VOL. 19, NO. 1, JANUARY 2024 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

COMPOSITIONAL KRIGING ANALYSIS: A SPATIAL INTERPOLATION METHOD FOR DISTRIBUTIONS

Ahram Kim, Bashir Busahmin and Stephen Tyson
Department of Petroleum and Chemical Engineering, Universiti Teknologi Brunei, Brunei Darussalam
E-Mail: ahram.kim0805@gmail.com

ABSTRACT

Effective development of subsurface petroleum resources relies on the estimation of spatially distributed parameters at undrilled locations. Established geostatistical algorithms based on kriging exist for the estimation of scalar spatial variables such as porosity and permeability. This may not be suitable for the estimation of distributions of properties, such as grain size, whose whole distribution varies spatially. The current conventional approach is to fit a normal or log-normal distribution to the available data, and then to estimate the parameters of the distribution, the mean and standard deviation, spatially using kriging, taking care to consider any dependence between the mean and standard deviation. The assumption that all the grain size distributions can be approximated by a single distribution type is unsatisfactory, since datasets have very different-looking distributions, with variations in skewness, kurtosis, and modality. This paper presents an alternative approach that can handle significant variability in the distribution shape by separating the distribution into bins, like a histogram, and treating these bins as percentages in a composition. Compositional data needs to be mapped onto a simplex to overcome spurious correlations between those components, in addition, spatial estimation methods for compositional data have already been developed. However, the contribution to this field is the mapping of continuous data from distributions into a composition that enables the compositional kriging method to predict distributions at new locations. Moreover, the results showed the prediction distributions in the presence of varying modality, skewness, and kurtosis. Therefore, the grain size datasets in this paper have been working with the confidentiality restrictions so it explains the technique with a dataset of population ages from the US census 2010 for the state of Texas, which shows similar variability in distribution.

Keywords: kriging, spatial, data, variables, interpolation.

Manuscript Received 14 October 2023; Revised 16 February 2024; Published 12 March 2024

1. INTRODUCTION

Spatial interpolation of distributions with varying numbers of modes is a useful technique in the estimation of the distributions [1] and [2] discussed calculating the spatial continuous data using the interpolation methods required undivided disciplines relative to the earth's surface. In addition, the importance of the application was investigated using the geostatistical method in the environment of continuous data [3]. Moreover, several studies related to the geostatistical method showed a better result than the interpolation method, and the process of estimating the values at unsampled locations was based on the observations of the samples taken from different locations, however, there are different types of spatial interpolation methods such as the Kriging method [4]. Furthermore, [5] studied the continuous data, and population density and found that the ordinary kriging technique is more accurate than other methods such as the areal weighting (AW) method to estimate the population of a new location spatially, thus, choosing the suitable interpolation method for a given input dataset may pose the challenges but only when the recommendation for estimating the continuous data is a kriging method, however, the spatial interpolation using kriging method is preferred, but there is some limitation in the situation where the form of the distribution is different. In addition, it was explained as a conditional density estimator with a mean of a function specified by local polynomial smoother and if the data are not spatially independent then it might be incorrect. Moreover, the compositional kriging is delivered as an extension of the ordinary kriging that complies with the constraints, [6] Moreover, stated that the compositional kriging method presented a simple extension of the ordinary kriging, Thus, compositional kinging CK is an innovative approach to investigate spatial interpolation for the population data. [7] introduced the comparison of spatial interpolation methods for the estimation of air temperature in Sydney, Australia, where it improved the precision in predicting the air temperature measurements However, another example in terms of continuous data is soil properties, and [8] examined that the better result of interpolation is the kriging method rather than other techniques like IDW and spline related to soil properties as the continuous data on the map, where further emphasized that the kriging method tends to exhibit superior performance compared to other methods specifically the spatial correlation within the data. Based on this, [9] stated that the compositional kriging considered the spatial correlation between the transformed parameters generated the predictions. In addition, [10] discussed and found that the method for using the multivariate statistical analysis on the compositional data helped the correlations with different components and increased the accuracy of the estimations; therefore, the method was classified to be an introduction of an uncomplicated method used for the extension of the

ordinary kriging. Furthermore, a comprehensive overview of the geostatistical methods was studied with the application for modelling the spatial uncertainty [11], where it was stated that the compositional data has shown as a remarkable value where in particular, the method is capable of the interpolation and the estimation of continuous data that are integrally compositional. Therefore, [12] came across the advantage of the CK technique as the result is dependable using the spatial interpolation without any limited area, specifically, when dealing with the proportions of density data and to overcome any of the limitations, the compositional kriging was introduced. [13] Stated that Python is a useful tool to analyse the density data and Python for compositional kriging because of its adaptability and ability to be customized. In addition, Python, being a versatile programming language, provides the freedom to modify and tailor the code to suit specific needs, additionally, Python offers a wide selection of libraries and modules that can be utilized for various tasks involved in compositional kriging, including data preprocessing, variogram modelling, kriging estimation, and result visualization. This compatibility enables easy incorporation of compositional kriging workflows into existing data pipelines and allows for smooth integration of the results with other spatial analysis processes. Therefore, it is allowed to offer powerful libraries for data analysis and visualization, including NumPy, pandas, and Matplotlib. These libraries provide efficient tools for handling and analysing large datasets, conducting statistical calculations, and generating meaningful visualizations of the compositional kriging results. The objective of this paper is to analyse the data selected from Texas population density data for spatial estimation region using compositional kriging method (CK) by applying Python programming language, also, this paper focuses on the understanding of CK along the variograms and covariance to comprehend the spatial dependency and the correlation between each age group.

2. MATERIALS AND DATA PROCESSING

The total number of data is 254 by city within 18 age groups. The data for this paper is collected from the public domain related to the census Texas 2010, as shown Table-1.

Table-1. Group number and age group [Texas population density data in 2010].

Group number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Age group	< 5	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-50	50-54	55-59	60-64	65-69	70-74	75-79	80-84	> 85

Based on proportion density data, Python code is created to obtain a reliable result in this research paper. Since the data is already in compositional data format it is allowed to directly check the variogram. One advantage of using Python for compositional kriging is its adaptability and ability to be customized. utilizing Python for compositional kriging is its seamless integration with other commonly used software and tools in geospatial analysis, including GIS software, databases, and web frameworks. This compatibility enables easy incorporation of compositional kriging workflows and allows for smooth integration of the results with other spatial analysis processes. Python programming is a convenient option for comprehensive data analysis workflows and is flexible to customize. In addition, it was applied to enable visualizations as to which group cluster of density data belongs to an area of the state of Texas. Moreover, this ensures an easier understanding ability to detect the proportion of each group age where exactly centered in the map.

3. RESULTS AND DISCUSSIONS

Figure-1 presents the proportion of 5 to 9 years, as per group 2 based on the Texas map, which means that people lived near the border area that is called Starr County. Figure-2 similarly shows the populations under group 5 on the map of Brazos County, however the highest proportion in age ranged between 20 and 24 as a result, most of the people are living near the cities, the reason behind this is because of the young generations need for education, jobs etc. in the cities. Interestingly, other groups like group 10 as presented in Figure-3 with the proportion in the age of 45 to 49 years old based on the Texas map is called Loving County and is located in the southwestern region which is mostly identified as an urban area, where people just moved after their material status has changed. In addition, it was said that when people get married and want to settle down, usually they move to locations close to families and workplaces hence large population density ends in the urban area. Moreover, in an urban area job opportunities might be available alongside the educational system for growing families.

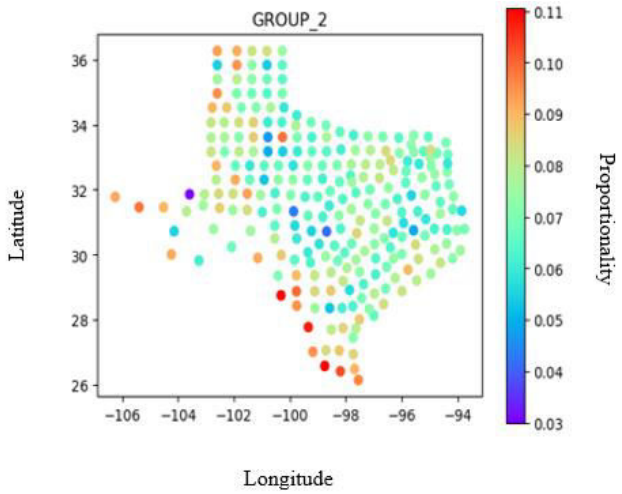

Figure-1. The proportion density of 5 to 9 years old based on the Texas map.

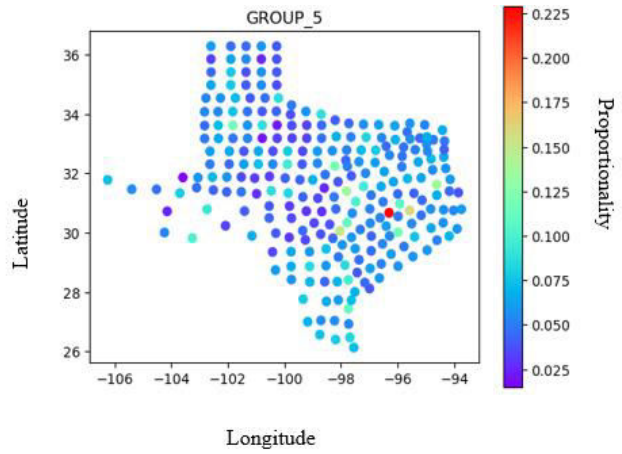

Figure-2. The proportion of 20 to 24 years old on the Texas map.

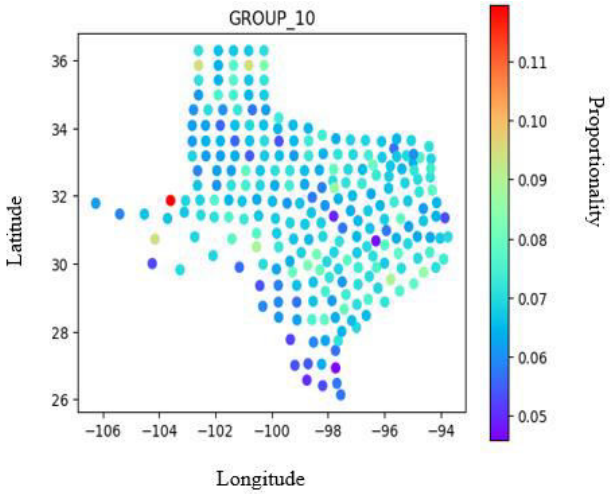

Figure-3. The proportion of 45 to 49 years old on the Texas map.

Figure-4 presents different shapes of histogram distribution on the Texas map. The accuracy of the interpolation is affected by the density data and the shape in the distribution pattern therefore strategies are suggested to manage the interpolation. However, many places on the map have interestingly different distribution shapes and this finding is aligned with [13]. Moreover, the diverse distribution shapes are detected by expanding the map. In addition, [14] compared the regression and the interpolation method for mapping the groundwater quality parameters and explained that the performance of the interpolation method changed the impact of different distribution patterns and affected its shape as a result providing an insight into the challenges that occurred due to the variations and emphasized the importance of checking the covariance between the two variables to estimate the interpolation.

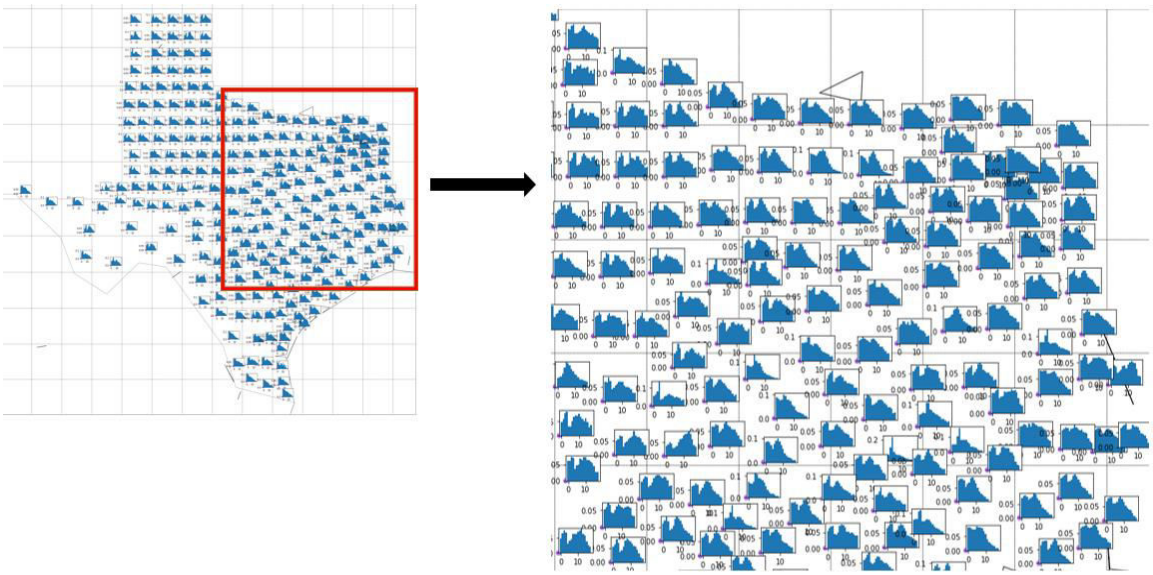

Figure-4. Different shapes of histogram distribution on Texas map.

Figure-5 presents the correlation between age group 5 (20-24 years old) and age group 6 (25-29 years old). It shows that both groups are similar and are strongly dependent correlated, however, age group 2 (5 - 9 years old) and age group 11 (50-54 years old) are large age groups, nonetheless, there is still a weak negative correlation between the two age groups as shown in Figure-6, as a result, the interpolation scheme should honour the dependency between the diverse groups as well as the modality of the populations at each point.

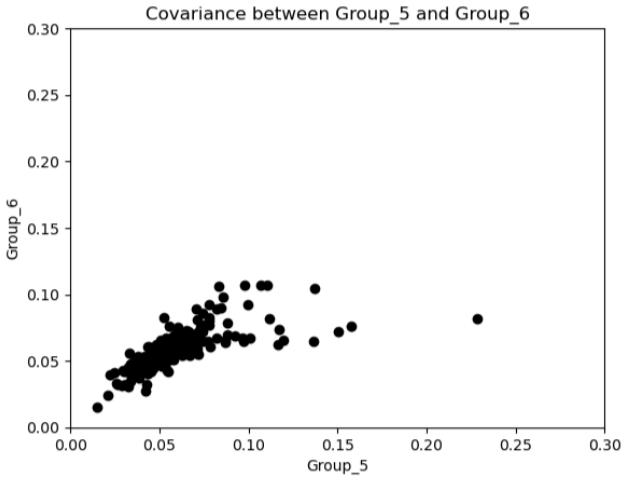

Figure-5. A covariance between group 5 and group 6.

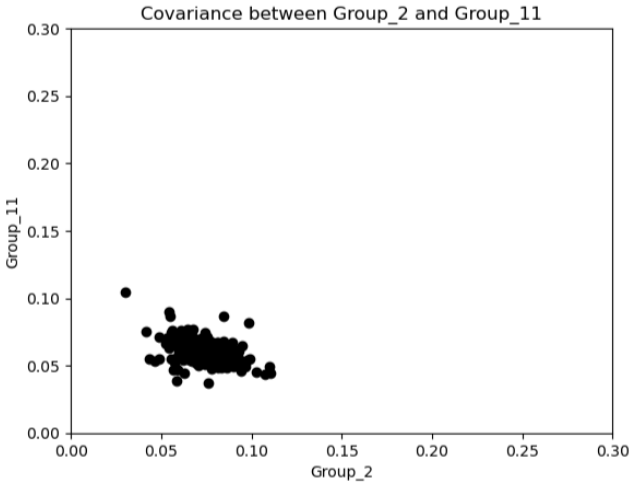

Figure-6. A covariance between group 2 and group 11.

Geostatistics relies on the analysis of the underlying spatial model using a variogram. Therefore, the variogram for 25-29 years old in age Group 6 as shown in Figure-7 has a range of 200 km and beyond that distance the data points behaved as independent, however, the closer the data points are, the more likely they are similar. As a result, the variogram shows a sill of 0.001, the value approaches and stabilizes at 0.001 as the lag. In other words, spatial dependency in the data to a certain degree, but the variability is limited to 0.001 as the maximum value. In addition, there is also a small nugget which indicates a small element of the randomness in the locations that are remarkably close to each other.

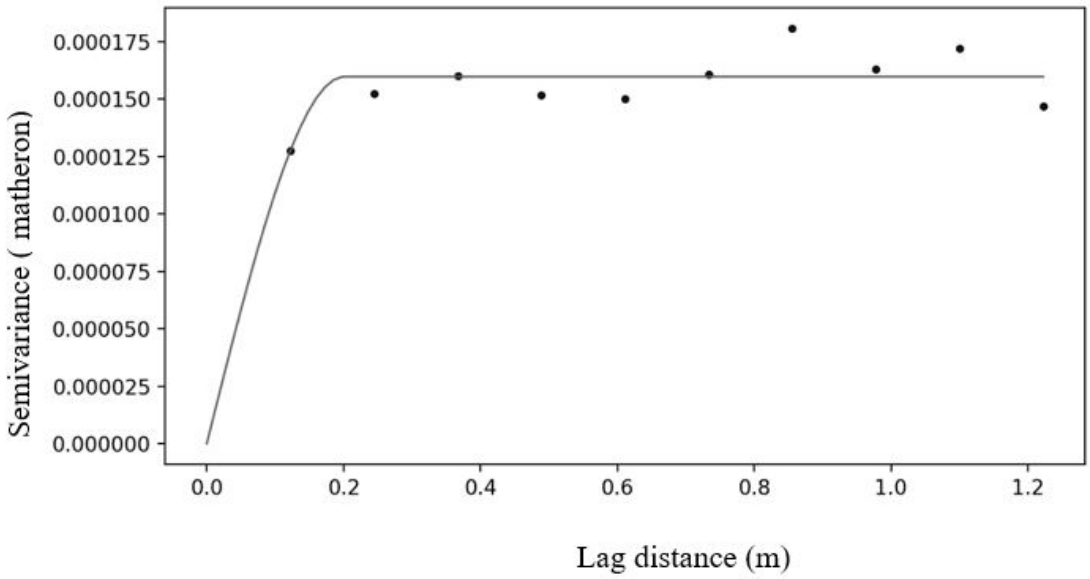

Figure-7. The variogram for 25-20 years old (Group 6).

Since the variogram and covariance are spatially dependent, it is allowed to interpolate density data in spatial. Moreover, a comparison between the kriging data points and the original data points is presented in Figures 8, 9 10, and 11 with each county among the eighteen age groups, furthermore, there are two picks from the shape of the distribution, where it is challenging to estimate the type of the skewness by applying the CK interpolation method and it is possible to predict the values of the population in each group. Interestingly, County ID 159 as

shown in Figure-8, and County ID 14 in Figure-11 where do not exactly follow the same trend as the original data distribution, but still show a similar shape of the distribution. Thus, this kind of technique assisted in increasing the accuracy of the estimation by age group. Moreover, Figures 9 and 10 showed the same behaviour.

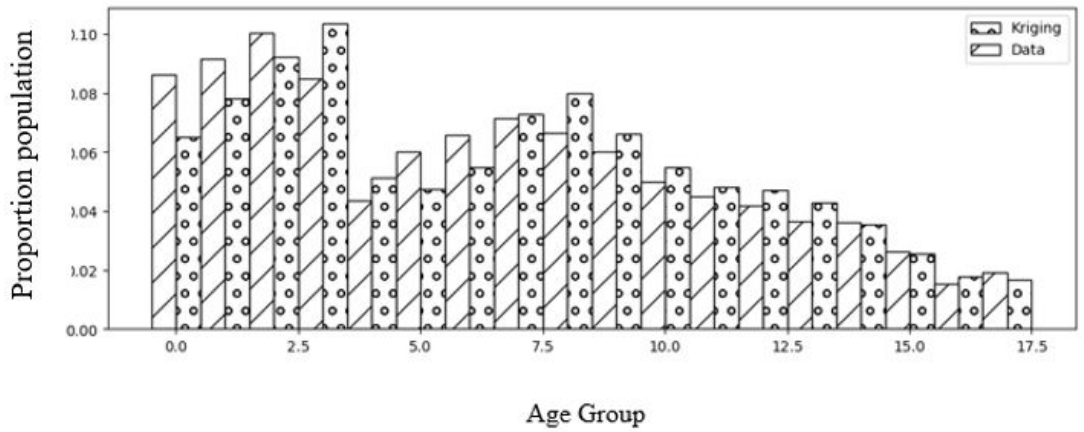

Figure-8. The distribution for Kriging and original data in Marion County ID 159.

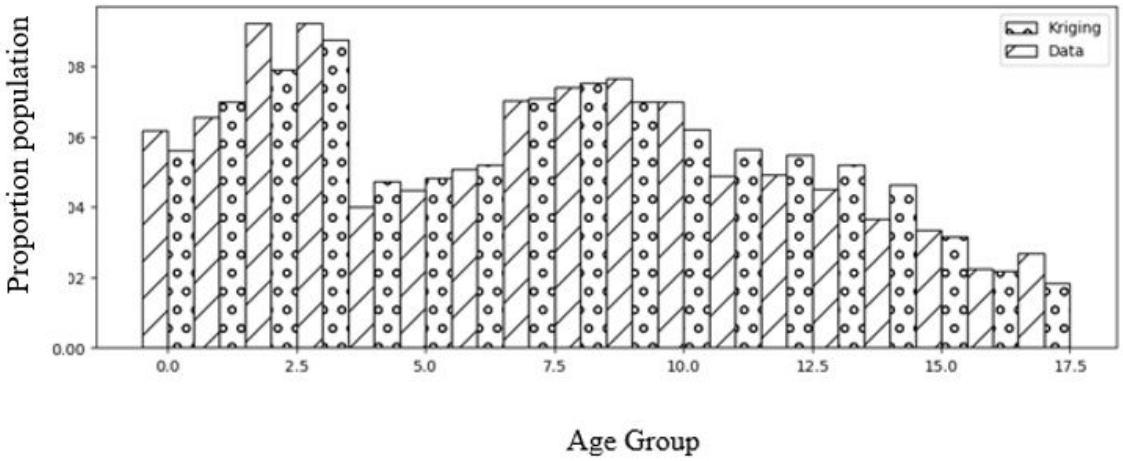

Figure-9. The distribution for Kriging and original data in Scurry County ID 207.

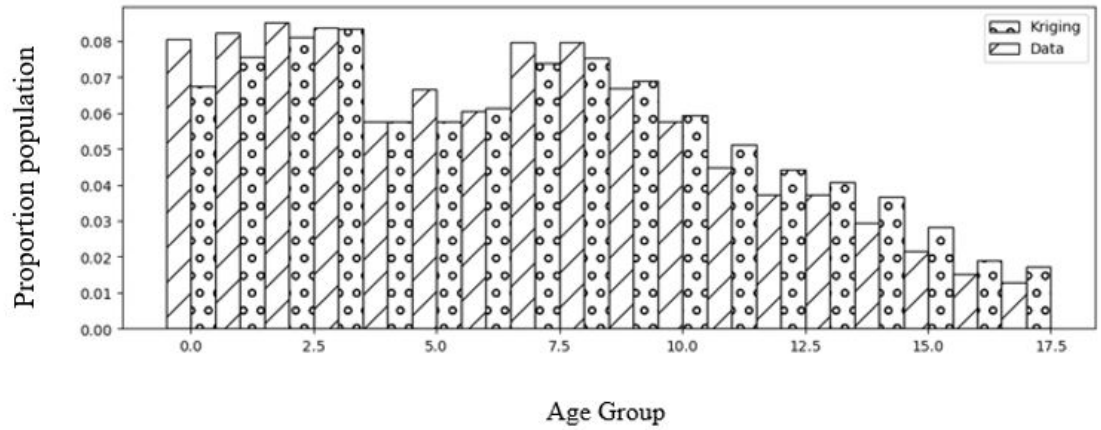

Figure-10. The distribution for Kriging and original data in Oldham County ID 179.

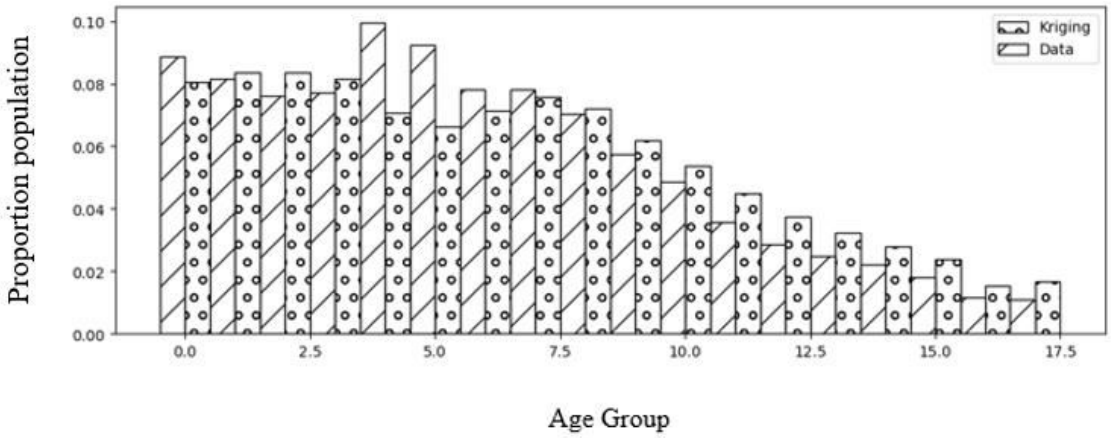

Figure-11. The distribution for Kriging and original data in Bexar County ID 14.

4. CONCLUSIONS

This paper presented a spatial interpolation technique of compositional kriging (CK) to interpolate the different spatial distribution forms on the Texas map by age groups. The compositional kriging method was applied to continuous data obtained from an international domain of Texas by using, for example, python coding and its application. Covariance and variogram were applied to detect the spatial dependency on continuous data, it was concluded that the spatial estimation of the data distribution which has a modality changed over space was obtained similarly matched to the origin data. In addition, Texas population density was predicted using compositional kriging through the shape of the distribution and locations in the population data. Furthermore, the interpolation method presented a significant role in accomplishing high accuracy of the data predicted from various locations. It is to conclude that the compositional kriging is more accurate in interpolating with the compositional data due to its simplicity.

Funding

This research received no funding.

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

[1] Li J. and Heap A. D. 2008. A review of spatial interpolation methods for environmental scientists.

[2] Armando Montanari. 2019. From Geographic Information Systems (GISy) to Geographic Information Science (GISc). In book: Laboratori Geografici in Rete: ricerca, didattica, progettualità, publisher: Labgeo Caraci, Roma.

[3] Alan T. Murray. 2020. Spatial Analysis and Modeling: Advances and Evolution. Geographical Analysis 53(2), DOI:10.1111/gean.12263.

[4] Huang J., Huang Y., Sun Y. and Zhang F. 2019. A comparative study of Kriging and Inverse Distance Weighting interpolation methods to estimate soil organic carbon. 18, e 00225, 2019.

[5] Cai Q., Rushton G., Bhaduri B., Bright E. and Coleman P. 2006. Estimating small-area populations by age and sex using spatial interpolation and statistical inference methods. Transactions in GIS. 10(4): 577-598.

[6] Heyang Wan, Jiang Li, Songhao Shang, Rahman. 2021. Exploratory factor analysis-based co-kriging method for spatial interpolation of multi-layered soil particle-size fractions and texture. Journal of Soils and Sediments 21(12). DOI:10.1007/s11368-021-03044-4.

[7] Li J. and Heap A. D. 2013. A review of spatial interpolation methods for environmental scientists.

[8] Mengmeng Wang, Guojin He, Zhaoming Zhang, Guizhou Wang, Zhengjia Zhang, Xiaojie Cao, Zhijie Wu, Xiuguo Liu. 2017. Comparison of Spatial Interpolation and Regression Analysis Models for an Estimation of Monthly Near Surface Air Temperature in China. 9(12): 1278; doi.org/10.3390/rs9121278.

[9] Menafoglio A., Guadagnini A. and Secchi P. 2014. A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. DOI:10.1007/s00477-014-0849-8, Springer, Stochastic Environmental Research and Risk Assessment. 28(7).

[10] Han Zhang, You Tian, Zhao. 2023. Dispersion Curve Interpolation Based on Kriging Method. Applied Sciences 13(4):2557. DOI: 10.3390/app13042557, License: CC BY 4.0.

[11] Jinkyung Yoo, Zequn Sun, Qin Ma, Young Min Kim. 2022. A Guideline for the Statistical Analysis of Compositional Data in Immunology. License: CC BY 4.0.

[12] Filzmoser P., Hron K. and Thompson K. 2012. Linear regression with compositional explanatory variables. Journal of Applied Statistics. 39(5): 1115-1128.

[13] C. Ozgur, T. Colliau, G. Rogers, Z.Hughes and B. Myer-Tyson. 2017. MatLab vs. Python vs. R. Journal of data science: JDS. 15(4).

[14] Franke J. and Theis N. 2013. Comparison of regression and interpolation methods for mapping groundwater quality parameters. Environmental Monitoring and Assessment. 185(6): 4685-4702.