Gender Equality Analysis In The Census Report 2011 For Tamil Nadu State By Using Outlier Detection

R.Vijay Anand[1], R.Manicka chezian[2]

ABSTRACT

Spatial outliers are those observations which are inconsistent with their surrounding neighbors. Identification of spatial outliers can lead to the discovery of unexpected, interesting, and useful spatial patterns for further analysis. One draw-back of existing methods is that normal objects tend to be falsely detected as spatial outliers when their neighborhood contains true spatial outliers. Chen D et al propose a suite of spatial outlier detection algorithms to overcome this disadvantage[8]. In this paper, the census report 2011 for Tamil Nadu state is taken to analyze gender equality (i.e.) number of females per 1000 males using spatial outlier detection.

Keywords – Spatial Outlier, Data Mining, Tamil Nadu Census Report, Outlier Detection

I. INTRODUCTION

Spatial data mining [1] [4] is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. There has been a lot of research focusing on various spatial/spatio-temporal data mining tasks, including spatial clustering , spatial classification , spatial outlier detection ,spatial hotspot detection, spatial/spatio-temporal association rules (such as collocation and co-occurrence) discovery etc. Among them, spatial outlier detection, which aims at finding out locations where data observed are significantly different from other locations in the proximity in a large data set. An outlier [2] is a observation which deviates so much from other observation as to arouse suspicions that it was generated by different mechanism. Outliers in the spatial data can be classified into three categories: set-based outliers, multi-dimensional space-based outliers and graph-based outliers. A set-based outlier is a data object whose attributes are inconsistent with the attribute values of other objects in a given data set regardless of the spatial relationships. Both multidimensional and graph-based outliers are spatial outliers, that is, data objects that are significantly different in attribute values from the collection of data objects among spatial neighborhoods. The goal of outlier detection is to uncover the different mechanism. The identification of outliers can lead to the discovery of useful knowledge and has a number of practical applications in areas such as credit card fraud detection, severe weather prediction etc. Spatial outliers are those observations which are inconsistent with their surrounding neighbors. In identification of spatial outliers, attribute space is generally divided into two parts, non-spatial attributes and spatial attributes. Spatial attributes record the information related to locations, boundaries, directions, sizes, and volumes, which determine the spatial relationships between neighbors. Based on the neighborhood relationship, non-spatial attributes can be processed to identify abnormal observations. They are different from traditional outliers in the following aspects. Traditional outliers focus on global comparison with the

[1]Research Scholar, Department of Computer Science, NGM College, Pollachi [1]vijayanand.r86@gmail.com

[2]Associate professor, Department of computer science, NGM College, Pollachi. [2]chezian_r@yahoo.co.in

whole data set while spatial outliers pay more attention to local differences among spatial neighborhood. Traditional outlier detection mainly deals with numbers, characters, and categories, whereas spatial outlier detection processes more complex spatial data such as points, lines, polygons, and 3D objects. Spatial outlier detection plays an important role in many applications including weather forecast, military image analysis and traffic management.

Recent work by Shekhar et al. introduced a method for detecting spatial outliers in graph data set [6]. The method is based on the distribution property of the difference between an attribute value and the average attribute value of its neighbors. Several spatial outlier detection methods are also available in the literature of spatial statistics. These methods can be generally grouped into two categories, namely graphic approaches and quantitative tests.

Major drawback of existing system is that some true outliers are ignored and false outliers are identified. To avoid this here, we are using the Median algorithm [8]. So, this algorithm can detect only the true spatial outliers.

II. Algorithms

(Median Algorithm)

Algorithm is an non-iterative algorithm which uses median as neighborhood, thus reducing the negative impact caused by presence of neighboring points with very high/low attribute values.

1. For each spatial point xi , compute the k nearest neighbor set $NNk(xi)$, the neighborhood function $g(xi)$ = median of the data set $\{f(x) : x \in N N_k(xi)\}$, and the comparison function

   $$hi = h(x_i) = f(x_i) - g(x_i).$$

2. Let μ and σ denote the sample mean and sample Standard deviation of the data set $\{h1, h2, \ldots, hn\}$. Standardize the data set and compute the absolute values

   $$yi = \left| \frac{hi - \mu}{\sigma} \right|$$

   for $i = 1, 2, \ldots, n$.

3. For a given positive integer m, let $i1, i2, \ldots, im$ be the m indices such that their y values in $\{y1, y2, \ldots, yn\}$ represent the m largest. Then the m

   S-outliers are $xi1, xi2, \ldots, xim$.

Table 2 shows the results using the algorithm with parameters $k = m = 2$, compared with the existing approaches. As seen, the three proposed algorithm is accurately detect $S1$, $S2$ as spatial outliers. In this table, the rank of the outliers is defined in an obvious way. For example, in iterative r and z algorithms, the rank is the order of iterations, while in both z and Median algorithms; the rank is determined by the y value.

Rank in 2011	District	Sex- Ratio (Number of Females per 1000 males) 2011
1	The Nilgiris	1041
2	Thanjavur	1031
3	Nagapattinam	1025
4	Thoothukkudi	1024
5	Tirunelveli	1024
6	Thiruvarur	1020
7	Ariyalur	1016
8	Pudukkottai	1015
9	Karur	1015
10	Tiruchirappalli	1013
11	Kanniyakumari	1010
12	Virudhunagar	1009
13	Perambalur	1006
14	Vellore	1004
15	Coimbatore	1001
16	Sivaganga	1000
17	Dindigul	998
18	Tiruvannamalai	993
19	Erode	992
20	Madurai	990
21	Theni	990
22	Tiruppur	988
23	Namakkal	986
24	Chennai	986
25	Viluppuram	985
26	Kancheepuram	985
27	Cuddalore	984
28	Thiruvallur	983
29	Ramanathapuram	977
30	Krishnagiri	956
31	Salem	954
32	Dharmapuri	946

Table1.Ranking of Districts by Sex-Ratio for the Year 2011

The graphical representation of table1 for ranking of districts of Tamil Nadu state, India by sex-ratio for the year 2011 is shown in figure1.The true outliers detected by using the Median algorithm is illustrated in figure 2.

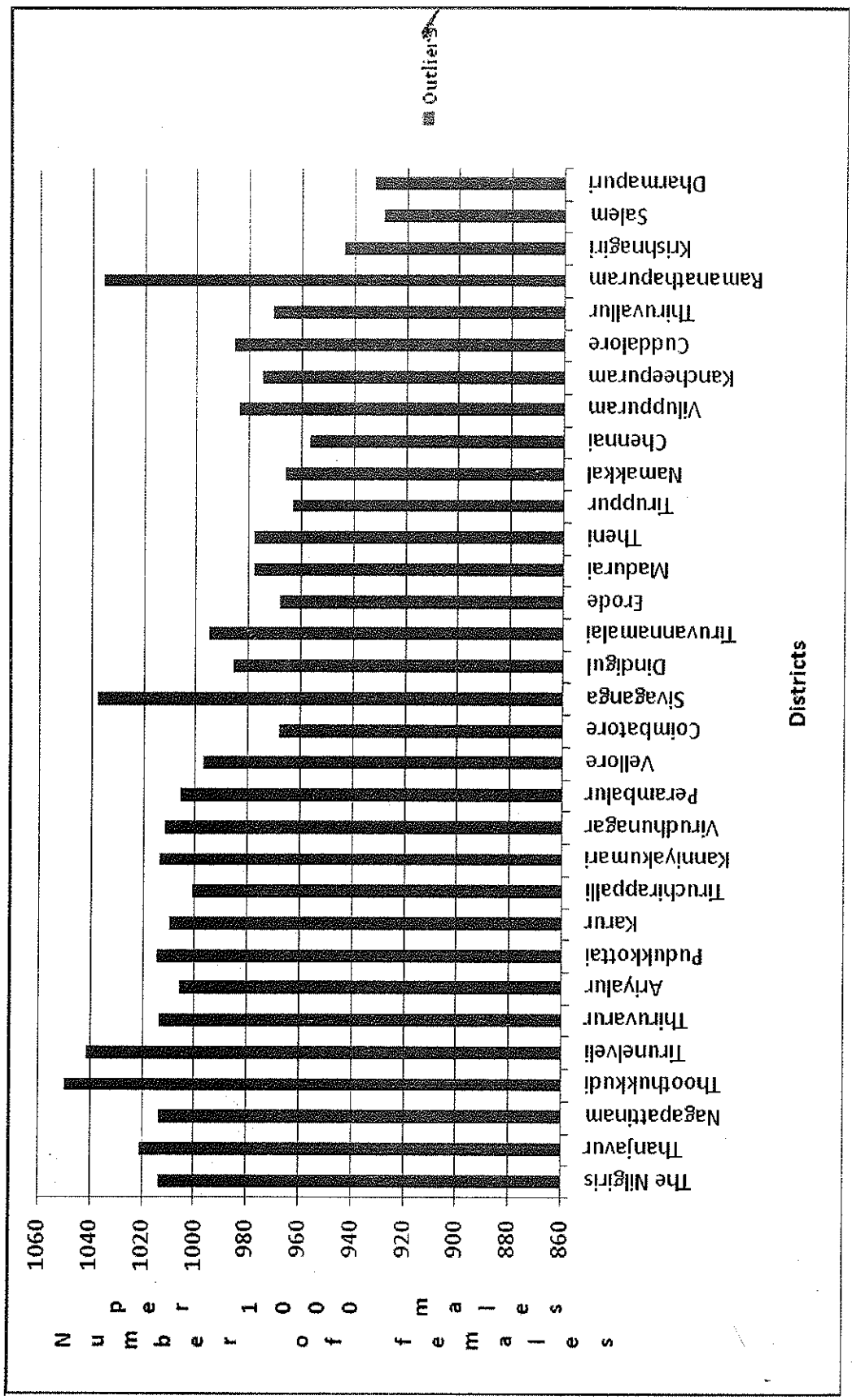

Figure1. No of females per 1000 males as per Tamil Nadu Districts

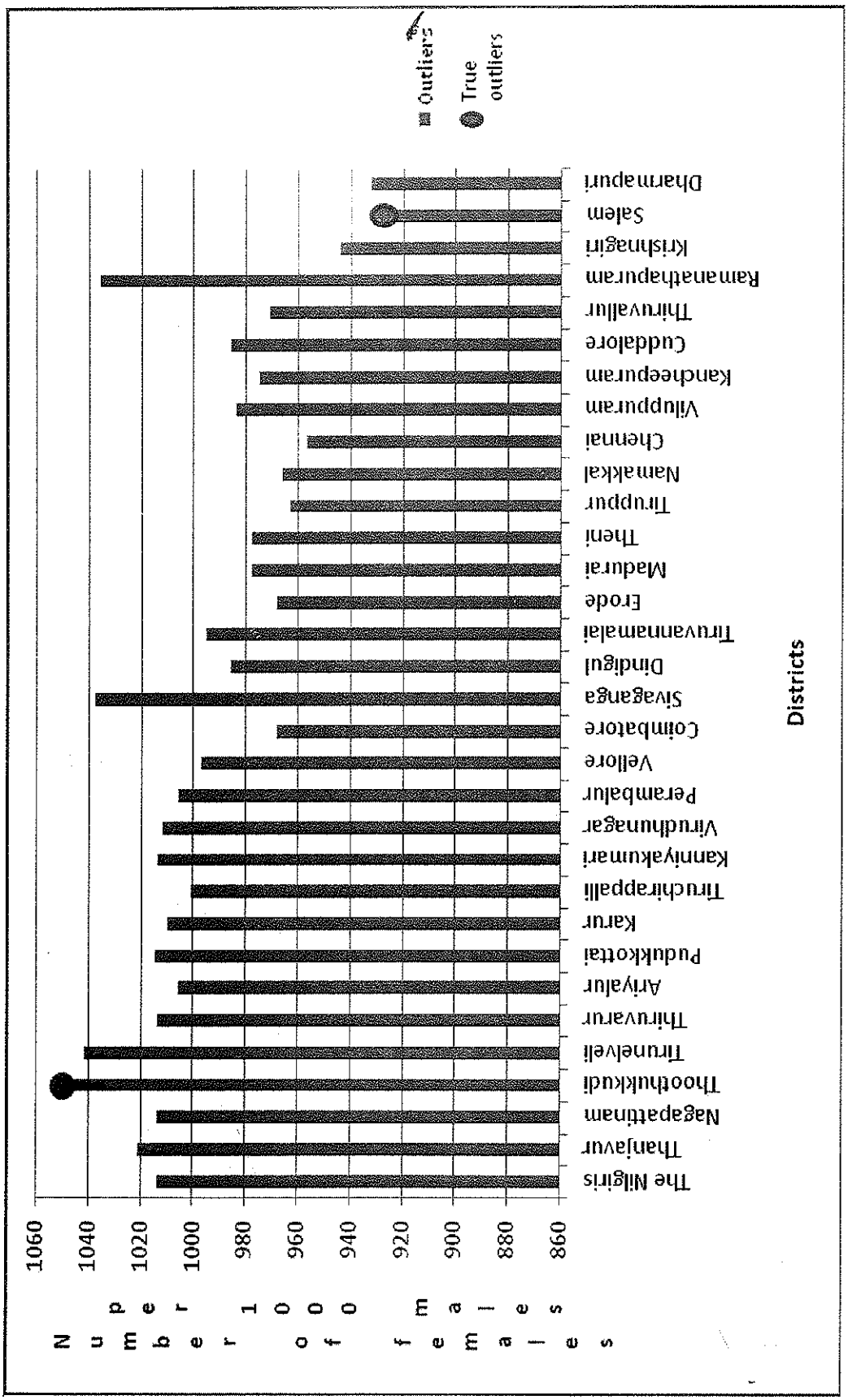

Figure2. Ranking of Districts by Sex-Ratio for the year 2011 with Detected Outliers

III. Experiments

We empirically compared the detection performance of Chen D et al. proposed methods with the z algorithm through mining a real-life census data set. The experiment results indicate that the algorithm can successfully identify spatial outliers ignored by the z algorithm and can avoid detecting false spatial outliers.

In this experiment, we tested various attributes from census data compiled by Indian Census Bureau [9]. The attributes tested include the ranking of districts by sex-ratio, 2011. We first ran the two algorithms (z, median algorithm) to detect which districts have abnormal sex ratio. There are 32 districts in the Tamil Nadu, India. We show the top 10 districts which are most likely to be the spatial outlier's in Table2.

Experimental results from other attributes also show that the median method is more accurate than the non-iterative algorithm in terms of falsely detected spatial outliers [6]. For running the algorithms and generating more results, Chen D et al. refer interested readers to, where they developed one software package which implements all the existing and proposed algorithms.

Rank in 2011	District	Sex- Ratio (Number of Females per 1000 males) 2011
1	The Nilgiris	1041
2	Thanjavur	1031
3	Nagapattinam	1025
4	Thoothukkudi	1024
5	Tirunelveli	1024
6	Tiruppur	988
7	Namakkal	986
8	Chennai	986
9	Viluppuram	985
10	Kancheepuram	985

Table2.The Top Ten Spatial Outliers detected by Median

IV. Conclusions

In this paper we have analyzed the census report 2011 for every district in Tamil Nadu state, the number of female births per 1000 male births by using spatial outlier detection. Chen D et al., propose spatial outlier detection algorithm to analyze spatial data: algorithm based on median. By using this algorithm, we have found the true outliers by analyzing number of females per 1000 males in the district of Tamil Nadu, India and generated the graphical representations. The experimental results confirm the effectiveness of the approach in reducing the risk of falsely claiming regular spatial points as outliers, which exists in commonly used detection methodologies. Furthermore, it carries the important bonus of ordering the spatial outliers with respect to their degree of outlierness.

REFERENCES

[1] M.Hemalatha., Naga saranya. "A Recent Survey On Knowledge Discovery in Spatial Data Mining", IJCSI International Journal of Computer Science Issue, Vol.8, Issue8, No.1, May 2011.

[2] Ben-Gal. Outlier Detection. In: Maiman O. and Rockach L.(Eds) "Data Mining and Knowledge discovery handbook: A Complete Guide for Practitioners and Researchers", Kluwer Academic Publishers, 2005.

[3] S.Shekar., Chang-Tien .Lu. and P.Zhang."A Unified Approach To Spatial Outlier Detection". Geoinformatica, An International Journal on advances of computer science for GIS, Dec 2001.

[4] Li.D.R., Wang S.L., Li D.Y."Theories and Applications of Spatial Data Mining". (Beijing: science press), 2005.

[5] Ramaswamy.s., Rastogi.R. And Kyuseok Shim."EfficientAlgorithms for Mining Outliers from Large Data Sets". ACM Special Interest Group of Management of Data, 2000.

[6] S.Shekar., Chang-Tien Lu. and P.Zhang. "Detecting Graph-Based Spatial Outlier Detection". Intelligent Data Analysis:An International Journal, 6(5):451-468,2002.

[7] D.Hawkins. "Identification of Outliers." Chapman and Hall, 1980.

[8] Chang-Tien Lu ., Dechang Chen ., Yufeng Kou. "Algorithm For Spatial Outlier Detection". Proceedings of the third IEEE International Conference on Data Mining (ICDM'03),0-7695-1978-4/03. 2003.

[9] Indian Census Bureau,India.http:// www.census.tn.nic.in