Qualitative Regionalization of Bay of Bengal

R. Venkat Shesu[1*], M. Ravichandran[2], B. Venkateswara Rao[3], Venkata Jampana[1]
[1]Indian National Center for Ocean Information Services, India
[2]National Center for Polar and Ocean Research, India
[3]Jawaharlal Nehru Technological University, India
[*]Corresponding Author:- venkat@incois.gov.in

ABSTRACT

Clustering is an essential statistical technique to find subgroups within a dataset. Clustering helps to uncover hidden relationships among the observational variables in the data. Many researchers have applied clustering techniques across a number of knowledge domains with success. Here we focus on application to ocean science, more particularly to the north region of the Indian Ocean, i.e., the Bay of Bengal. In this article, Regionalization is carried out for the Bay of Bengal to group the regions with similar characteristics using a graph-based clustering technique called SKATER (Spatial 'K'luster Analysis by Tree Edge Removal). These subgroups are consistent with observations that characterize the near-surface features of the Bay of Bengal.

Key words: Spatial clustering, Regionalization, SKATER, Bay of Bengal.

1. INTRODUCTION

In a very general setting, it is of value to find organized structure in seemingly unrelated data. It helps to reduce the complexity of the problem and makes it more tractable. Identifying regions of similar structure helps provide more insight into the internal dynamics and hidden relationships among the variables. Several useful statistical techniques allow finding similarities and differences among the data variables. Across many domains, researchers were used supervised and unsupervised methods of data mining, machine learning, and deep learning techniques, such as weather forecasting [1]. Among these, data clustering is one of the widely used techniques to find subgroups in the data. In data clustering, we partition the data into distinct groups using a set of rules and associations and seek to see similar characteristics in each grouping. Background knowledge on the problem helps partition the data and provides additional insights on hidden relationships among the variables of the multivariate dataset. This clustering comes under the unsupervised classification problem class as we do not know about the distinct clusters in the data. Clustering helps to discover homogenous subgroups among a set of observations.

The focus of this manuscript is on the application of clustering techniques to the oceanography domain. Previously, deep learning object detection techniques have been used for automated tree identification on the shoreline from LiDAR data [2]. Here, we are demonstrating the data mining application using cluster analysis for the ocean water data. To the naive, oceans seem like vast expanses of water masses. However, progress in the science of oceanography shows that distinct water masses have a nearly uniform temperature, salinity, and density. Here data from the world ocean atlas is used to study the Bay of Bengal, which is landlocked from three sides and influenced by the Indian Ocean from the South. The rivers of the subcontinent drain off huge quantities of freshwater from the North and the east coast of Indian. From the west, the Indonesian throughflow brings different types of waters into the Bay [3]. This seemingly homogenous Bay is analyzed to group the water masses of similar spatial characteristics through Regionalization.   Spatially constrained multi-variate clustering to the method to group small spatial units into a contiguous region of homogenous properties is known as Regionalization [4], [5]. The primary outcome of this study is to delineate regions of similar water masses, which may help in decision-making for the deployment of various oceanographic platforms. In this study, Regionalization is carried out for the Bay of Bengal to group the regions with similar characteristics using a graph-based clustering technique called SKATER (Spatial 'K'luster Analysis by Tree Edge Removal)[6].   SKATER is an efficient regionalization technique that uses a minimum spanning tree consisting of a connected tree with no circuits.

In this paper, an approach for the Regionalization of the Bay of Bengal is briefly described. Section 2 briefly describes the water masses of the Bay of Bengal and the data used for this study. Section 3 gives the details about the algorithm used,

i.e., Regionalization using SKATER [6]. Section 4 describes the results of the SKATER for Regionalization. The main conclusions from this study are given in section 5.

2. DATA

The data for this analysis is taken from the World Ocean Atlas (WOA). This data product is released by the ocean Climate Laboratory of the National Oceanographic Data Center [7]. In-situ observations of essential climate variables in the WOA atlas are those of ocean temperature, salinity, dissolved oxygen, and a number of other variables. They are released as data products approximately every four years [7]. These atlases are very useful to weather and climate related studies. The WOA dataset is made up of objectively analyzed global grids at 1° spatial resolution. The data spans across the surface to the seafloor over 33 standardized vertical intervals in terms of spatial and vertical resolution. On temporal scales, the data is available over annual, seasonal, and monthly time scales. The dataset also has several statistical fields like standard errors and standard deviations, which help get insights into the natural variability of the climate and weather. The data is available in both compressed ASCII and netCDF formats.

The Bay of Bengal and the Arabian Sea are often viewed as two contrasting branches of the Indian Ocean. These water masses profoundly affect the regional climate of the Indian subcontinent. This article focuses on the Bay of Bengal, which has a steady influx of freshwater from the Ganges-Brahmaputra. Decades of observations using buoys, research ships, satellite imagery, and climate models show the Bay of Bengal (henceforth the Bay) as an agglomeration of distinct water masses [8].

The Southwest monsoon and the Northeast monsoon are two primary forcing factors that affect the Bay circulation. Other important drivers of the circulation and mixing in the Bay are the East India Coastal Current (EICC) which follows the east coast of India. Similarly, on the western side, the Rossby and Kelvin waves influence the dynamics and mixing. In the southern region of the Bay, there is the subtropical anticyclonic gyre which controls the mixing and circulation patterns. The water masses partly influence the regions of the Bay near the Andaman and Nicobar Islands from the Indonesian throughflow. Part of the inhomogeneity of the water masses is controlled by the so-called barrier layers in the Bay, which does not allow free mixing of the surface waters with the interior water masses. This constrained mixing is one reason for observing water masses of different properties in the Bay [9].

The near-surface density ρ of the water masses of the Bay is affected by both the temperature (T) and salinity (S) variables,

which need not follow a simple linear T-S correlation. In general, the horizontal gradients of temperature and salinity of the water masses compensate each other over the oceans, keeping the balance across the horizontal density gradients. Several processes like strong winds, solar insolation, and ocean currents influence density compensation. All these can be taken as potential categorical variables influencing the overall circulation of the Bay. However, the focus of this article is only on temperature and salinity data. The data fields are monthly climatologies from the WOA. For example, the climatology over November implies composite all previous November month data extracted and objectively analyzed from the data. These climatologies are constructed for the whole Bay of Bengal for all months using data from many sources [7].

The temporal gradients of temperature and salinity climatologies are computed by differencing successive months of data. The positive and negative gradients obtained by the time differencing are converted to categorical (binary) fields related to increasing and decreasing parameters. The differencing is based on the comparison with the previous month.
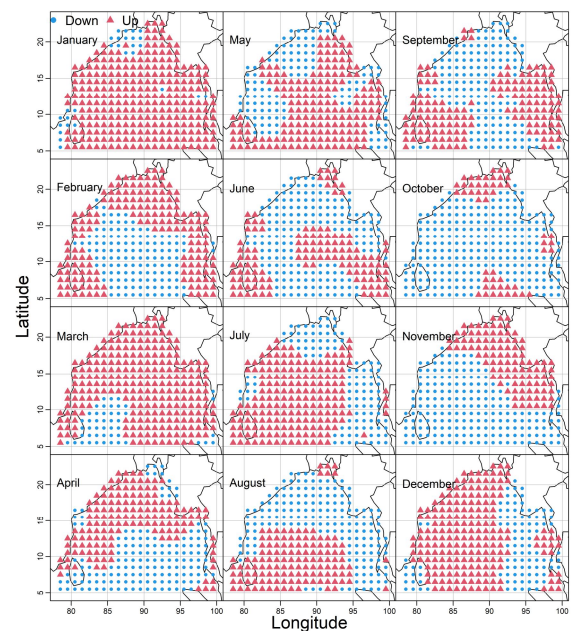
Figure 1: Categorization of salinity data over the Bay of Bengal
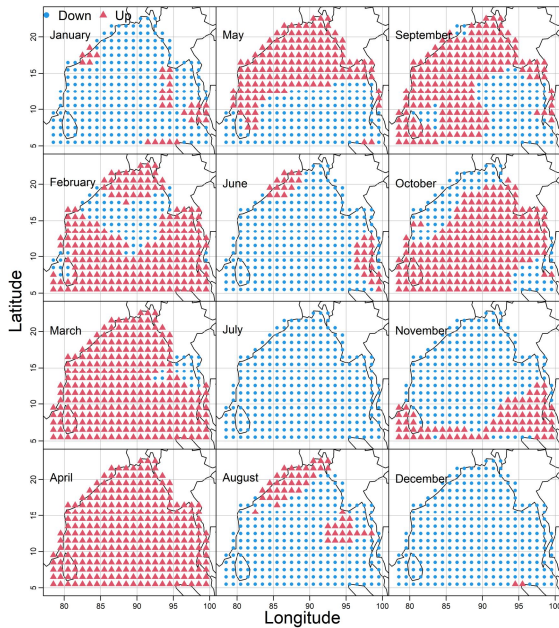
Figure 2: Categorization of temperature data over the Bay of Bengal

Figure 1 shows the parameter plot of salinity across all the twelve months of the year, with a legend showing regions of increasing (up) and decreasing (down) data compared with the corresponding previous month.

Figure 2 shows a similar plot for temperature. Based on the categorization presented in Figures 1 and 2, the binary dissimilarity matrix is computed for these two datasets.

3. REGIONALIZATION USING SKATER

Assunçáo et al., 2006 introduced a regionalization method using graph partitioning for efficient division of spatial objects while preserving neighborhood connectivity [6]. The SKATER builds off of a connectivity graph to represent spatial relationships between neighboring areas, where a node and edges represent each area represent connections between regions. Edge costs are calculated by evaluating the dissimilarity between adjacent regions. The connectivity graph is reduced by pruning edges with higher dissimilarity until we are left with n nodes and $n-1$ edges. At this point, any further pruning would create subgraphs, and these subgraphs become cluster candidates. SKATER first makes an undirected weighted graph $G(V,E)$ based on the node adjacency, where V is the set of nodes and E is the set of Edges. Given $G(V,E)$, edge weight $w:E \rightarrow R^+$ is defined with the dissimilarity of two connected nodes. Minimum Spanning Tree (MST) is then obtained by pruning edges to minimize the sum of dissimilarities over all nodes through Prims's algorithm [10]. Finally, the MST is partitioned into a

set of sub-trees that minimizes the within-cluster sum of squares. In our analysis, the V is the set of binary values representing the increasing and decreasing trend of each grid, and E is connectivity to adjacent grids. In brief, SKATER is performed in three steps:

i) Creation of Connectivity Graph with neighborhood relationship including calculation of distance.
ii) Creation of minimum spanning tree (MST).
iii) Partitioning of the MST by successive removal of edges.

The *spdep* package [11] in R is used to implement the SKATER algorithm for this study.

4. RESULTS

As data from WOA is available at 1° spatial resolution, each grid is going to be considered as a node. The connectivity graph $G(V,E)$ is constructed with V(vertices) as each grid and E(edges) as the connection between adjacent grids with the weight, which is the binary distance between the grids. Figure 3 shows all the connected components of the data, i.e., a connected graph with spatial relationships between the neighboring grids.
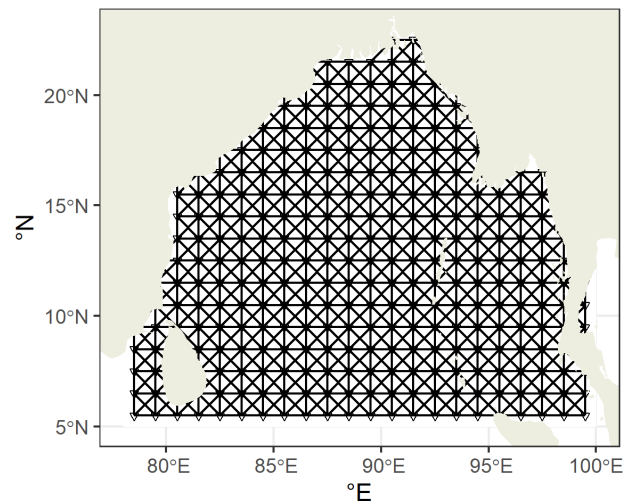
Figure 3: Connectivity Graph

In the second step, the minimum spanning tree (MST) is created on the connectivity graph $G(V,E)$ by pruning edges to minimize the sum of dissimilarities. Figure 4 represents the minimum spanning tree constructed over the graph illustrated in figure 3.
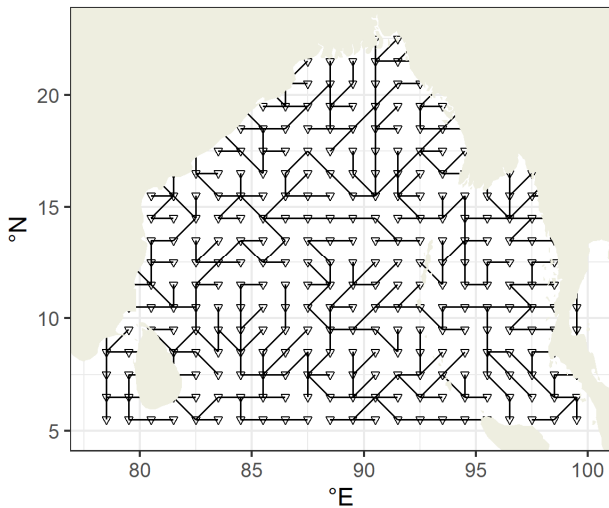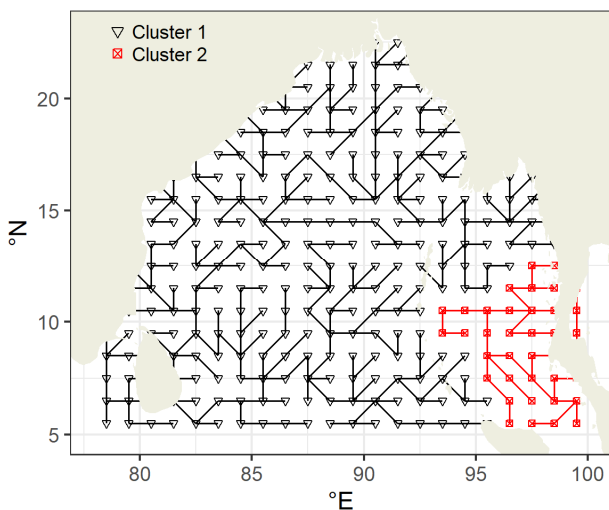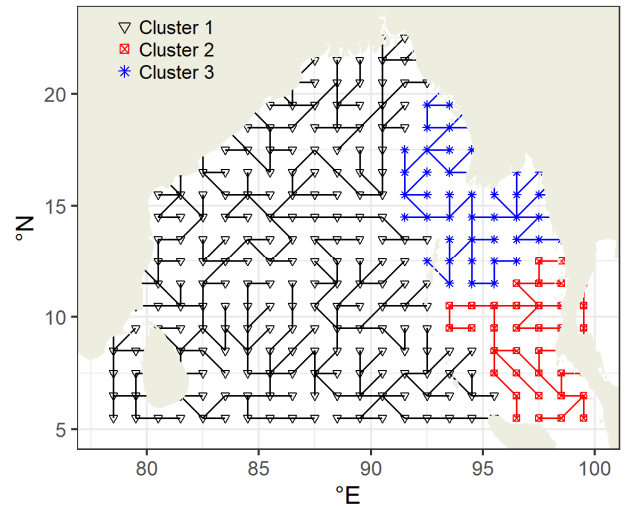
Figure 4: Minimum Spanning Tree

Figure 6: Two edge removal for the formation of three clusters

In the final step, the MST is partitioned into n-1 of groups by successive removal of edges. Five clusters were selected heuristically for this study. It depends on the user choice, which the geographical setting and operational requirements will dictate. For example, the cluster number should be five for deploying five moored buoys, and four edge removals are required. Thus by consecutive removal of four edges leads to the formation of the five clusters.
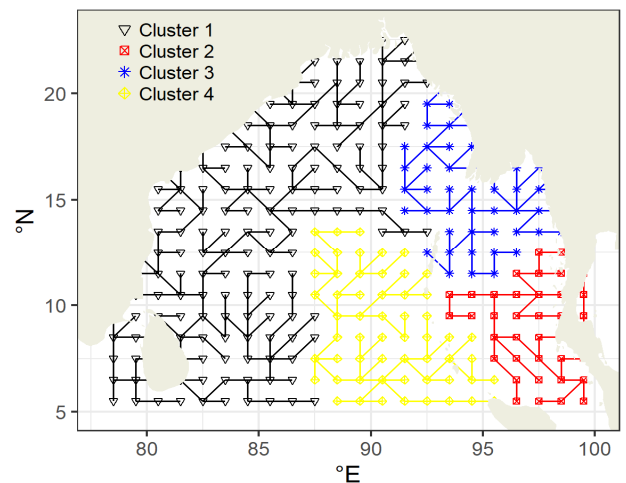
Figure 5 represents the Regionalization of the Bay by removing the most prominent edge. This edge removal divides the water mass, mainly affected by Indonesian throughflow and the Andaman Sea. Figure 6, 7 and 8 represents the removal of 2, 3 and 4 edges respectively.

Figure 7: Three edge removal for the formation of four clusters

Figure 5: One edge removal for the formation of two clusters
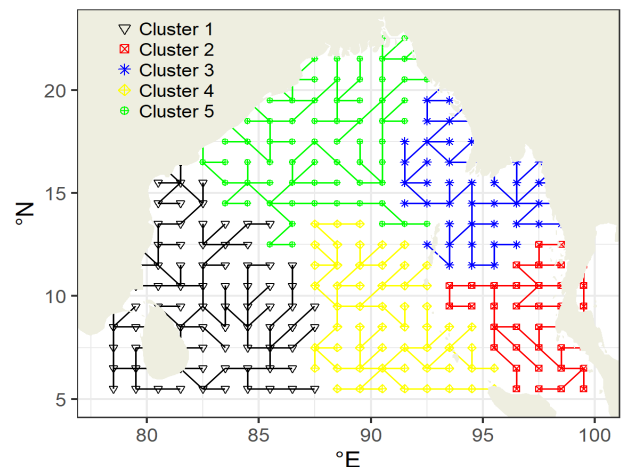
Figure 8: Four edge removal for the formation of five clusters

The monthly central tendency for five regions 1, 2, 3, 4, and 5 with 84, 39, 49, 57, and 75 grid points, respectively, are given in Tables 1 to 5. "Temp ↑" represents the percentage of points in a given cluster where the temperature is increasing. "Temp ↓" means the percentage of points in a given cluster

where the temperature is decreasing. Similarly, "Sal ↑" represents the percentage of grid points in a given cluster where salinity is increasing, and "Sal ↑" represents the percentage of points where salinity is decreasing.

Table 1: Cluster 1 with 84 grid points (in the percentage of grid points)

Month	Temp ↑	Temp ↓	Sal ↑	Sal ↓
January	0.00	100.00	88.10	11.90
February	89.29	10.71	58.33	41.67
March	100.00	0.00	52.38	47.62
April	100.00	0.00	47.62	52.38
May	42.86	57.14	46.43	53.57
June	0.00	100.00	50.00	50.00
July	0.00	100.00	89.29	10.71
August	1.19	98.81	76.19	23.81
September	84.52	15.48	80.95	19.05
October	84.52	15.48	0.00	100.00
November	30.95	69.05	0.00	100.00
December	0.00	100.00	80.95	19.05

Table 2: Cluster 2 with 39 grid points (in the percentage of grid points)

Month	Temp ↑	Temp ↓	Sal ↑	Sal ↓
January	30.77	69.23	94.87	5.13
February	94.87	5.13	84.62	15.38
March	94.87	5.13	84.62	15.38
April	100.00	0.00	25.64	74.36
May	12.82	87.18	58.97	41.03
June	51.28	48.72	84.62	15.38
July	0.00	100.00	20.51	79.49
August	5.13	94.87	10.26	89.74
September	17.95	82.05	92.31	7.69
October	43.59	56.41	10.26	89.74
November	89.74	10.26	28.21	71.79
December	0.00	100.00	41.03	58.97

Table 3: Cluster 3 with 49 grid points (in the percentage of grid points)

Month	Temp ↑	Temp ↓	Sal ↑	Sal ↓
January	18.37	81.63	100.00	0.00
February	65.31	34.69	79.59	20.41
March	65.31	34.69	100.00	0.00
April	100.00	0.00	65.31	34.69
May	71.43	28.57	65.31	34.69
June	0.00	100.00	24.49	75.51
July	0.00	100.00	42.86	57.14
August	28.57	71.43	4.08	95.92
September	57.14	42.86	65.31	34.69
October	100.00	0.00	4.08	95.92
November	0.00	100.00	100.00	0.00
December	0.00	100.00	8.16	91.84

Table 4: Cluster 4 with 57 grid points (in the percentage of grid points)

Month	Temp ↑	Temp ↓	Sal ↑	Sal ↓
January	8.77	91.23	100.00	0.00
February	80.70	19.30	0.00	100.00
March	100.00	0.00	98.25	1.75
April	100.00	0.00	0.00	100.00
May	5.26	94.74	100.00	0.00
June	0.00	100.00	40.35	59.65
July	0.00	100.00	80.70	19.30
August	1.75	98.25	87.72	12.28
September	38.60	61.40	5.26	94.74
October	91.23	8.77	31.58	68.42
November	31.58	68.42	0.00	100.00
December	3.51	96.49	70.18	29.82

Table 5: Cluster 5 with 75 grid points (in the percentage of grid points)

Month	Temp ↑	Temp ↓	Sal ↑	Sal ↓
January	8.00	92.00	78.67	21.33
February	48.00	52.00	65.33	34.67
March	98.67	1.33	100.00	0.00
April	100.00	0.00	89.33	10.67
May	100.00	0.00	42.67	57.33
June	18.67	81.33	18.67	81.33
July	0.00	100.00	56.00	44.00
August	45.33	54.67	12.00	88.00
September	90.67	9.33	10.67	89.33
October	44.00	56.00	34.67	65.33
November	0.00	100.00	56.00	44.00
December	0.00	100.00	100.00	0.00

5. CONCLUSION

In this paper, we used a SKATER (Spatial 'K'luster Analysis by Tree Edge Removal) for the Regionalization of the Bay of Bengal. It produces good partitions in a fast manner. This study can be helpful for a selection of appropriate locations for deploying a new oceanographic platform and also to decide the optimized number of the platform to deploy. This study shows that using a SKATER algorithm exists five distinct regions of the Bay of Bengal from long-term climatological surface temperature and salinity data. Further analysis can be taken up where other oceanographic differences could be highlighted using various other parameters and vertical profiles.

REFERENCES

1. Bhawsar, M., Tewari, V. and Khare, P. **A Survey of Weather Forecasting based on Machine Learning and Deep Learning Techniques**, *International*

Journal of Emerging Trends in Engineering Research, Vol. 9(7), 2021.

2. Alon, A.S., Festijo, E.D. and Casuat, C.D. **Tree Extraction of Airborne LiDAR Data Based on Coordinates of Deep Learning Object Detection from Orthophoto over Complex Mangrove Forest**, *International Journal of Emerging Trends in Engineering Research*, Vol. 8(5), 2020.

3. Schott, F.A. and McCreary Jr, J.P. **The monsoon circulation of the Indian Ocean**, *Progress in Oceanography*, Vol. 51(1), pp.1-123, 2001.

4. Guo, D. **Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP),** *International Journal of Geographical Information Science*, Vol. 22(7), pp.801-823, 2008.

5. Kugler, T.A., Manson, S.M. and Donato, J.R. **Spatiotemporal aggregation for temporally extensive international microdata**, *Computers, environment and urban systems*, Vol. 63, pp. 26-37, 2017.

6. Assunção, R.M., Neves, M.C., Câmara, G. and da Costa Freitas, C. **Efficient regionalization techniques for socio□economic geographical units using minimum spanning trees**, *International Journal of Geographical Information Science*, Vol. 20(7), pp.797-811, 2006.

7. Locarnini, M., Mishonov, A.V., Baranova, O.K., Boyer, T.P., Zweng, M.M., Garcia, H.E., Seidov, D., Weathers, K., Paver, C. and Smolyar, I. **World ocean atlas 2018, Vol. 1: Temperature**, 2018.

8. Sengupta, D., Bharath Raj, G. N., and Shenoi, S. S. C. **Surface freshwater from Bay of Bengal runoff and Indonesian Throughflow in the tropical Indian Ocean**, *Geophysical Research Letters*, Vol. 33 (22), 2006.

9. Thadathil, P., Muraleedharan, P.M., Rao, R.R., Somayajulu, Y.K., Reddy, G.V. and Revichandran, C. **Observed seasonal variability of barrier layer in the Bay of Bengal**, *Journal of Geophysical Research: Oceans*, Vol. 112(C2), 2007.

10. Prim, R.C. **Shortest connection networks and some generalizations**, *The Bell System Technical Journal*, Vol. 36(6), pp.1389-1401, 1957.

11. Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A. and Blanchet, G. **Package 'spdep'**, *The Comprehensive R Archive Network*, 2015.