Outlier Detection and Removal with Spatial Algorithm in Agriculture Data for Davangere Region

Arun Kumar G H[a] , Nirmala C R[b]
[a]Department of Computer Science and Engineering, Bapuji Institute of Engineering And Technology, Davanagere-577004, Karnataka, India
[b]Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere-577004, Karnataka, India

Abstract

Precision agriculture incorporates the deployment of farm equipment, principles in on-farm experiments and emerging technologies to assess and supervise temporal and spatial variability associated with the facets of agricultural production with the aim of improving cultivation quality and crop yield performance. Massive amount of agricultural data generated using sensors can be used for crop management decisions. Such data generated by the sensors may contain values of Soil physico-chemical parameters (n,p,k, pH and EC) in the form of spatial or temporal data and may possess inconsistent data observations that deviates crop management assessments. In this Research work, we mainly deal with the detection and removal of such outliers in the dataset prepared for Davangere Agricultural jurisdiction by applying Spatial Iterative-R algorithm. We also consider Iterative-R algorithm to reduce the effect of swamping and masking. The algorithm is checked for its efficiency in identifying the outliers and reducing the masking and swamping effects. The implementation is demonstrated with R-Studio/ R-Programming Language.

KEYWORDS – Precision Agriculture, Iterative-R, KNN, Outlier Detection, Swamping, Masking

I. INTRODUCTION

Precision Agriculture is a modern farming approach intends to bring the agricultural world towards a high efficiency with low physico-chemical inputs and sustainable agriculture using emerging techniques. The precision agriculture mainly includes the Global Positioning System (GPS), Soil sampling, Digital soil mapping, Management zones, Yield mapping and Variable rate technology [10] for the possible developments in crop yield performance. Soil sampling specifies the soil properties to define Management Zones (MZ) to apply inputs. Digital Soil Mapping (DSM) incorporates quantitative methods to process and interpret information on soils. Variable Rate Technology (VRT) allows agricultural inputs such as fertilizers (nitrogen, potassium and phosphorous), pesticides, seeding to be applied on-the-go throughout the field at appropriate rates according to the pre-set application map. Yield mapping [2] is the Process of collecting geo-referenced crop yield data to provide the idea of supervising the crop moisture and the soil fertility. Expert filtering software is used to remove yield errors present in the crop yield data collected. These software programs also reduces systematic / stochastic errors (spatial outliers) [10, 12] from unknown/known sources. For Site-Specific Crop Management, it is necessary to detect errors in the yield data. Datasets of crop yield may contain many errors raised from known and unknown sources which are classified as measurement, management and natural errors. All of these errors greatly impacts yield measurements by creating unrealistic measurements [3] [6].
Outliers are the observations those deviates so much from other observations. In a spatial-data context, Outlier may be any spatially referenced data-point/observation [13] whose non-spatial-attribute values are remarkably distinct from other spatially referenced data-observations in its spatial-proximity. Outliers, Usually in spatial-data-context, are collective and point. Many statistical-techniques are proposed to detect and remove outliers [4, 5, 7]. These techniques include Spatial-outlier-tests such as Graphical and Quantitative-approaches [10]. Graphical-outlier-detection methods like Moran-Scatterplot and Variogram-Cloud involves the visualization of spatial-data in order to analyze the outliers. The

Quantitative-Statistical-tests such as linear-regression/Scatterplot and Spatial-Statistic-Z distinguishes spatial-outliers from the rest of the dataset. Further, in the context of finding spatial outlying observations, Swamping (failing to identify outliers) and masking (mistaking clean observations for outliers) effects [9, 10, 15] are considered to be minimize and the Spatial-outlier detection techniques such as Iterative-R, Iterative-Z and Median-Z [10, 14] can be served for this purpose. Even-though, Iterative-R[14] was proposed few years -ago with extensive Background-work, ahead of its time, the utilization of such algorithm specifically intended for outlier-detection while analyzing the spatial-data for precision-farming has been remained as remarkably new concern.
In this paper, we propose to use Iterative-Ratio(R) [14] algorithm as a baseline for the detection of outliers and further the comparison of efficiency of Iterative–R with semivariogram in dealing with masking and swamping effect is made. The work carried-out is aimed at investigating the spatial-outlier-detection-methods (both the Statistical and Graphical-approach) by comparing their ability to deal with masking and swamping effects.

We performed Iterative-Ratio (R) [14] algorithm for the detection and removal of spatial outliers. Our research work mainly involves the three phases in the analysis of outlier detection algorithms. As a first step, in the Data Acquisition phase, spatial data observations that contains outliers / errors from many known/unknown sources are collected; Second, Data Preprocessing is the phase that involves Data cleaning, Data transformation and Classification; Lastly, We perform Iterative-Ratio (R) algorithm by applying it on the Pre-processed dataset to detect, remove outliers and to reduce the swamping and masking effects. The complete work is developed using R-Studio and R-language. The rest of the paper has the explanation on the Performed Iterative-Ratio (R) Outlier- detection-algorithm in section II with Experimental results described in section III. Concluding remarks and Future work are given in section IV and V respectively.

II. WORK DESCRIPTION

2.1 Iterative-Ratio (R) Algorithm for Outlier Detection

We are Considering Iterative-Ratio (R) [14] spatial-outlier-detection-algorithm on the pre-processed dataset. In each iteration, the algorithm finds only one inconsistent / nonconforming (very large or very small value in that range) value which can be considered as to be a spatial-outlier. First, for every data-observation / data-point, we calculate the ratio-value- which is the ratio of a data-point's attribute-value and the average-attribute value of its neighbors and then the data-point that has the largest R-value is identified as an outlying observation for that range. Next, we substitute the attribute-value of this detected spatial-outlier value with the average-attribute-value of its neighbors.

2.1.1 Problem Description:

We are given with a spatial-data-observation set of points S = {s_1,s_2,s_3,.....s_n}. Let $f_{attr}(\,)$ be the function that denotes the attribute-values of every spatial-data-point in a space with dimension ϱ ≥ 1. This $f_{attr}(\quad)$intrinsically is a function that also maps the attribute-values from S to R, where R is the set of real-number. $f_{attr}(S_i)$ is an attribute-function that represents the attribute-value of spatial-data-point 'S_i'. For each given spatial-data-point 'S_i', k-nearest-neighbors of data-point 'S_i' are denoted by $NN_k(S_i)$, where $k = k(S_i)$ depends on the value of S_i, for i= $1, 2, \ldots , n$. Next, we have$f_{aggr}(\quad)$, which is the neighborhood-function, considered as a mapping from S to R, so that for every individual 'S_i', a summary-statistic of attribute-values of all the spatial-data-points inside $NN_k(S_i)$is returned by $f_{aggr}(S_i)$. For example, $f_{aggr}(S_i)$ can be the average-attribute-value of the k-nearest-neighbors of a spatial-data-point 'S_i'. In order to detect spatial-outliers, we perform the comparison of the attribute-value of every spatial-data-point 'S_i' with those attribute-values of its neighbor-spatial-data-points $NN_k(S_i)$. This comparison can be performed by a comparison-function$f_{ratio}(\quad)$, which is a function of $f_{attr}(\quad)$and $f_{aggr}(\quad)$.Here, we consider $f_{ratio}(\quad)$as the ratio$f_{attr}(\,)/f_{aggr}(\,)$. Further, Let r_i = $f_{ratio}(S_i)$for i =$1, 2, \ldots , n$. Given the attribute function $f_{attr}(\,)$, function k, neighborhood function

$f_{aggr}(\ )$ and comparison function $f_{ratio}(\ )$, a spatial-data-point 'S_i' is considered as a spatial-outlier, if 'r_i' is an extreme-value (very large or very small value in that range) in the set $\{r_1, r_2, r_3, \ldots\ldots r_n\}$.We consider Iterative-Ratio (R) algorithm to find the outliers in spatial-data context. We assume all $k(S_i)$ are equal to a fixed-number 'k'. Further, the algorithm can be generalized by replacing fixed 'k' with dynamic $k(S_i)$. The neighborhood function $f_{aggr}(\ )$ evaluated at a spatial point 'S_i' is taken to be the average attribute value of all the k-nearest neighbors of 'S_i'. Comparison function $f_{ratio}(S_i)$ is taken to be the ratio of $f_{attr}(S_i)$ to $f_{aggr}(S_i)$. For $f_{ratio}(S_i)$,if the value is very large or very small (detected by the given threshold 'θ') then it is an indication that 'S_i' might turn out to be a Spatial-outlier.

Input: 1. A Spatial Dataset of various farm-fields with coordinates of soil sample values collected Across Davangere region.
2. An attribute function $f_{attr}(\,)$.
3. A Number '***k***' of nearest-neighbors and an expected number '***m***' of spatial-Outliers.

Output: Spatial-Outlier-Detection based upon sample-value of referred coordinate and Crops Coordinate.

Iterative- Ratio Algorithm:

1. For every spatial-data-point 'S_i', calculate the k-nearest-neighbor-set $NN_k(S_i)$, the neighborhood-function $f_{aggr}(s_i) = \frac{1}{k}\sum_{S \in NN_k(S_i)} f_{attr}(s)$ and the comparison-function $r_i = f_{ratio}(S_i) = \frac{f_{attr}(s_i)}{f_{aggr}(s_i)}$.
2. Let r_q or r_q^{-1} denote the extreme of $r_1, r_2, r_3, \ldots r_n, r_1^{-1}, r_2^{-1}, \ldots r_n^{-1}$, where $r_q = f_{ratio}(S_q)$ or $r_q^{-1} = f_{ratio}^{-1}(S_q)$. For a given threshold θ, if r_q or $r_q^{-1} \geq \theta$, treat the Corresponding $'S_q'$ as a Spatial-outlier.
3. Update $f_{attr}(S_q)$ to be $f_{aggr}(S_q)$. For each spatial-point $'S_i'$ whose $NN_k(S_i)$, contains S_q, Update $f_{aggr}(S_i)$. And r_i i.e $f_{ratio}(S_i)$.
4. Repeat steps 2 and 3 until either total number of S-outliers equals 'm' or the threshold condition is not met.

III. Experiment and Result

The Research work first focuses on the preparation of dataset of Davangere region with Soil parameters like Nitrogen (*N*), phosphorus (*P*), potassium (*K*), potential Hydrogen (*pH*) and Electrical Conductivity (*EC*) and then it applies Iterative technique such as Iterative-R algorithm to detect and remove the outliers present in the spatial data observations. Further, we reduce the Swamping and Masking effects (i.e. the risk of falsely claiming regular spatial points as outliers and ignoring true outliers) in subsequent iterations. The Real dataset consists of spatial data records from Davanagere region and concentrations of Nitrogen (*N*) and Phosphorus (*P*), Potassium (*K*) contents, potential Hydrogen (*pH*) content and sampling site latitude and longitude are used.

The Entire Davangere area lies in between latitude of 75.88 to 75.95E and longitude of 14.44 to 14.48N. We have collected below shown Soil sample values within Davanagere region based on N, P, K and pH content.

	Location	Y	X	N	P	K	Ph	EC
141	attigere(4)	14.321665	75.9943051	102.31	40.25	37.96	6.49	0.16
142	belavanur(2)	14.398330	75.9202116	107.04	22.31	91.43	6.80	0.11
143	Bilchodu	14.491642	76.1595788	64.05	14.01	82.13	8.06	0.42
144	Talavatti	14.103281	76.5579808	98.15	16.15	83.24	6.88	0.93
145	Talavatti(2)	14.102485	76.5599066	60.01	10.32	64.07	7.86	0.76
146	halebathi(2)	14.476671	75.8595142	102.31	16.75	102.31	7.36	0.54
147	halebathi(3)	14.476070	75.8591699	124.02	18.75	78.15	7.09	0.57
148	halebathi(4)	14.475790	75.8599172	114.31	18.73	94.07	6.67	0.62
149	Kod	14.514800	75.4440594	58.32	9.32	79.07	7.68	0.16
150	Kod(2)	14.514319	75.4447326	62.31	15.32	83.01	7.81	0.63
151	Sokke	14.683239	76.2519837	108.15	17.32	83.01	7.63	0.46
152	Bullapura(2)	14.386691	75.5805478	105.71	20.06	102.77	6.93	0.45

Table -1: Raw Dataset of Soil Sample with N, P, K, pH and EC Values

Soil contains large amount of organic materials. It is observed that the survey [12] made for the analysis of the physico-chemical parameters has resulted the following ranges: The availability of Nitrogen (*N*), represents the fraction of total nitrogen content absorbable by the crops and it varies between 48 to 236 kg ha^{-1}. The availability of Phosphorous (*P*), in soil represents the part of total phosphorous Susceptible to crops and this content varies between 8 to 78 kg ha^{-1}.The availability of Potassium (*K*) represents the exchange reactions of other organic matters in the Soil; the potassium content in soil varies between 9 to 312 kg ha^{-1}. Potential Hydrogen (*pH*) is the concentration of hydrogen in the soil, it measures the alkalinity of substance and the *pH* varies from 6.7 to 7.7 in the Davanagere region. The physico-chemical parameters obtained from the prepared dataset indicates that most of the soil samples are within the permissible limits for agricultural purposes and few are turned out to be outlying observations. We have observed *n,p,k*, *pH* and *EC* values for 685 locations within entire Davangere District's agricultural jurisdiction and it can be summarized that nitrogen content ranges between 7.08 to 230.10 kgha^{-1},Phosphorous content varies between 4.32 to 78.90 kg ha^{-1} and potassium content in soil varies between 5.06 to 319.56 kg ha^{-1}.

The Exploratory Data Analysis (EDA) for the acquired data has been made to study the nature of the data. We have plotted histograms and scatterplots individually, for the values of *N*, *P* and *K* that shows the regions where most of the data is concentrated. These plots also highlight those values which are deviated from the normal observations. The following histograms shows that most of the Nitrogen (*N*) values are concentrated between 50 to 130 kg ha^{-1}, Phosphorous (*P*) concentration ranges between the 5 to 24 kg ha^{-1} and Potassium (*K*) values ranges between 10 to 120 kg ha^{-1}.

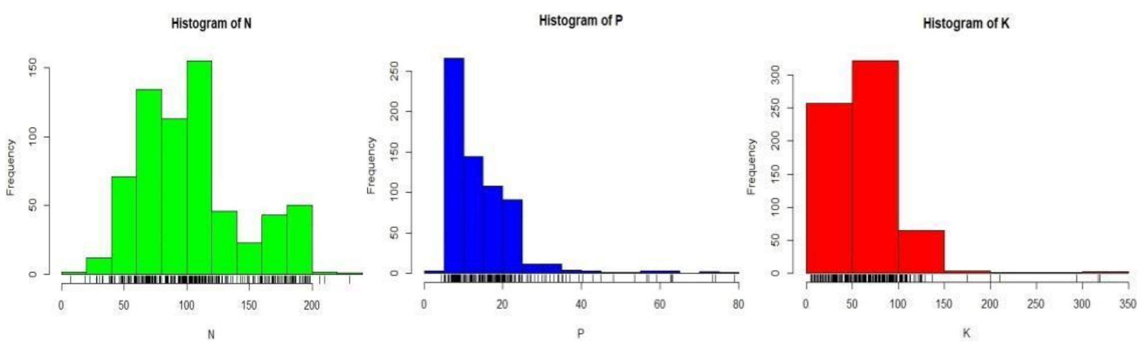


Fig-1: Histograms for 'N','P','K' Values

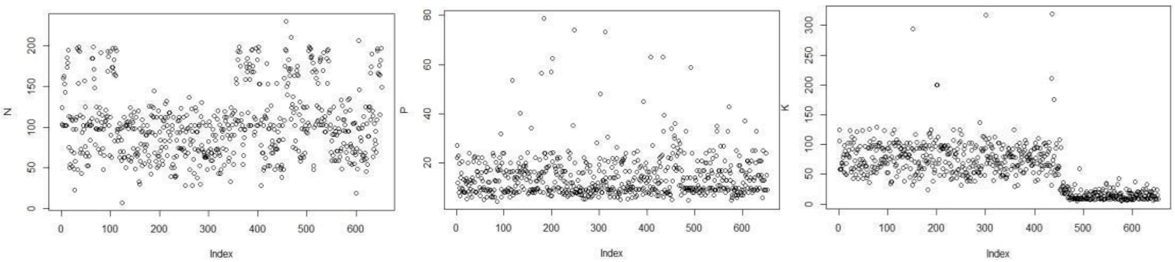


Fig-2: Scatterplots for ‘N’,’P’,’K’ Values

If, we check the conformity of each of the n,p,k values in our prepared dataset with the above mentioned standard range of n,p,k values obtained from the Survey [12], out of 652 observations, We will get 23 nonconforming values for ‘n’, 77 nonconforming values for ‘p’ and 29 nonconforming values for ‘k’.

We have performed a Graphical S-outlier method such as ‘Variogram’, which is based on the visualization of spatial data to assess outliers. A variogram is a description of the spatial continuity of the data. The Semi-variance is a measure of the spatial dependence between two observations as a function of the distance between them. The variogram plot is used to fit a model of the spatial correlation of the observations. The Semivariogram depicts the spatial autocorrelation among the measured data points. The distance where the model first flattens out is known as the range. The locations separated by distances closer than the range are spatially auto-correlated, whereas locations farther apart than the range are not. The value that the Semivariogram model attains at the range (the value on the y-axis) is called the sill. The Nugget effect can be attributed to measurement errors or spatial sources of variation at distances smaller than the sampling interval or both. Measurement error occurs because of the error inherent in measuring devices. Variation at micro-scales smaller than the sampling distances will appear as part of the nugget effect. For the values of Nitrogen (*N*), the Nugget value is 343, which indicates that there are outlying observations due to measurement error. The Sill value is 2298, beyond which there is no correlation, the variable is a purely random variable and the variogram is flat. The range value is 11361 at which the model first flattens out.

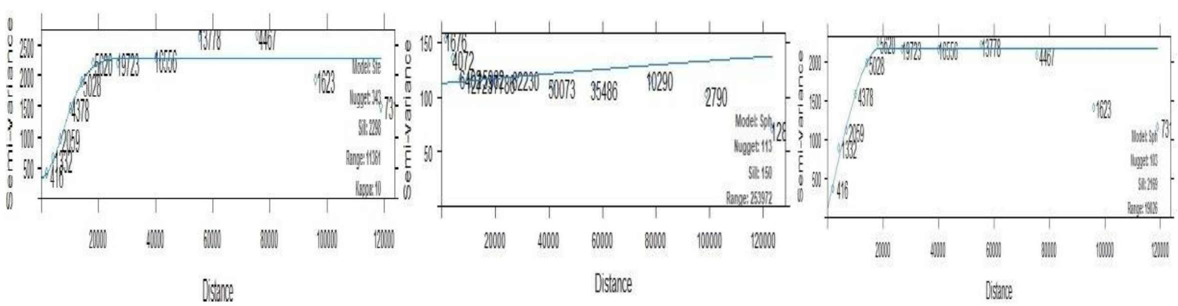
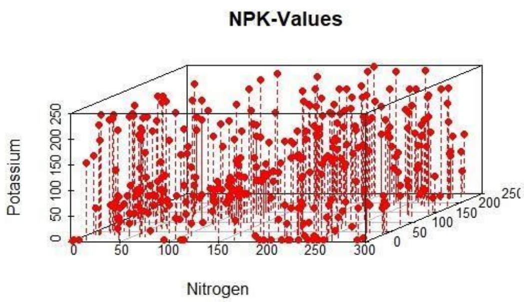


Fig-3:Semivariogram on (a) Nitrogen (b) Phosphorous (c) Potassium

For the values of Phosphorus (*P*), the Nugget value is 113. The Sill value is 150, beyond which there is no correlation and the Variogram is flat. The range value is 253972 at which the model first flattens out. For the values of Potassium (*K*), the Nugget value is 103. The Sill value is 2169, beyond which there is no correlation and the Variogram is flat. The range value is 19026 at which the model first flattens out.

We have used ScatterPlot3D to plot the graph which represents the *N*, *P*, *K* values of 652 locations of various villages from all six Taluks of Davangere district. As a first step, before applying Iterative-R (ratio) technique, we need to find nearest-neighbors for each Spatial-data-point ‘S_i’.



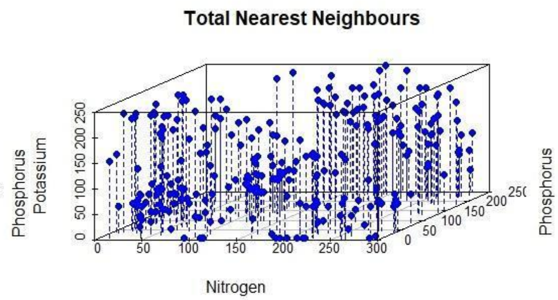


Fig-4: The N, P, K values for the Raw Dataset

Fig-5: Nearest-neighbors Identified by KNN Algorithm

We have a lot of points in a low dimensional space. So, *k-NN* is a good Choice and Euclidean is a good distance measure to use if, the input variables are similar in type. As a first step we calculate distances between every single data point 'S_i' and rest of the data-points using Euclidean-distance-formula as follows.

[1] 0.000 44.890 57.680 10.550 2.360 1.660 40.660 16.060 36.490 29.500 4.950 72.640 46.990 22.220 36.070 97.710 90.745 86.690 3.560 19.670 66.250 95.309 14.030 11.370

[1] 44.890 0.000 102.570 34.340 47.250 43.230 85.550 60.950 81.380 74.390 39.940 27.750 2.100 22.670 80.960 142.600 135.635 131.580 48.450 64.560 111.140 140.199 30.860 56.260

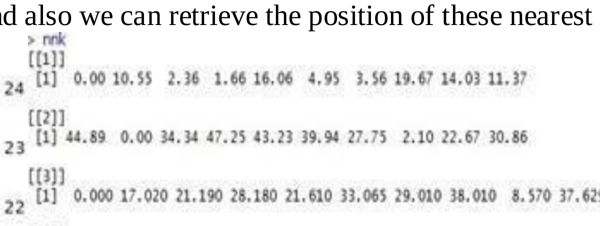
[1] 57.680 102.570 0.000 68.230 55.320 59.340 17.020 41.620 21.190 28.180 62.630 130.320 104.670 79.900 21.610 40.030 33.065 29.010 54.120 38.010 8.570 37.629 71.710 46.310

We choose the value of *k*-input to be 10, so that we can retain only those $NN_k(S_i)$points which are more near to every 'S_i' than remaining points and also we can retrieve the position of these nearest neighbors.

```
> nnkpos
[[1]]
 [1]  1  4  5  6  8 11 19 20 23 24

[[2]]
 [1]  1  2  4  5  6 11 12 13 14 23

[[3]]
 [1]  3  7  9 10 15 17 18 20 21 22
```

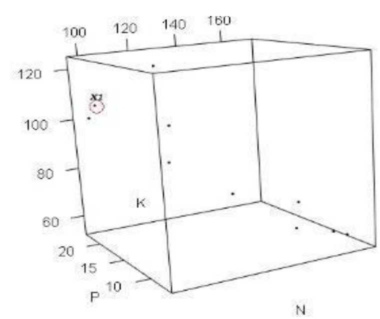


```
> nnk
[[1]]
[1]  0.00 10.55  2.36  1.66 16.06  4.95  3.56 19.67 14.03 11.37

[[2]]
[1] 44.89  0.00 34.34 47.25 43.23 39.94 27.75  2.10 22.67 30.86

[[3]]
[1]  0.000 17.020 21.190 28.180 21.610 33.065 29.010 38.010  8.570 37.62
```

(a) (b)



(c)

Fig-6: Retaining the *NNk(S_i)* Data-Points closest to 'S_i' (a) Position of Data Points (b) Distance between every 'S_i' and its closest Data Points

(c) Plot of Retained Data Points

Next, we calculate a neighborhood function $f_{aggr}(\ \ )$ which is defined as a map from S to R such that for each 's_i',
$f_{aggr}(S_i)$returns a summary-statistic of attribute-values of all the spatial-points inside $NN_k(S_i)$.For instance, $f_{aggr}(S_i)$ can be the average-attribute-value of the k-nearest-neighbors of 's_i'.

[1] 8.4210 29.3030 23.4284 11.5740 8.4210 8.6760 17.4510 12.2160 15.7830 14.7380

For the instance, for each 's_i' its summarized attribute value is assigned as attribute function $f_{attr}(S_i)$ as follows.

$$f \leftarrow (76.94\ , 91.91\ , 57.72\ , 80.46\ , 76.16\ , 77.5\ , 63.39\ , 71.59\ , 64.78\ , 67.11)$$

We compute the ratio (r-value), of a point S_i's attribute value $f_{attr}(S_i)$ and the average attribute values of its neighbors $f_{aggr}(S_i)$ for every point. i.e. Comparison function $f_{ratio}(S)$ is taken to be the ratio of $f_{attr}(S) to\ f_{aggr}(S)$.

[1] 9.136682 3.136539 2.463677 6.951788 9.044057 8.932688 3.632457 5.860347 4.104416 4.553535

Then we declare that data-point as an outlier only when it has extreme r–value *(i.e. max(r) =9.136682* at position 1) within the proximity of point 'S_1'. Next, We Substitute the attribute-value of this detected spatial-outlier with the average-attribute-value of its neighbors i.e. $f_{aggr}[s_1]$=8.421.So, $f_{attr}[s_1]$= 76.94is replaced with the value of $f_{aggr}[s_1]$which is8.421.

$$f_{attr}[max] \leftarrow f_{aggr}[max]$$

[1] 8.421 91.910 57.720 80.460 76.160 77.500 63.390 71.590 64.780 67.110

After Updating $f_{attr}(S_q)$to $f_{aggr}(S_q)$,for each and every spatial point 'S_i' whose $NN_k(S_i)$contains S_q, we will update $f_{aggr}(S_i)$and r_i. The steps are repeated until either the total number of Spatial-outliers equals number of expected Outliers *(m)* or the threshold condition is not met. Figure-7, Shown below represents all those data points

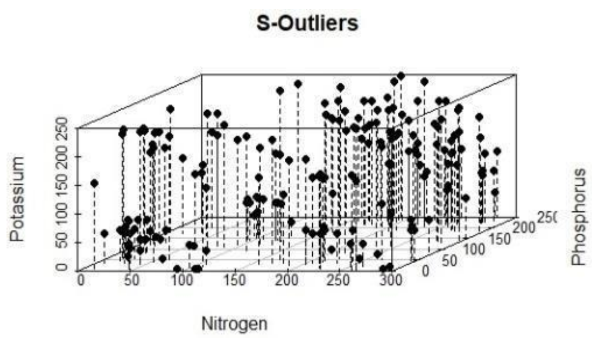


Fig-7: Outliers Removed by Iterative –R algorithm

Which are free from the outliers with reducing the effect of ignored true outliers and minimizing the influence of falsely identified outliers.

IV.CONCLUSION

Outliers can skew any analysis performed on the dataset. Considering outliers as true observations and mistaking clean observations to be outliers are masking and swamping effects. So, it is necessary to reduce masking and swamping effects along with the detection and removal of inconsistent observations from the dataset. In this Research work, we have made an attempt to find the outliers using Iterative–R algorithm to reduce the Masking and Swamping effects. Iterative–R (ratio) algorithm works in multi-iterations i.e. Iterative–R algorithm identifies only one outlier in each iteration, so that this outlier will not impact the subsequent iterations negatively. We performed a graphical spatial-outlier method i.e. Semivariogram, which assess outliers, based on the visualization of spatial data. The Semivariogram considers most of those N,P and K values as nonconforming which fall below the nugget value. But these values are actually true observations and may comply with normal inlying observations. Further, the data-points which are free from outliers analyzed using Iterative-R are more in number and the data-points free from outliers, identified by the graphical outlier methods such as Semivariogram are less. From this, it is illustrated that masking and swamping effects are reduced by our implementation. The values that have no correlation between the neighbors are found as outliers. Our work has shown that among those 680 observations and considering the actual ranges of N, P and K, only 214 values are found as conforming values.

REFERENCES

[1] B.A.Sabarish & R.Karthi, "Spatial Outlier Detection Algorithm for Trajectory Data"-International Journal of Pure and Applied Mathematics Volume 118 No.7,325-331 ISSN: 1311-8080; ISSN:1314-3395 -2018

[2] Andrei B. B. Torres, Jos´e Adriano Filho, Atslands R. da Rocha, Rubens Sonsol Gondim, Jos´e Neuman de Souza-"Outlier detection Methods and sensor data fusion for precision agriculture"- July 2017

[3] Jharna M, Sneha N, Shilpa A -"Analysis of agriculture data using data mining techniques: application of big data", Big Data -2017

[4] Mrs.NitaM.Dimble and Mr.BharatTidke, "A Framework for Outlier detection in Geographic Spatial data", International Journal in Foundations of Computer Science & Tech, (IJFCST), Vol.5, No.2, March 2015.

[5] T. Jagadeeswari, N.Harini, "Identification of outliers by Cook's distance in agriculture Datasets", International Journal of Engineering and Computer Science ISSN: 2319-7242 Volume 2 Issue 6 June 2013

[6] Liying Cao, HelongYu,,Li Ma, Guifen Chen, "Research Fertility Spatial-temporal Evolution Pattern and Crop Precision Fertility Based on Spatio-Temporal Data Mining"- 2nd International Conference on Computer Science and Network Technology -2013

[7] V.S. Jagadeeswaran , P. Uma,"Detection Of Noise By Efficient Hierarchical Birch Algorithm For Large Data Sets", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 2, February 2013

[8] Andrew Wright, Darren Hudson, Maria Mutuc,"A Spatial Analysis of Irrigation Technology-Natural Resources"-307-318 http://dx.doi.org/10.4236/nr.2013 44037 August 2013

[9] Robert Serfling, ShanshanWang,"General Foundations for Studying Masking and Swamping Robustness of Outlier Identifiers", Nov- 2012

[10] Peter Chu Su,"Statistical Geocomputing: Spatial Outlier Detection in Precision Agriculture" - Waterloo, Ontario, Canada, 2011

[11] Basavaraja S, Chandrashekarappa K.N, Anil N Patel, "Soil Quality Status of Tunga Left Bank Command area,Shimoga&Davanagere Districts, Karnataka", Global Journal of Human Social Science,Volume 11 Issue 2 Version 1.0 March 2011, ISSN: 0975-587

[12] ShashiShekhar, RangaRajuVatsavai, Mete Celik, "Spatial and Spatiotemporal Data Mining: Recent Advances"-2008

[13] S. Subramaniam U C Riverside, T. Palpanas, D. Papadopoulos, V. Kalogeraki, D. Gunopulos, "Online Outlier Detection in Sensor Data Using Non-Parametric Models", University of California, Riverside, VLDB '06 , September 12-15, 2006, Seoul, Korea

[14] Chang-Tien Lu, Dechang Chen, Yufeng Kou," Algorithms for Spatial Outlier Detection", Third IEEE International Conference on Data Mining (ICDM'03), 2003

[15] M. I. Petrovskiy, "Outlier Detection Algorithms in Data Mining Systems", Programming and Computer Software, Vol. 29, No. 4, 2003, pp.228–237