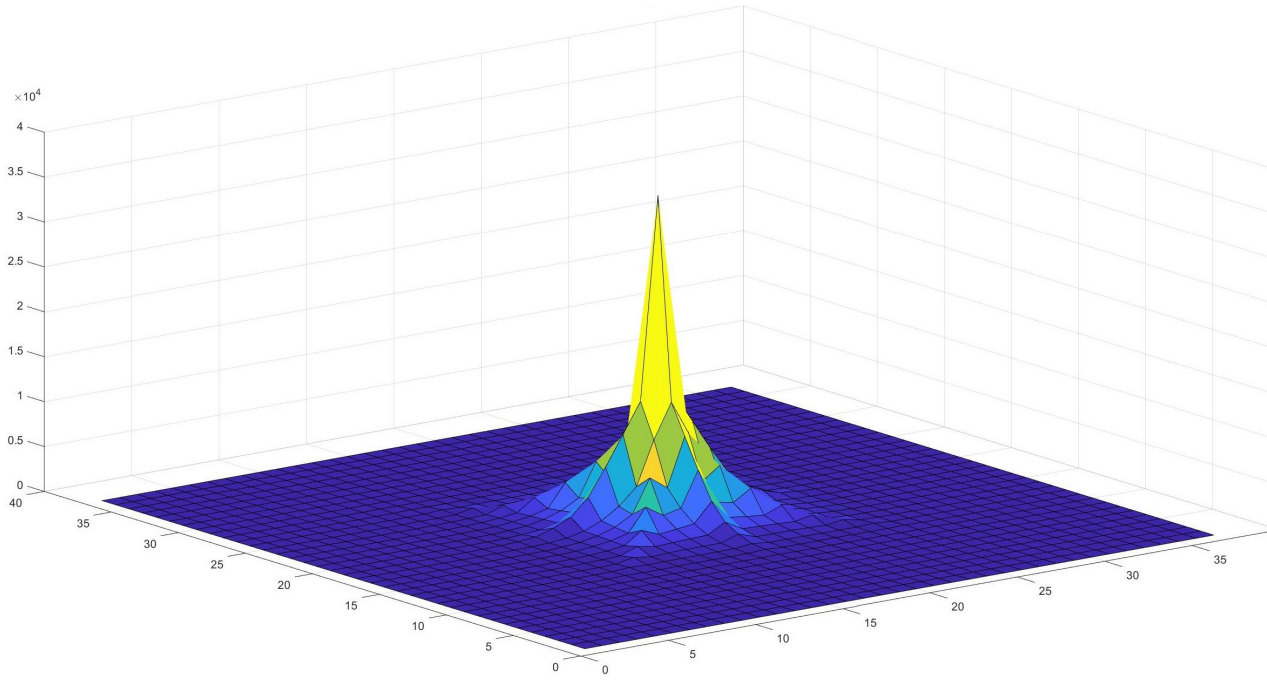
Add Operations on Comparison Results

Still taking the sum of data from Australia and Canada, Brazil and Japan, Italy and the United States, China and France as examples, it can be found that the sum of data from Australia and Canada and the sum of Brazil and Japan show the richest colors, and the distribution of the two images is very similar, not as obvious as the comparison chart. Italy and the United States have distinctive summing characteristics. The image is composed of many cones, without too much dense data, and is distributed more in the blue and green areas. The combined images of China and France are almost evenly distributed, mainly in dark blue areas, with little color fluctuation. The specific overall distribution of the added comparison maps is shown in Fig. 3.

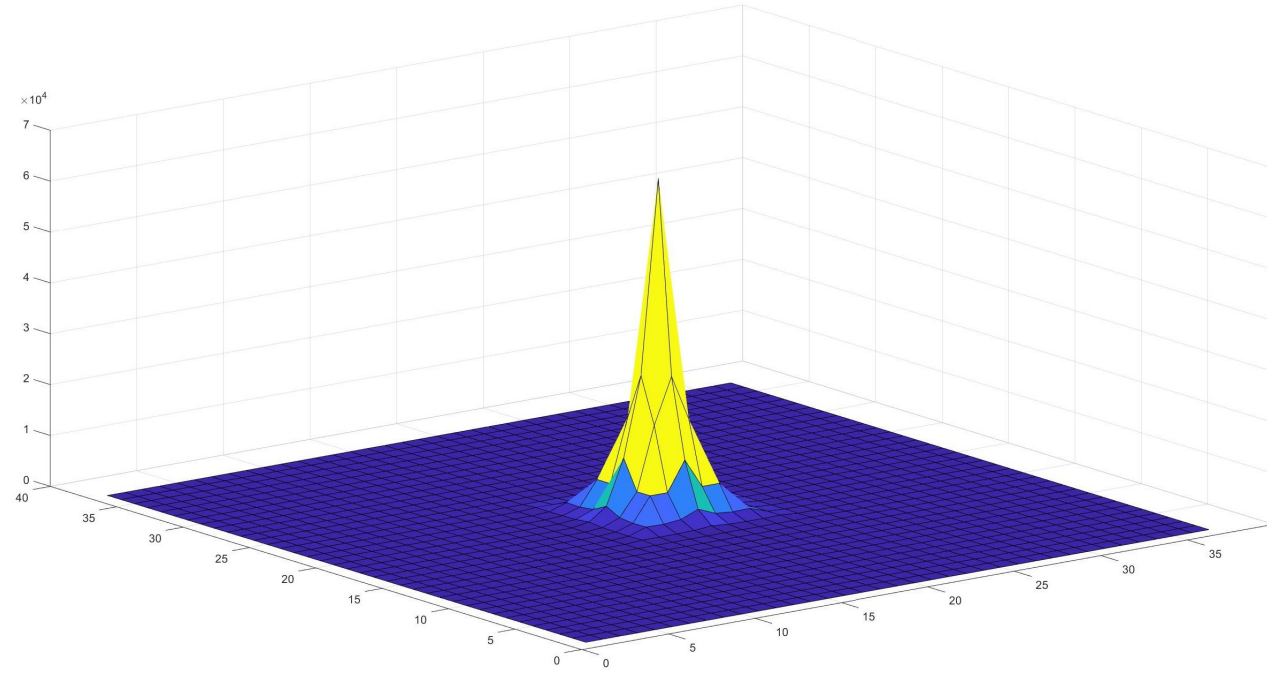
Conclusion

Bioinformation data are increasing with the development of life sciences. Due to the complexity of living organisms, biological information often has a larger quantity with levels of more uncertainty, and complicated relationships. Results of biological information is more secret and complex. People's research on biological information is in the stage of exploration and discovery. The analysis of massive and complex biological information data and the development of new visual analysis tools are more practical.

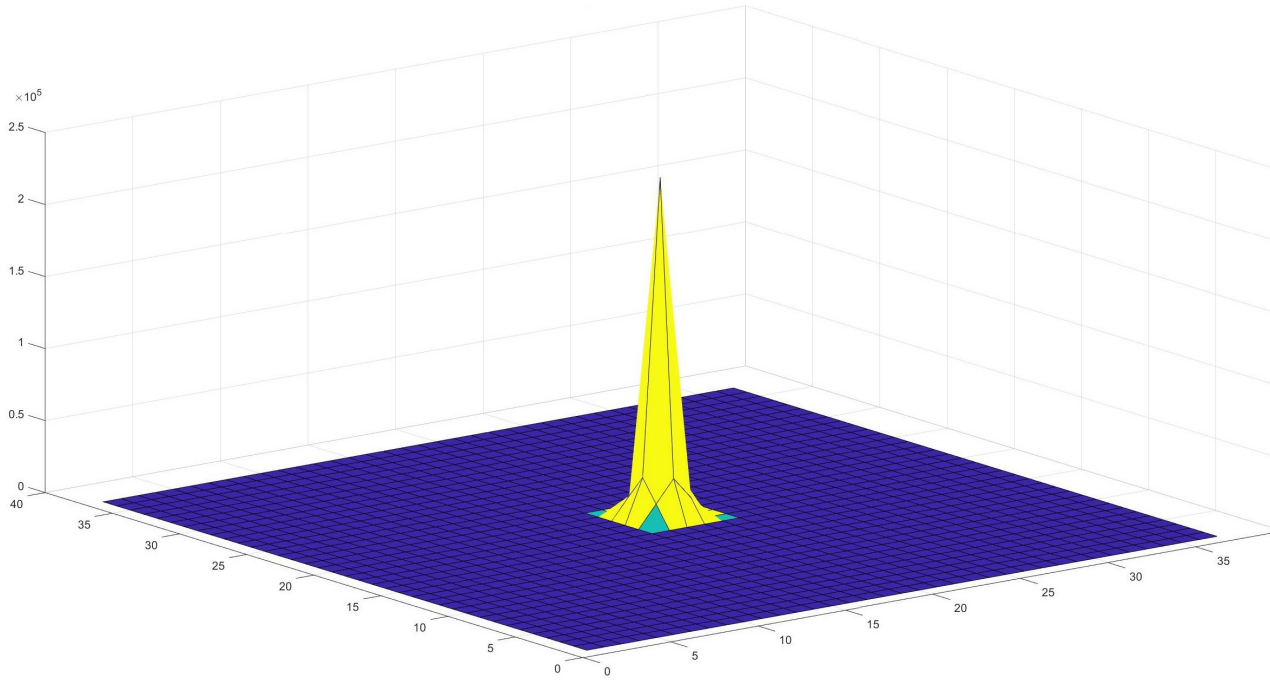
This paper forms a quantization matrix based on the four base arrangements in the genome sequence. Using the most concise method of calculation of difference and addition, the sequence differences of the genome sequence of SARS-cov-2 under different samples are displayed. The analysis and visualization methods used


×10^4
4
3.5
3
2.5
2
1.5
1
0.5
0
40
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35

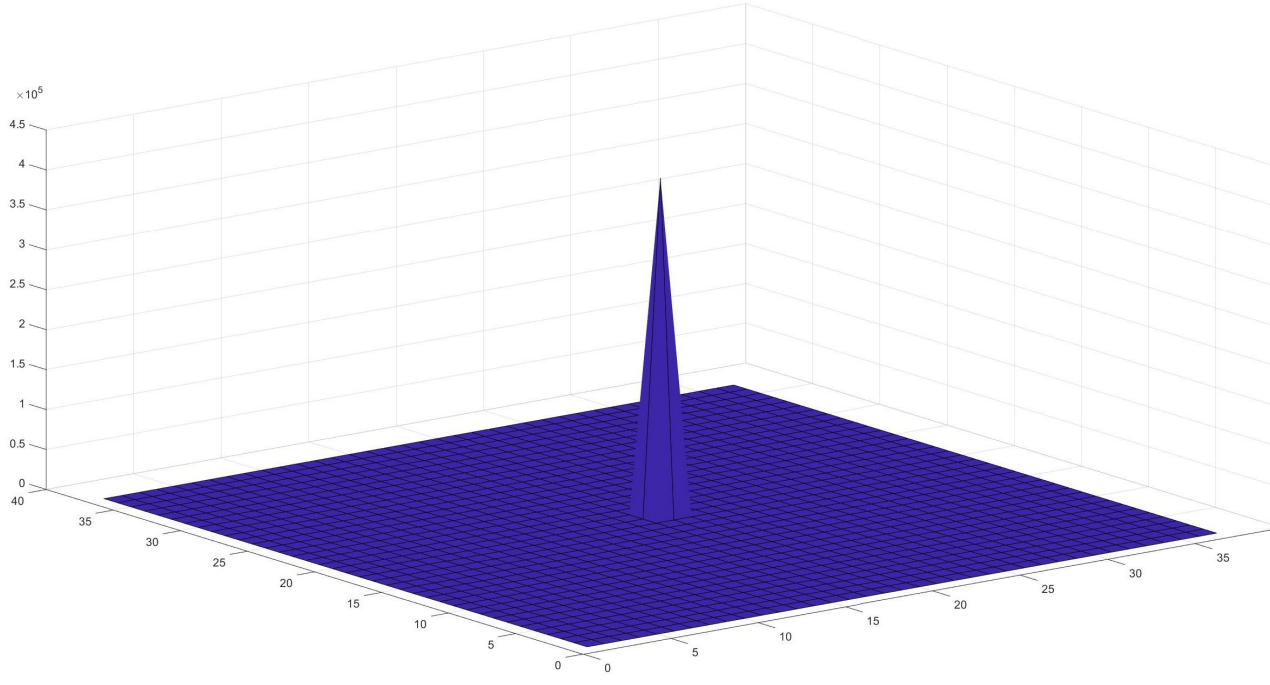
(*a*)


×10^4
7
6
5
4
3
2
1
0
40
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35

(*b*)

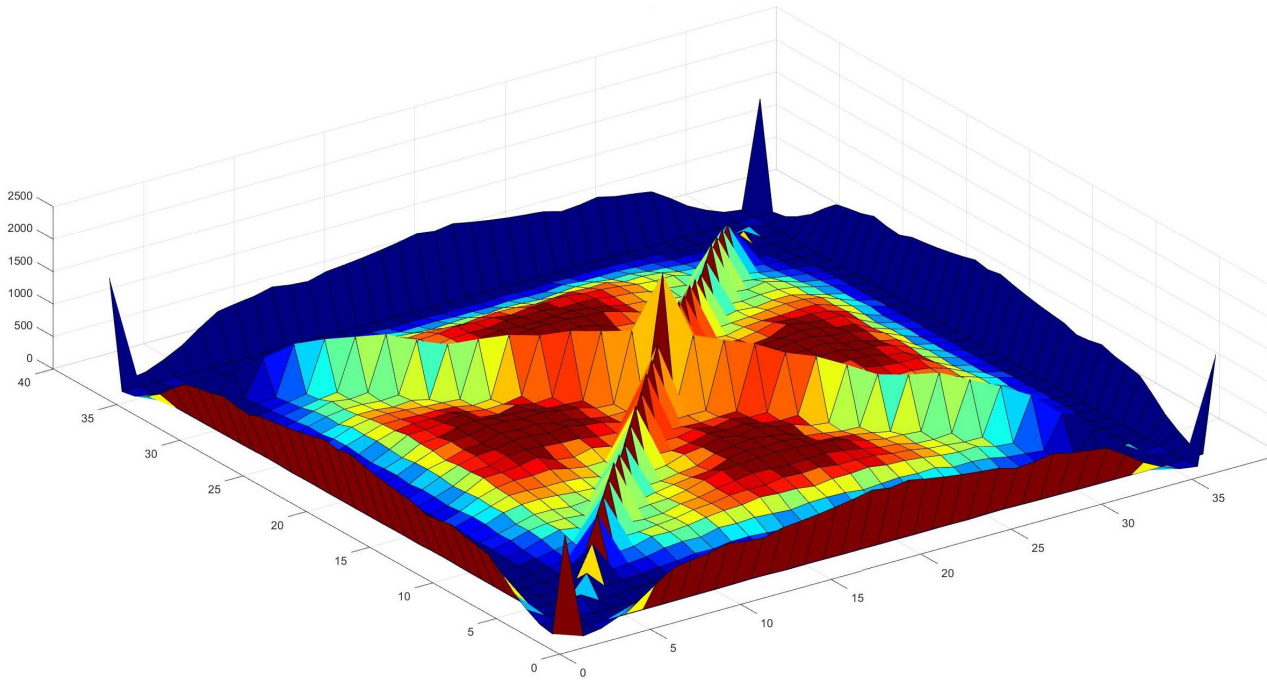


(*c*)

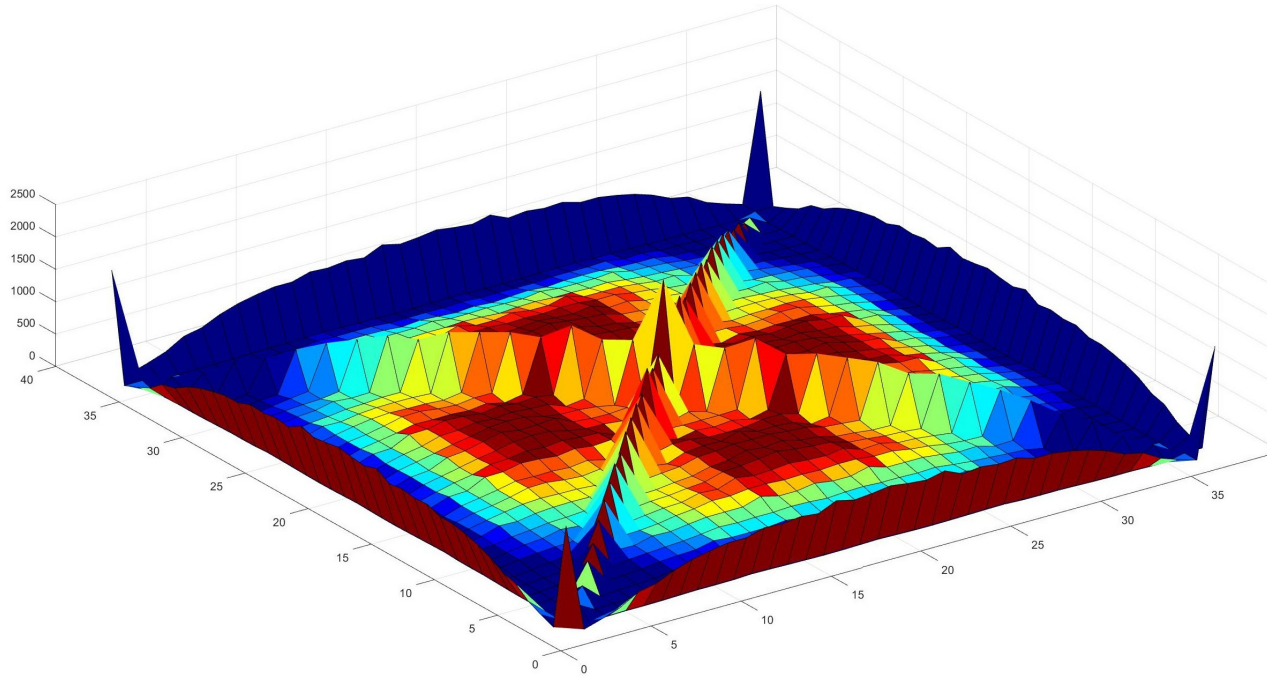


(*d*)

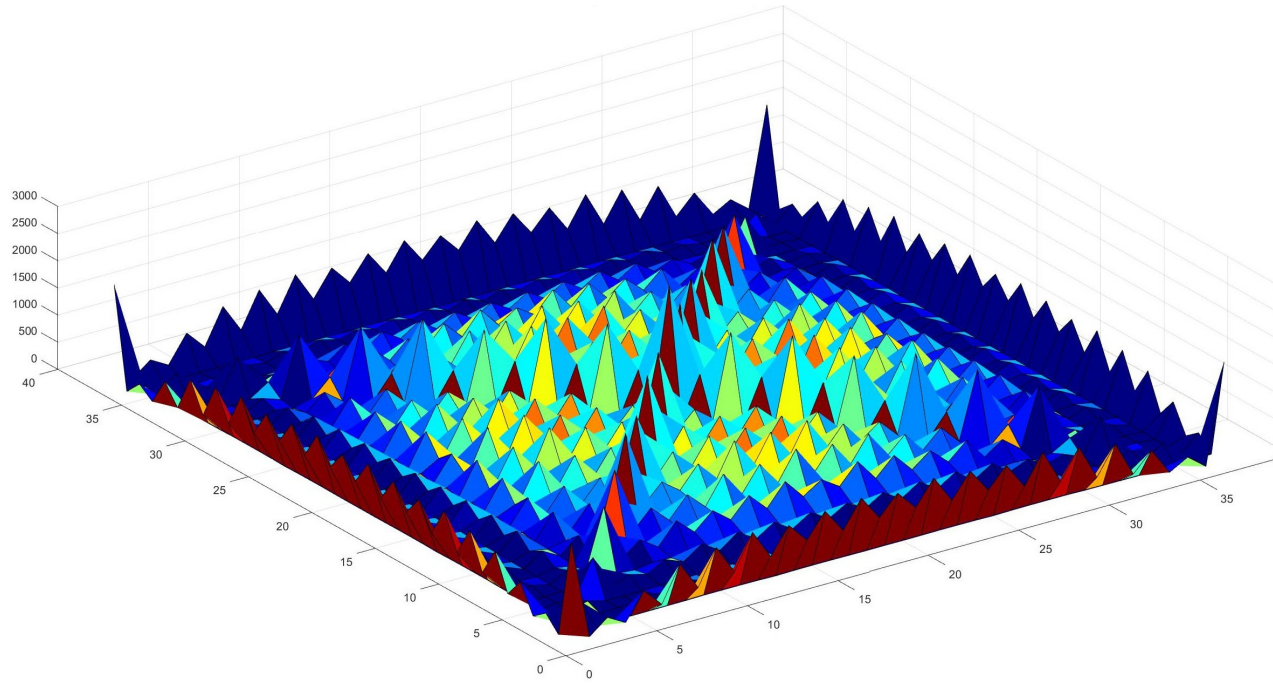
Fig. 3 Difference comparison in four maps (a)-(d), m=18; (a) Brazil-differ-Japan (b) Australia-differ-Canada (c) Italy-differ-America (d) China-differ-France


2500
2000
1500
1000
500
0
40
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35

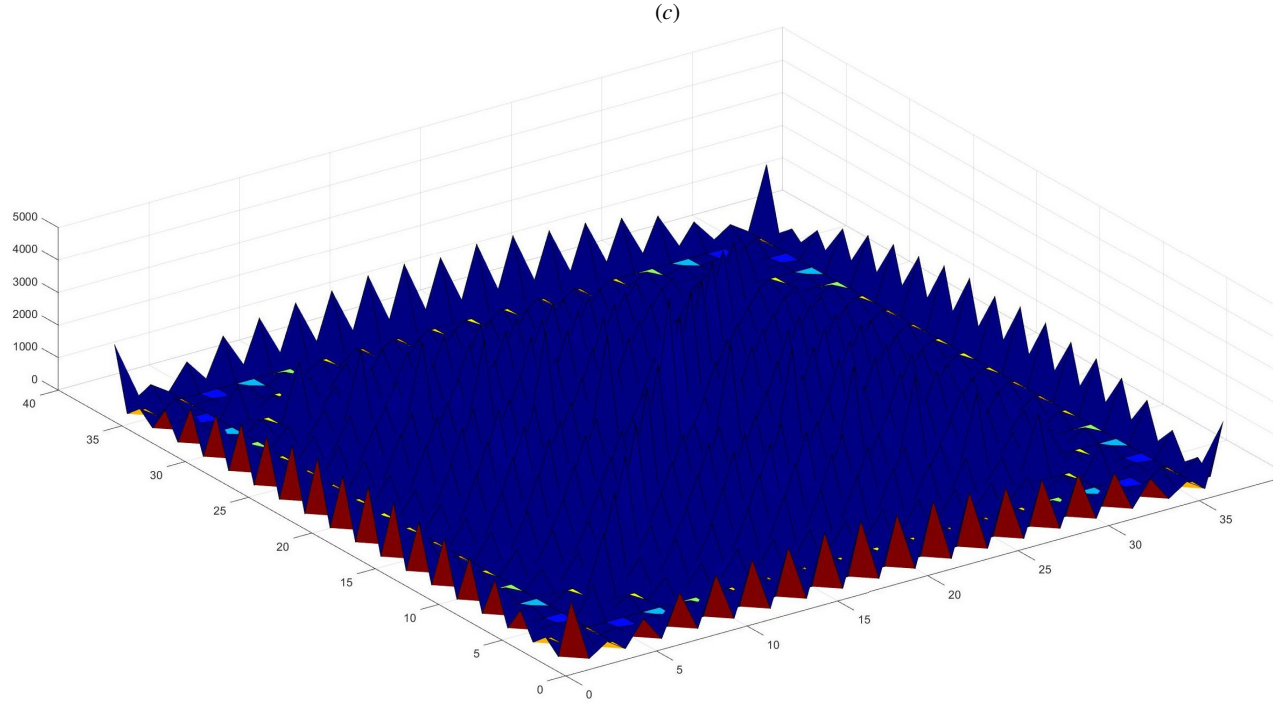
(*a*)


2500
2000
1500
1000
500
0
40
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35

(*b*)



(*c*)



(*d*)

Fig. 4 Add operations on comparison results in four maps (a)-(d), m=18; (a) Brazil-differ-Japan (b) Australia-differ-Canada (c) Italy-differ-America (d) China-differ-France

provide information support for medical research on a theoretical basis for the fields of computers, big data, artificial intelligence, and machine learning.

Conflict Interest

No conflict of interest has been claimed.

Acknowledgements

The authors would like to thank NCBI, GISAID, Nextstrain for providing invaluable information on the newest dataset collections of SARS-CoV-2 and other virus genomes to support this project working smoothly.

References

1. Z. J. Zheng, A. Maeder, The conjugate classification of the kernel form of the hexagonal grid, Modern Geometric Computing for Visualization, Springer-Verlag, 73-89, 1992.
2. Z. J. Zheng,Conjugate transformation of regular plan lattices for binary images, PhD Thesis, Monash University, 1994.
3. Jeffrey Z. J. Zheng, Christian H. H. Zheng, A framework to express variant and invariant functional spaces for binary logic, Frontiers of Electrical and Electronic Engineering in China, 5(2):163-172, Higher Educational Press and Springer-Verlag, 2010.
4. Jeffrey Z.J. Zheng, Christian H.H. Zheng and Tosiyasu L. Kunii. A Framework of Variant Logic Construction for Cellular Automata, Cellular Automata - Innovative Modeling for Science and Engineering, Dr. Alejandro Salcido (Ed.), InTech Press, 2011.
5. Jeffrey Zheng, Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019 https://www.springer.com/in/book/9789811322815
6. Jeffrey Zheng, Variant Construction Theory and Applications, Vol.1: Theoretical Foundation and Applications, Science Press 2020 (Chinese, Formal Publishing Soon). 郑智捷, 变值体系理论及其应用 第 1 册：理论基础及其应用，科学出版社 2020 (即将正式发行)
7. Jeffrey Zheng, Research Gate: http://researchgate.net/pprofile/JeffreyZheng
8. Jeffrey Zheng, Chris Zheng, Biometrics and Knowledge Management Information Systems, Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019, 193-202 https://link.springer.com/chapter/10.1007/978-981-13-2282-2 11 被斯普林格－自然杂志出版社，选入抗击新型冠状病毒肺炎研究(Research of COVID-19)资料汇集。推荐给 PMC 和 WHO（PubMed Central PMC and the World Health Organization WHO）以方便全球科学研究人员免费使用。
9. Jeffrey Zheng, Jianzhong Liu, A Visual Framework of Meta Genomic Analysis on Variations of Whole SARS-CoV-2 Sequences[J]
10. Jeffrey Zheng, Minghan Zhu, Input-Outpout Types of Fifteen Modules on Discrete and Real Measurements for COVID-19[J]
11. GISAID: Open access to influenza virus data https://gisaid.org
12. NCBI: Open access to dataes https://www.ncbi.nlm.nih.gov
13. Yanni Li et al. Similarities and Evolutionary Relationships of COVID-19 and Related Viruses[J].arxiv.2020.

14. Yu WB, Tang GD, Zhang L, et al. Decoding evolution and transmissions of novel pneumonia coronavirus using the whole genomic data[J]. ChinaXiv. 2020:202002.00033.
15. Peter Forster, Lucy Forster, Colin Renfrew, et al. Phylogenetic network analysis of SARS-CoV-2 genomes[J]. PANS. 2020.

Figures

samples	NO.	*Locality*
SARS-COV-2	$(2019-nCoV)$	
	$NC-045512$	China
	$LC-528233$	Japan
	$EPI-ISL-412974$	Italy
		America
	$EPI-ISL-413016$	Brazil
	$EPI-ISL-410720$	France
	$EPI-ISL-4089771$	Australia
	$EPI-ISL-413014$	Canada

Figure 1

Datasets of SARS-CoV-2 and other cases worldwide

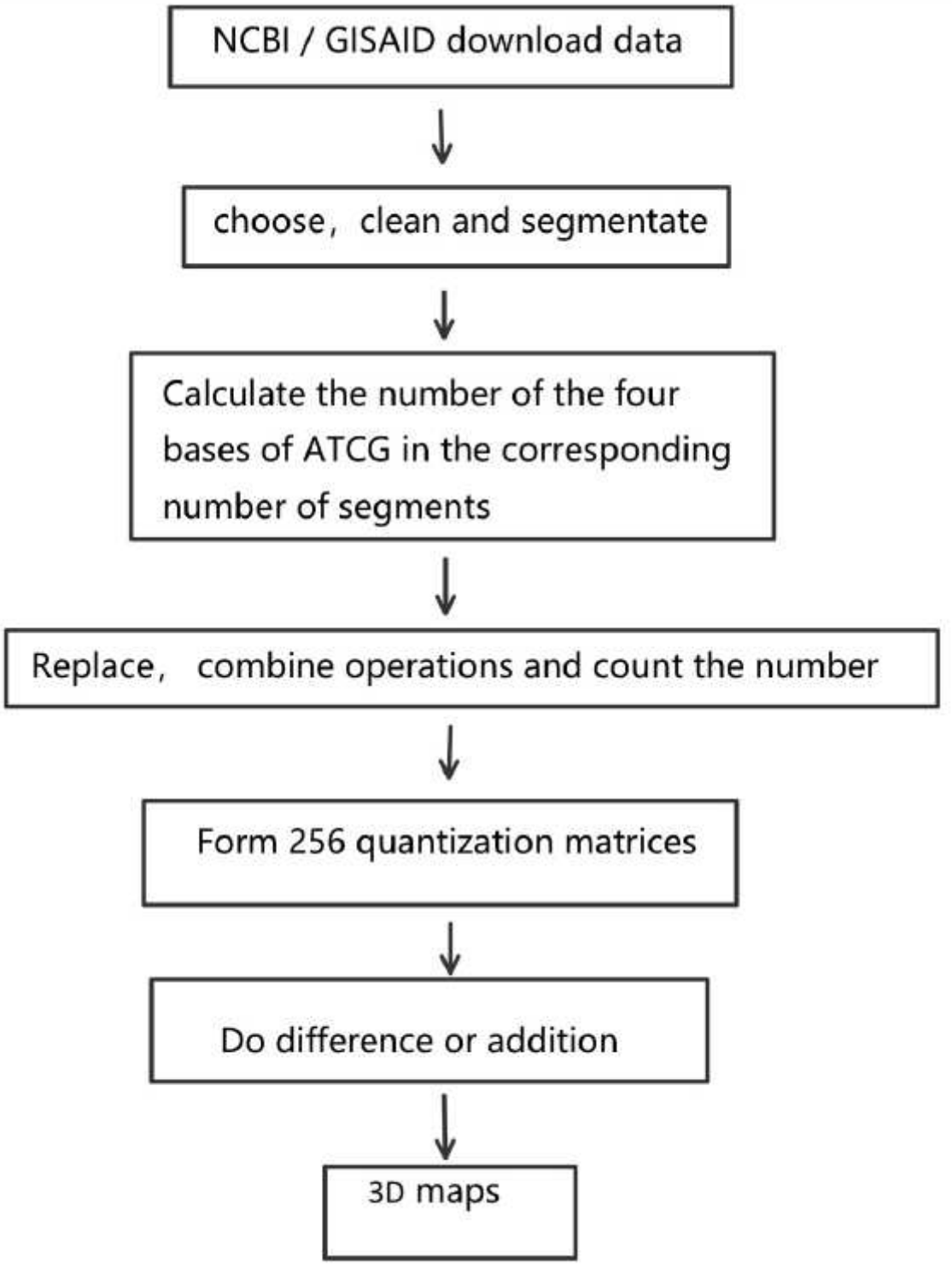


Figure 2

The flow of the method maps

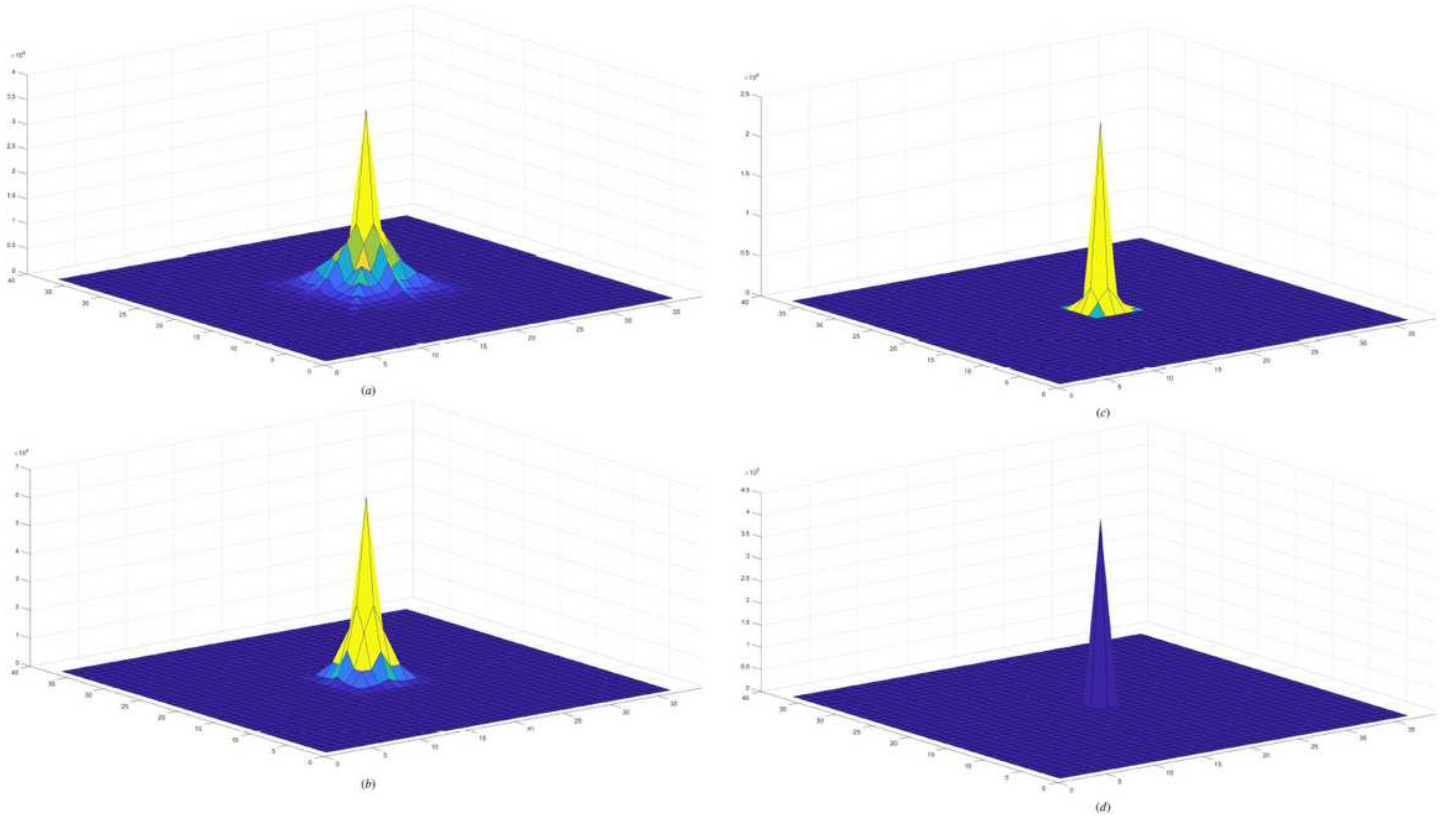


Figure 3

Difference comparison in four maps (a)-(d), m=18; (a) Brazil-differ-Japan (b) Australiadiffer- Canada (c) Italy-differ-America (d) China-differ-France

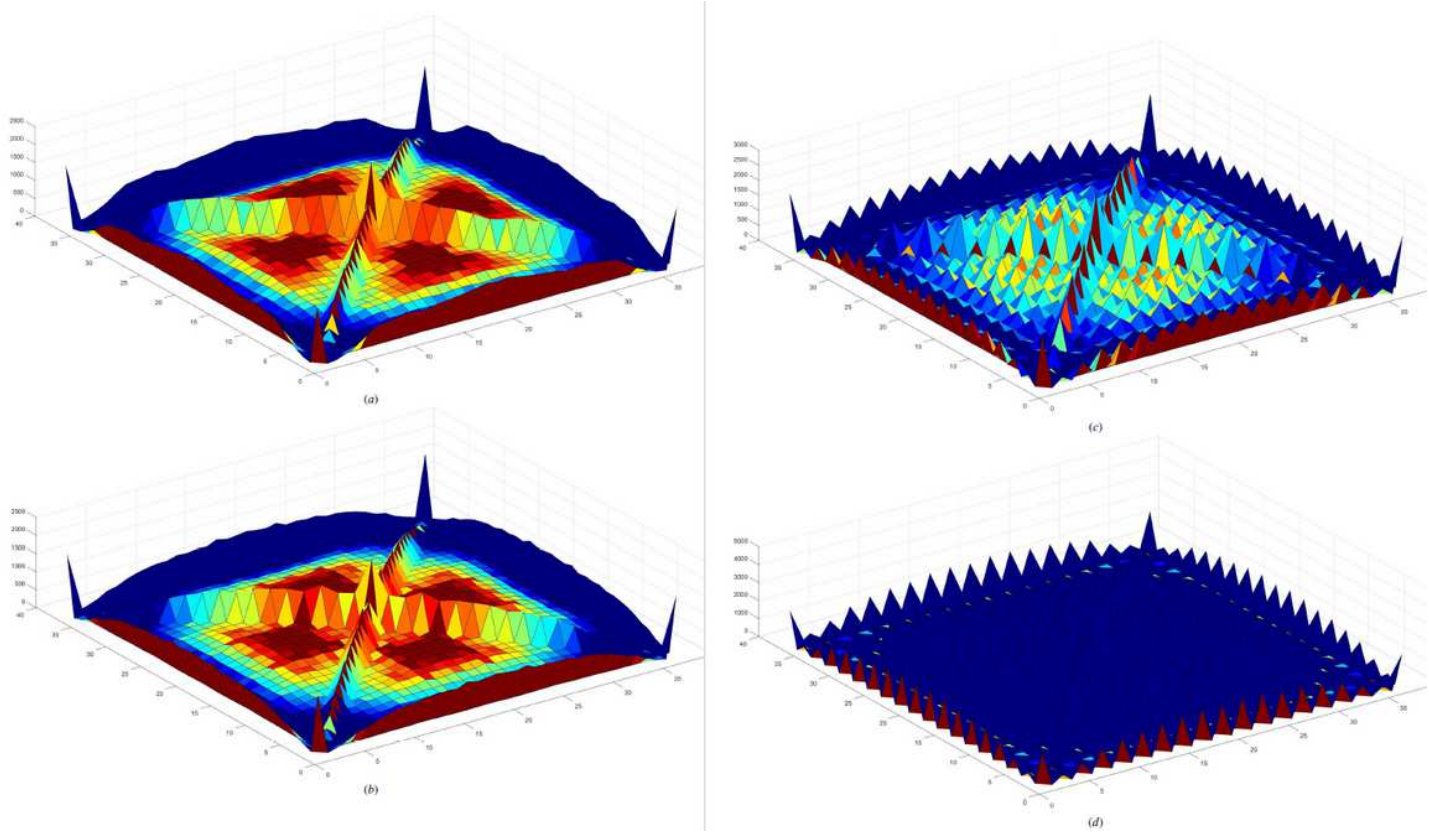


Figure 4

Add operations on comparison results in four maps (a)-(d), m=18; (a) Brazil-differ-Japan (b) Australia-differ-Canada (c) Italy-differ-America (d) China-differ-France