Knowl Inf Syst (2006) 9(4): 412–429
DOI 10.1007/s10115-005-0200-2

**Knowledge and Information Systems**

**REGULAR PAPER**

**Sanjay Chawla · Pei Sun**

# SLOM: a new measure for local spatial outliers

Received: 1 June 2004 / Revised: 11 January 2005 / Accepted: 30 January 2005 /
Published online: 27 September 2005


**Abstract** We propose a measure, spatial local outlier measure (SLOM), which captures the local behaviour of datum in their spatial neighbourhood. With the help of SLOM, we are able to discern local spatial outliers that are usually missed by global techniques, like "three standard deviations away from the mean". Furthermore, the measure takes into account the local stability around a data point and suppresses the reporting of outliers in highly unstable areas, where data are too heterogeneous and the notion of outliers is not meaningful. We prove several properties of SLOM and report experiments on synthetic and real data sets that show that our approach is novel and scalable to large datasets.

**Keywords** Spatial local outlier · Spatial neighbourhood · Oscillating parameter · R-trees index · Complexity

## 1 Introduction and related work

Of all the data-mining techniques, outlier detection seems closest to the definition of discovering nuggets of information in large databases. When an outlier is detected and determined to be genuine, it can provide insights that can radically change our understanding of the underlying process. We give a historical example of how the discovery of outliers led to a better understanding and prediction of global weather patterns known as El Niño and La Niña.

In the early 1900s, Sir Gilbert Walker, a British meteorologist, discovered that extreme variations in surface pressure over the equator close to Australia are correlated with monsoon rainfall and drought in India and other parts of the world. This variation is captured in a measure, which is now called the Southern Oscillation

S. Chawla · P. Sun (✉)
School of Information Technologies, University of Sydney, New South Wales, Australia
E-mail: Psun2712@it.usyd.edu.au

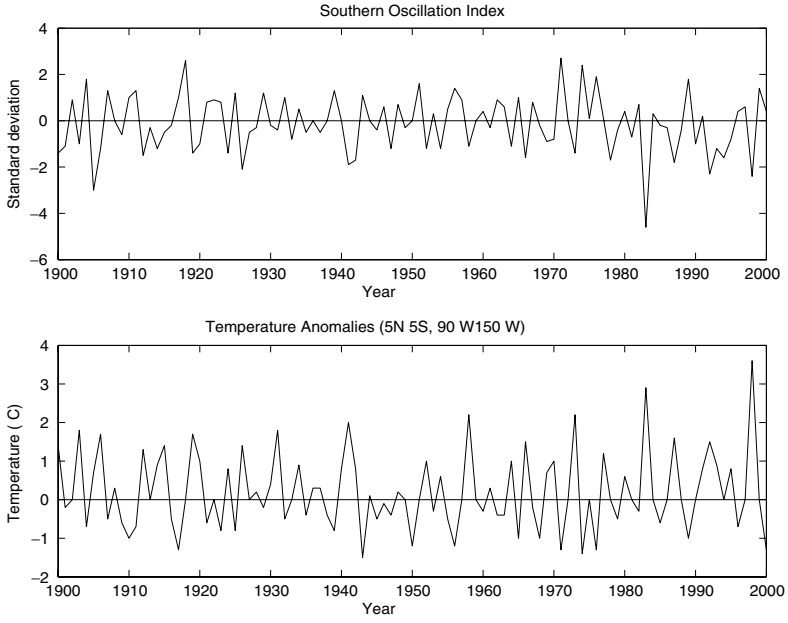



**Fig. 1** The relationship between the Southern Oscillation Index(SOI) and sea surface temperature. High temperature anomalies correspond to El Niño and low to La Niña. The relationship was discovered by Sir Gilbert Walker and clearly shows how outlier detection can provide penetrating insights about the underlying phenomenon, global weather patterns in this case

Index (SOI). The SOI is defined as the normalized surface air-pressure difference between the islands of Tahiti and Darwin, Australia. As shown in the upper graph in Fig. 1(reprinted from McFadden (2002)), when the SOI index attains outlier values, i.e. when it is two or more standard deviations away from the mean, the sea surface temperature over the Pacific Ocean also rises and falls sharply (lower graph). Thus, a SOI of two standard deviations below the mean corresponds to a rise in surface temperature and is known as El Niño. The opposite phenomenon, i.e. when SOI is two or more standard deviations above the mean, which corresponds to a fall in surface temperature, is known as La Niña. Notice how, in 1998, the sea surface temperature reached more than 3°C above normal and was one of most dramatic El Niño years in recorded history. Also notice that the relationship between SOI and El Niño is sharper than that between SOI and La Niña.

Thus, an automated or partially automated system of outlier detection can serve as a trigger for unlocking secrets about the underlying process that has generated the data.

The classic definition of an outlier is due to Hawkins (1980), who provides the definition "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". Several different approaches have been taken in order to operationalize this definition. For example, it is standard to use variations of the Chebyshev's inequality,

$$P(\mid X - \mu \mid \geq k\sigma) \leq \frac{1}{k^2},$$

where $\mu$ and $\sigma$ are the mean and variance of a random variable, $X$, which models the underlying mechanism. When additional information is available, like the distributional assumption of $X$, this inequality can be sharpened. For example, when $X$ follows a normal distribution, it can be shown that 99.7% of the data lies between three standard deviations, as opposed to 88.8% given by the general Chebyshev's inequality.

Knorr and Ng (1998) were the first to propose the definition of distance-based outlier, which was free of any distributional assumptions and was readily generalizable to multidimensional datasets. They gave the following definition of *DB(p, D)* outlier: "An object $o$ in a dataset $T$ is a $DB(p, D)$-outlier if at least fraction $p$ of the objects in $T$ lies at a greater distance than $D$ from $o$".

The authors proved that this definition generalized the folk definition of outliers "three standard deviations away from the mean". For example, if the dataset $T$ is generated from a normal distribution, with mean $\mu$ and standard deviation $\delta$, and $t \in T$ is such that $\frac{t-\mu}{\delta} > 3$, then $t$ is a $DB(p, D)$ outlier with $p = 0.9988$ and $D = 0.13\delta$. Similar extensions were shown for other well-known distributions, including the Poisson.

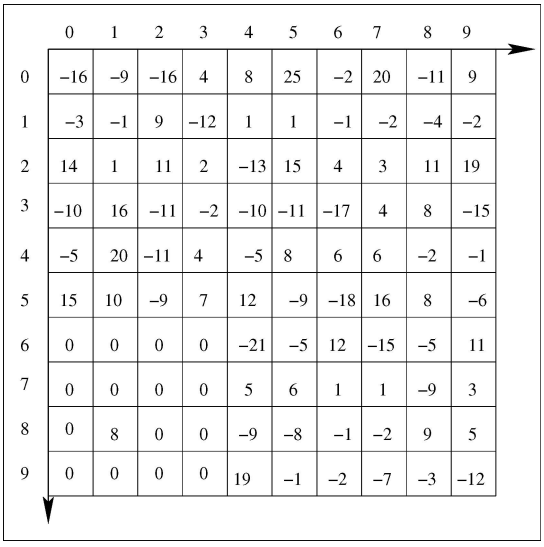
Following the definition of distance-based outlier introduced by Knorr and Ng, several methods and algorithms (Aggarwal and Yu 2001; Bay and Schwabacher 2003; Angiulli and Pizzuti 2002; Knorr and Ng 1998; Ramaswamy et al. 2000) have been proposed to detect distance-based outliers. However, the outliers detected by these methods and algorithms are global outliers. Breunig et al. (2000) argued that, in some situations, local outliers are more important than global outliers. They proposed the concept of a local outlier factor (LOF), which defines how isolated an object is with respect to its surrounding neighbourhood rather than the whole dataset. Papadimitriou et al. (2003) have also proposed a method, local correlation integral (LOCI), to detect local outliers. Their method is similar to LOF except that they use a different definition to define the local neighbourhood.

For spatial data, both statistical and data-mining approaches have to be modified because of the qualitative difference between spatial and nonspatial dimensions. The attributes that comprise the nonspatial dimensions intrinsically characterise the data while the spatial dimensions provide a locational index to the object and are not intrinsic to the object. However, the physical neighbourhood plays a very important role in analysis of spatial data. For example, in Fig. 2, the data value 8 indexed at location (8, 1) is an outlier; however, the same value 8 indexed at (3, 8) is not an outlier.

Shekhar et al. (2001) proposed the following definition of spatial outlier: "A spatial outlier is a spatially referenced object whose nonspatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighbourhood".

A spatial neighbourhood may be defined based on spatial attributes, e.g., location, using spatial relationships such as distance or adjacency. Comparisons between spatially referenced objects are based on nonspatial attributes.

There are two types of spatial outliers: multidimensional space-based outliers and graph-based outliers. The only difference between them is that they use different spatial neighbourhood definitions. Multidimensional space-based outliers use Euclidean distances to define spatial neighbourhoods, while graph-based outliers use graph connectivity.



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | −16 | −9 | −16 | 4 | 8 | 25 | −2 | 20 | −11 | 9 |
| 1 | −3 | −1 | 9 | −12 | 1 | 1 | −1 | −2 | −4 | −2 |
| 2 | 14 | 1 | 11 | 2 | −13 | 15 | 4 | 3 | 11 | 19 |
| 3 | −10 | 16 | −11 | −2 | −10 | −11 | −17 | 4 | 8 | −15 |
| 4 | −5 | 20 | −11 | 4 | −5 | 8 | 6 | 6 | −2 | −1 |
| 5 | 15 | 10 | −9 | 7 | 12 | −9 | −18 | 16 | 8 | −6 |
| 6 | 0 | 0 | 0 | 0 | −21 | −5 | 12 | −15 | −5 | 11 |
| 7 | 0 | 0 | 0 | 0 | 5 | 6 | 1 | 1 | −9 | 3 |
| 8 | 0 | 8 | 0 | 0 | −9 | −8 | −1 | −2 | 9 | 5 |
| 9 | 0 | 0 | 0 | 0 | 19 | −1 | −2 | −7 | −3 | −12 |

**Fig. 2** Original data matrix

Thus, given a function $f$ defined on a spatial grid $S$, a natural approach is to transform $f$ into $g$ such that $g(o) = f(o) - \frac{1}{|N(o)|} \sum_{p \in N(o)} f(p)$, where $N(o)$ is the spatial neighbourhood of $o$. Now, a Chebyshev inequality-like approach can be undertaken in order to identify those points, $o$, that are candidate outliers. Indeed, this is the state of the art (Lu et al. 2003a, 2003b; Shekhar et al. 2001, 2003).

However, the approach of using a statistical test is useful for discovering global outliers but may not be able to discover local outliers, which are likely to be of more interest. For example, again consider the data value 8 indexed at location (8, 1) in Fig. 2. Clearly, this point is a local outlier as it forms a local maxima in its neighbourhood; however, the value 8 is not a global outlier in the sense that, even after transformation, it still is within three standard deviations from the mean.

Thus, clearly an approach is needed that can efficiently capture spatial local outliers. In fact, our method will go further and associate a SLOM score with each data point. The SLOM defines the degree of outlierness of each point very much along the lines proposed by Breunig et al. (2000). However, besides the qualitative difference between spatial and nonspatial attributes, spatial data exhibits spatial autocorrelation (nonindependence) and heteroscedasticity (nonconstant variance), both of which must be factored into SLOM.

## 1.1 Problem definition

**Given:** A large spatial database with multidimensional, nonspatial attributes.
**Design:** A measure that assigns a degree of outlierness to each element in the database.
**Constraints:**

**Spatial autocorrelation**: The value of each element in the database is affected by its spatial neighbours.

**Spatial Heteroscedasticity**: The variance of the data is not uniform and is a function of the spatial location.

Together, these two constraints imply that the IID (identical and independent distribution) assumption cannot be assumed to hold in the context of spatial data.

### 1.2 Key insights and contributions

1. The first insight that guides our approach can be described with the help of an example. Consider the cell with value 8 indexed at location (8, 1) in Fig. 2. Clearly, in the local neighbourhood, 8 is an outlier. An obvious way to capture the relationship between a point and its neighbours is to define a measure $d(o)$ for each point $o$ as

$$d(o) = \frac{1}{\mid N(o) \mid} \sum_{p \in N(o)} \text{dist}(o, p),$$

where dist$(o, p)$ is a definition of (Euclidean) distance between the nonspatial components of $o$ and $p$ and $N(o)$ are the neighbouring points of $o$.

In Fig. 2, the value of $d(o)$ is 8 for object $o$ located at (8, 1). However, for a point $p$ in the neighbourhood of $o$, which is not an outlier, the influence of $o$ can overwhelm $p$'s relationship with its other neighbours. In order to factor out the effect of $o$ on $p$, a modified measure, $\tilde{d}(o)$, is defined as follows:

First, define $maxd(o) = max\{\text{dist}(o, p) \mid p \in N(o)\}$ as the maximum nonspatial distance between $o$ and its neighbours. Then define

$$\tilde{d}(o) = \frac{\sum_{p \in N(o)} \text{dist}(o, p) - \text{maxd(o)}}{\mid N(o) \mid -1}.$$

*Now notice that, for the point* 8 *(location (8,1)) in Fig. 2,* $\tilde{d}(o) = d(o)$*, but for points in the neighbourhood of this point,* $0 = \tilde{d}(p) < d(p) = 1$.

Thus, the advantage of using $\tilde{d}(o)$ instead of $d(o)$ is that, if $o$ is an outlier, then $\tilde{d}$ suppresses the effect of $o$ in its neighbourhood. The following two theorems hold under some conditions. They formalise the relationships between $d$ and $\tilde{d}$. A proof is given in the Appendix.

**Theorem 1** $\tilde{d}(o) - \tilde{d}(p) > d(o) - d(p)$

**Theorem 2** $\frac{\tilde{d}(o)}{\tilde{d}(p)} > \frac{d(o)}{d(p)}$.

The definition of $\tilde{d}$ is similar to that of *trimmed mean*, where a certain percentage of the largest and smallest values around the mean are removed (Wilcox 2003). The trimmed mean is less sensitive to outliers like the median but retains some of the averaging behaviour of the mean.

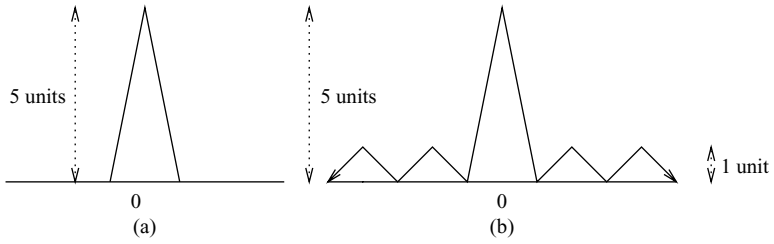



**Fig. 3** Both **a** and **b** have the same $\tilde{d}$ value; however, the $\beta$ values in **a** are higher than **b** because of the instability around **b**

2. The second insight that underpins our approach is that outliers that are in unstable areas should have lower precedence than outliers in stable areas. Stability around a point $o$ can be captured using the variance; however, we have used a statistic that can be deterministically bounded. In particular, we have defined a statistic $\beta$ that captures the net oscillation with respect to the average value around $o$ (details in Sect. 2). For example, Fig. 3 shows the plot of $\tilde{d}$ around the point $o$. For both the figures, $\tilde{d}(o)$ is the same but $\beta(o)$ in Fig. 3**a** is higher than Fig. 3**b**.
3. Another novel contribution of our work is related to system integration. All the spatial data remain database in situ. We manage this by exploiting the growing list of spatial features that are now standard features in commercial database systems, such as Oracle9i. In particular, we use R-trees to access the database and spatial sql to retrieve data based on spatial relationships.

The rest of the paper is as follows. In Sect. 2, we introduce a series of definitions that will culminate in the definition of the spatial local outlier measure (SLOM). Along the way, we will explain how each component of SLOM addresses spatial autocorrelation and heteroscedasticity. In Sect. 3, we analyse the complexity of our method and describe two database strategies to efficiently interact with the database in order to reduce the I/O overhead. In Sect. 4, we report the results of our experiments on synthetic and real data sets. In Sect. 5, we conclude with a summary and directions for future work.

## 2 Definitions

We now formally define SLOM and prove several properties. Recall that our objective is to design a measure that can capture both spatial autocorrelation and heteroscedasticity (nonconstant variance). We have already defined $\tilde{d}$ in Sect. 1, which factors out the effect of spatial autocorrelation, and now define $\beta$, which penalises for oscillating behaviour around a potential outlier.

For an object $o$, we can define its SLOM value as $\sum_{p \in N(o)} \frac{\tilde{d}(o)}{\tilde{d}(p)} / \mid N(o) \mid$, just like the method used in Breunig et al. (2000). However, this definition has two drawbacks.

1. First, one extreme (small) value of $\tilde{d}(p)$ will result in a very large SLOM value of an object $o$.

2. Second, it is possible that the value of $\tilde{d}(p)$ is zero, which will make the value of SLOM $= \infty$.
   We begin by quantifying the average of a $\tilde{d}$ in its neighbourhood.
   1. Let $N_+(o)$ denote the set of all the objects in $o$'s neighbourhood and $o$ itself and $\text{avg}(N_+(o)) = \sum_{p \in N_+(o)} \tilde{d}(p)/\mid N_+(o) \mid$.
   2. Oscillating parameter $\beta(o)$: For an object $o$, if it has large value for $\tilde{d}(o)$ and small $\tilde{d}$ value in $o$'s neighbourhood, then this means it is a good candidate for an outlier. On the other hand, even though it may have the largest value in its neighbourhood, if all neighbours also have large values, this means that $o$ inhabits an unstable (oscillating) area, so it is a poor candidate for an outlier. We define a parameter $\beta(o)$, which can capture the oscillation of an area, which intuitively is the net number of times the values around $o$ are bigger or smaller than $\text{avg}(N_+(o))$. We calculate $\beta(o)$ using the following pseudo-code:

      1. $\beta(o) \leftarrow 0$
      2. For each $p \in N_+(o)$
         if $\tilde{d}(p) > \text{avg}(N_+(o))$
         $\beta(o) + +$
         else if $\tilde{d}(p) < \text{avg}(N_+(o))$
         $\beta(o) - -$
      3. End for
      4. $\beta(o) = \mid \beta(o) \mid$
      5. $\beta(o) = \frac{\max(\beta(o),1)}{(\mid N_+(o) \mid -2)}$
      6. $\beta(o) = \frac{\beta(o)}{1+\text{avg}\{\tilde{d}(p) \mid p \in N(o)\}}$

      While steps 1–4 are self-explanatory, we explain steps 5 and 6. There are two reasons why we divide $\beta(o)$ by $\mid N_+(o) \mid -2$ in step 5. First, we need to correct for boundary terms where the number of neighbours is fewer than that in the interior. The second motivation is that, for a local region like that in Fig. 2, where the data value 8 at location $o = (8, 1)$ is surrounded by constant values, $\beta(o) = 1$, the highest value $\beta$ can assume.
      However, if we have stopped at step 5, then $\beta$ cannot distinguish between the two cases shown in Fig. 3. In order to do that, we divide $\beta(o)$ by $1 + \text{avg}\{\tilde{d}(p) \mid p \in N(o)\}$. This allows us to penalise the situation where large values of $\tilde{d}$ exist around the point $o$. However, in order to bound this term, we have to normalise the original data so that the maximum value that the denominator can assume is $1 + \sqrt{\text{d}}$, where $d$ is the dimensionality of the nonspatial attributes. Thus, in Fig. 3a and b, the $\beta$ values are 1 and 0.5, respectively.
3. We are ready to define SLOM. For a point $o$,

$$\text{SLOM(o)} = \tilde{d}(o) * \beta(o).$$

A high value of SLOM indicates that the point is a good candidate for an outlier. The $\tilde{d}$ term is analogous to the expectation of the first derivative of a smooth

random variable, while the $\beta$ term is analogous to the standard deviation of the first derivative of a smooth random variable.

**Lemma 1** *For all* $o \in S$, $1/(\mid N_+(o) \mid -2)(1+\sqrt{d}) < \beta(o) \leq 1$, *where* $d$ *is the dimensionality of the nonspatial attributes.*

*Proof* After step 4 of computing $\beta(o)$, the maximum value of $\beta(o)$ is $\mid N_+(o) \mid -2$. This happens when $\tilde{d}(p)$ is the only value that is greater than (or smaller than) $\text{avg}(N_+(o))$. The minimum value of $\beta(o)$ is 0. After step 5, the maximum value of $\beta(o)$ becomes 1, and the minimum value becomes $1/\mid N_+(o) \mid -2$. In step 6, the maximum value of $\text{avg}\{\tilde{d}(p) \mid p \in N(o)\}$ is $\sqrt{d}$ and the minimum value is 0. So, after step 6, the maximum value of $\beta(o)$ becomes 1 and the minimum value becomes $1/(\mid N_+(o) \mid -2)(1+\sqrt{d})$. □

**Lemma 2** *For all* $o \in S$, $0 \leq \text{SLOM(o)} \leq \sqrt{d}$.

*Proof* The value of $\tilde{d}(o)$ is between 0 and $\sqrt{d}$. From Lemma 1, we know that the value of $\beta(o)$ is between $\frac{1}{(|N_+(o)|-2)(1+\sqrt{d})}$ and 1. SLOM(o) is the product of $\tilde{d}(o)$ and $\beta(o)$, so its value must be between 0 and $\sqrt{d}$. □

## 3 Complexity analysis

For distance-based outlier detection (Knorr and Ng 1998), the key step is a method to search for nearest neighbours. This search must be performed on the complete dataset and is the computational bottleneck, especially in high-dimensional space.

However, for spatial outlier detection, the neighbourhood is defined by it spatial information, which is usually bounded by three dimensions. Here, we can use a spatial R-tree index in order to perform this step efficiently.

Given that we have $N$ objects and each object has a maximum of $k$ spatial neighbours ($k \leq 8$ for a 2D grid), the calculation of SLOM for the full data set involves the following steps:

1. The first step is to normalise the nonspatial attributes to between [0, 1]. Here we can take advantage of the summary statistics (min, max, avg) that are stored in the database catalogue. Thus, this step can be done in one database pass and the computational cost is $O(Nd)$, where $d$ is the number of nonspatial dimensions.
2. To compute $\tilde{d}(o)$, we need to find the spatial neighbours of each object and calculate the distance between them. The cost of a single *k-NN* query using an R-tree is $O(k \log N)$, and the cost of computing the nonspatial distance is $O(kd)$. Thus, the cost of this step is $O(Nk \log N + kdN)$.
3. After the computation of $\tilde{d}(o)$, we need to compute $\beta(o)$. This involves another round of nearest neighbour queries followed by a summation to compute the neighbourhood average of $\tilde{d}(o)$. The cost of this step is thus $O(Nk \log N + dN)$.
4. To compute the SLOM, we multiply the $\tilde{d}$ and $\beta$ for each object. The cost is $O(N)$.
5. Finally, we sort the objects by SLOM and report the top-$n$ outliers, for which the cost is $O(N \log N)$.
6. Thus, the final cost of the whole operation is $O(Nk \log N + kdN)$.

While the spatial dimensionality is bounded by three, the spatial part of the data set can be quite large and complicated, especially if the spatial objects are complex polygons (like the boundaries of countries). Even though the R-tree index can speed up the processing of the nearest neighbour search, finding the $k$ nearest neighbours is still a very time-consuming task. We have two options in order to avoid accessing the original spatial data twice (steps 2 and 3 above).

The first option is that we store the neighbourhood information in main memory when we compute $\tilde{d}$. Then, when computing $\beta(O)$, we can access this information from memory rather than the database. The prerequisite for this is that the main memory should be large enough to hold all relevant information.

The second option is that we use an R-tree index to generate the neighbourhood information and store it into a table beforehand. When computing $\tilde{d}$ and $\beta$, we visit this table instead of the original table that stores the spatial information. Because spatial data enjoys slow updates, this is a very attractive option and can result in huge savings in the running time as we can amortise the cost of creating the neighbourhood table over subsequent *k-NN* queries.

## 4 Experiments, results and analysis

We have carried out detailed experiments on synthetic and real datasets in order to

1. Test whether SLOM can pick up *local outliers* and suppress the reporting of *global outliers* in unstable areas. Intuitively, a point is a global outlier, if its determination, that it is an outlier, depends on a comparison with all other points in the data set. A point is classified as a local outlier if its determination is based on a comparison with points in its neighbourhood.
2. Compare the SLOM approach with the family of methods to discover spatial outliers proposed in Lu et al. (2003a) and Shekhar et al. (2001, 2003).
3. Test how the running time changes as we vary the number of nearest neighbours used in the experiments.

One of the strengths of our approach is that the data always remains database in situ, i.e. we never have to extract the data from the database into a flat file in order to carry out the data-mining exercise. In particular, all the spatial *k-NN* queries are carried out inside the database.

We accomplish this using the set of spatial features that are increasingly becoming a standard component in commercial and open source DBMS, like Oracle and Postgres, respectively. In particular, these systems provide an R-tree structure to index spatial data and also support extensions of SQL to formulate queries that involve spatial relationships. We used Oracle9i to store all spatial and nonspatial data.

In our experiments, all the spatial objects are polygons. For an object $o$, all the spatial objects that directly touch its boundaries are defined to be its neighbours. We use the following SQL statement to generate the neighbourhood information:

```
select  a.id, b.id
from    spatial a, spatial b
where   sdo_relate(a.geom,b.geom,
        'mask=touch querytype=window')='true'
```

In the table `spatial`, $GEOM$ is a special column that stores the boundary information of each object and an R-tree index is created on it to speed up the processing of this query.

If the spatial objects are points and the neighbourhood is defined to be the $k$ nearest neighbours, then the following SQL statement can be use to generate the neighbourhood information:

**select** a.id, b.id
**from** spatial a, spatial b
**where** sdo_nn(a.geom,b.geom,'sdo_num_res=k')=1true'

### 4.1 Results on a synthetic dataset

We have created a synthetic data set consisting of one nonspatial attribute in order to explain and compare our method with the prototype method proposed in Lu et al. (2003b) and Shekhar et al. (2001, 2003). We will refer to this method as SLZ (Shekhar, Lu and Zhang). The core idea of SLZ is that, given a function $f$ defined on the spatial set $S$, the neighbourhood effect can be captured by the transformation $g(x) = f(x) - \sum_{y \in N(x)} f(y)$. This is followed by an application of a statistical test on $g$, inspired from Chebyshev's inequality, to determine the outliers of $f$.

Our synthetic data set consists of 100 spatial objects organised as a $10 \times 10$ matrix. We used a Gaussian generator to produce the values of nonspatial attribute, and they are listed in Fig. 2. The location of some values were deliberately changed so that all the zeros appeared at the lower-left corner and an 8 showed up at the location $(8, 1)$.

The top five outliers detected by SLZ (at a confidence interval of 95%) and SLOM are listed in Table 1. The $\tilde{d}$, SLOM and the SLZ matrices are shown in Figs. 4, 5 and 6, respectively. The objects located at positions $(0, 7)$, $(0, 5)$ and $(3, 9)$ are marked as outliers by both methods. This means that they are both global and local outliers. The objects located at position $(6, 4)$ and $(9, 4)$ are captured as one of the top five outliers by SLZ but not by SLOM. This means that they are global outliers but not local outliers, as they are located in unstable areas. This can be seen from their SLOM values, which are 0.06 and 0.10. The objects located at position $(8, 1)$ and $(2, 5)$ are captured as outliers by SLOM but not by SLZ. This means they are local but not global outliers. Again, their SLOM values are 0.17 and 0.25, respectively.

**Table 1** Outliers found by different method on the same dataset

| Position (SLZ method) | $g(x)$ value (SLZ method) | Position (Our method) | SLOM value |
|---|---|---|---|
| (0, 7) | 24.0 | (0, 5) | 0.4277 |
| (0, 5) | 23.6 | (2, 5) | 0.2479 |
| (6, 4) | −23.0 | (0, 7) | 0.2061 |
| (9, 4) | 22.6 | (8, 1) | 0.1739 |
| (3, 9) | −22.0 | (3, 9) | 0.1727 |

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.21 | 0.15 | 0.25 | 0.15 | 0.19 | 0.49 | 0.14 | 0.48 | 0.24 | 0.26 |
| 1 | 0.13 | 0.19 | 0.22 | 0.27 | 0.16 | 0.13 | 0.15 | 0.11 | 0.21 | 0.19 |
| 2 | 0.25 | 0.18 | 0.22 | 0.18 | 0.18 | 0.41 | 0.12 | 0.09 | 0.20 | 0.34 |
| 3 | 0.35 | 0.31 | 0.30 | 0.15 | 0.18 | 0.20 | 0.43 | 0.06 | 0.13 | 0.41 |
| 4 | 0.33 | 0.41 | 0.28 | 0.21 | 0.16 | 0.30 | 0.21 | 0.14 | 0.16 | 0.13 |
| 5 | 0.21 | 0.24 | 0.23 | 0.21 | 0.26 | 0.25 | 0.40 | 0.31 | 0.18 | 0.13 |
| 6 | 0.05 | 0.05 | 0.04 | 0.10 | 0.46 | 0.23 | 0.30 | 0.31 | 0.18 | 0.23 |
| 7 | 0.0 | 0.0 | 0.0 | 0.04 | 0.16 | 0.18 | 0.11 | 0.12 | 0.22 | 0.13 |
| 8 | 0.0 | 0.17 | 0.0 | 0.04 | 0.20 | 0.17 | 0.05 | 0.06 | 0.23 | 0.15 |
| 9 | 0.0 | 0.0 | 0.0 | 0.04 | 0.46 | 0.08 | 0.03 | 0.10 | 0.11 | 0.28 |

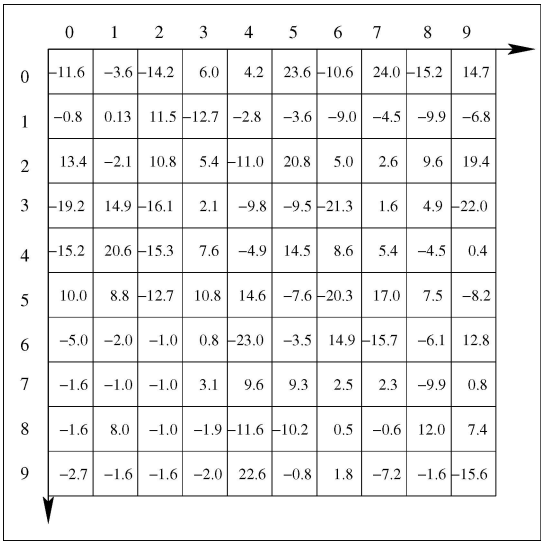
**Fig. 4** The matrix of the values of $\tilde{d}$.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.09 | 0.06 | 0.05 | 0.03 | 0.08 | 0.43 | 0.06 | 0.21 | 0.10 | 0.11 |
| 1 | 0.03 | 0.02 | 0.03 | 0.03 | 0.06 | 0.08 | 0.05 | 0.01 | 0.02 | 0.04 |
| 2 | 0.05 | 0.02 | 0.08 | 0.02 | 0.11 | 0.25 | 0.05 | 0.03 | 0.02 | 0.14 |
| 3 | 0.14 | 0.03 | 0.03 | 0.05 | 0.11 | 0.07 | 0.15 | 0.02 | 0.05 | 0.17 |
| 4 | 0.06 | 0.05 | 0.03 | 0.02 | 0.06 | 0.10 | 0.02 | 0.05 | 0.10 | 0.11 |
| 5 | 0.04 | 0.09 | 0.09 | 0.02 | 0.03 | 0.03 | 0.05 | 0.04 | 0.07 | 0.03 |
| 6 | 0.02 | 0.02 | 0.02 | 0.01 | 0.06 | 0.08 | 0.03 | 0.04 | 0.02 | 0.05 |
| 7 | 0.0 | 0.0 | 0.0 | 0.02 | 0.02 | 0.06 | 0.01 | 0.01 | 0.03 | 0.03 |
| 8 | 0.0 | 0.17 | 0.0 | 0.02 | 0.03 | 0.02 | 0.01 | 0.02 | 0.09 | 0.03 |
| 9 | 0.0 | 0.0 | 0.0 | 0.02 | 0.10 | 0.02 | 0.02 | 0.02 | 0.05 | 0.12 |

**Fig. 5** The SLOM matrix

## 4.2 Results on a real dataset—part one

The first real data set that we have used is from the U.S. Census Bureau and consists of spatial and nonspatial information about all the counties in the United States. The two-dimensional spatial information is used to define the spatial neighbourhood. For a specific county, all the counties that directly touch its boundaries are defined as its neighbours in this experiment. In order to make the results easily



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | −11.6 | −3.6 | −14.2 | 6.0 | 4.2 | 23.6 | −10.6 | 24.0 | −15.2 | 14.7 |
| 1 | −0.8 | 0.13 | 11.5 | −12.7 | −2.8 | −3.6 | −9.0 | −4.5 | −9.9 | −6.8 |
| 2 | 13.4 | −2.1 | 10.8 | 5.4 | −11.0 | 20.8 | 5.0 | 2.6 | 9.6 | 19.4 |
| 3 | −19.2 | 14.9 | −16.1 | 2.1 | −9.8 | −9.5 | −21.3 | 1.6 | 4.9 | −22.0 |
| 4 | −15.2 | 20.6 | −15.3 | 7.6 | −4.9 | 14.5 | 8.6 | 5.4 | −4.5 | 0.4 |
| 5 | 10.0 | 8.8 | −12.7 | 10.8 | 14.6 | −7.6 | −20.3 | 17.0 | 7.5 | −8.2 |
| 6 | −5.0 | −2.0 | −1.0 | 0.8 | −23.0 | −3.5 | 14.9 | −15.7 | −6.1 | 12.8 |
| 7 | −1.6 | −1.0 | −1.0 | 3.1 | 9.6 | 9.3 | 2.5 | 2.3 | −9.9 | 0.8 |
| 8 | −1.6 | 8.0 | −1.0 | −1.9 | −11.6 | −10.2 | 0.5 | −0.6 | 12.0 | 7.4 |
| 9 | −2.7 | −1.6 | −1.6 | −2.0 | 22.6 | −0.8 | 1.8 | −7.2 | −1.6 | −15.6 |

**Fig. 6** Result from the SLZ method

comprehensible, we have selected two nonspatial attributes: area and population density.

The two nonspatial attributes have different absolute magnitudes and may have a different effect on the SLOM values. We have standardized the values of each attributes to [0, 1] by using the formula $\frac{\text{value-min}}{\text{max-min}}$, where $min$ and $max$ are the minimum and maximum values of that attribute, respectively.

The information about the top five outlier counties and their neighbours is listed in Table 2. Not surprisingly, the top five outliers consist of counties that have large areas or large population densities. However, they are truly local outliers. For example, the area of the Yukon–Koyukuk county in Alaska is almost twice as big as the area of any of its neighbouring counties, and its population density (except for the North Slope county) is three times smaller. For urban areas, notice that Philadelphia is more outlierish compared with the Bronx even though it has a bigger area and smaller population density, again because its neighbourhood is relatively more stable.

### 4.3 Results on a real dataset—part two

The second real data set that we have used is from the U.S. Census Bureau as well. The two-dimensional spatial information is the same as the data set one. We have selected four nonspatial attributes: the proportion of people identified as African American, American Indian (including Eskimo and Aleut), Asian (including Pacific Islander), and of Hispanic origin. We standardized them by using the same method introduced in Part One.

The information about the top five local outlier counties with the highest SLOM values and their neighbouring counties is listed in Table 3. Menominee

**Table 2** Top five outliers and their neighbours. The two attributes used in the experiment are area and population density

| County name (Area, population density) | SLOM value | Neighbours | (Area, population density) |
|---|---|---|---|
| Yukon–Koyukuk, Alaska (157094.25, 0.05) | 0.2896 | Bethel Census Area, Alaska | (41080.34, 0.33) |
| | | Wade Hampton, Alaska | (17121.14, 0.34) |
| | | Southeast Fairbanks, Alaska | (25989.64, 0.23) |
| | | Fairbanks North S.B., Alaska | (7361.16, 10.56) |
| | | Matanuska-Susitna B., Alaska | (24689.41, 1.61) |
| | | Nome Census Area, Alaska | (23008.59, 0.36) |
| | | North Slope B., Alaska | (87845.38, 0.07) |
| | | Northwest Arctic B., Alaska | (35856.31, 0.17) |
| Philadelphia, Pennsylvania (135.11,11735.78) | 0.1884 | Burlington, New Jersey | (804.63, 490.99) |
| | | Delaware, Pennsylvania | (184.19, 2973.23) |
| | | Montgomery, Pennsylvania | (483.06, 1403.78) |
| | | Bucks, Pennsylvania | (607.54, 890.76) |
| | | Camden, New Jersey | (222.29, 2261.98) |
| | | Gloucester, New Jersey | (324.81, 708.36) |
| Suffolk, Massachusetts (58.51,11347.16) | 0.1831 | Essex, Massachusetts | (497.98, 1345.59) |
| | | Norfolk, Massachusetts | (399.54, 1542.01) |
| | | Middlesex, Massachusetts | (823.40, 1698.40) |
| Bronx, New York (42.02,28645.75) | 0.1633 | Bergen, New Jersey | (234.16, 3524.80) |
| | | Nassau, New York | (286.72, 4489.89) |
| | | New York, New York | (28.37, 52428.34) |
| | | Westchester, New York | (432.81, 2021.36) |
| | | Queens, New York | (109.38, 17842.16) |
| Northwest A.B., Alaska (35856.31,0.17) | 0.1489 | Nome Census Area, Alaska | (23008.59, 0.36) |
| | | Yukon–Koyukuk, Alaska | (157094.25, 0.05) |
| | | North Slope B., Alaska | (87845.38, 0.07) |

*Note*. The figures shown in the table are original values (not standardized values).

of Wisconsin, Rolette of North Dakota, Glacier of Montana, Thurston of Nebraska and Petersburg of Virginia are flagged as local outliers by our method. The first four counties are dominated by the population of American Indian, Eskimo, or Aleut. The proportions are 0.942, 0.703, 0.595, 0.464, respectively, while in their neighbouring counties, the proportion of American Indian, Eskimo, or Aleut is very low. The fifth county (Petersburg, Virginia) is dominated by African American, with proportion 0.836.

Even though the proportion of African Americans in Petersburg is higher than the proportion of American Indian, Eskimo, or Aleut in the second, third and fourth counties in the Table 3, the SLOM value is lower than those of these three counties. This is because Petersburg is located in an unstable region—its neighbouring counties also have a high proportion of African Americans.

The top five global outliers flagged by the SLZ method are listed in Table 4. Not surprisingly, Menominee and Rolette appear in this list. This means that they are both global and local outliers.

We also applied a DB-based approach on the nonspatial attributes to detect the top five outliers. None of the counties listed in Table 3 and Table 4 were flagged as outliers.

**Table 3** Top five outliers flagged by SLOM method and their neighbours. The four attributes used in the experiment are the proportions of people identified as African American Indian (including Eskimo, or Aleut), Asian (including Pacific Islander), and of Hispanic origin, respectively. The figures shown in the table are standardized values

| County name (Value in percentage) | SLOM value | Neighbours | Values (Percentage) |
|---|---|---|---|
| Menominee, Wisconsin (0.000, 0.942, 0.000, 0.015) | 0.8822 | Langlade, Wisconsin | (0.001, 0.007, 0.002, 0.005) |
| | | Oconto, Wisconsin | (0.001, 0.007, 0.002, 0.004) |
| | | Shawano, Wisconsin | (0.001, 0.050, 0.003, 0.004) |
| Rolette, North Dakota (0.003, 0.703, 0.002, 0.005) | 0.6506 | Bottineau, North Dakota | (0.001, 0.008, 0.003, 0.002) |
| | | Pierce, North Dakota | (0.000, 0.005, 0.005, 0.000) |
| | | Towner, North Dakota | (0.001, 0.015, 0.002, 0.001) |
| Glacier, Mon. (0.001, 0.595, 0.001, 0.007) | 0.4914 | Flathead, Mon. | (0.001, 0.016, 0.006, 0.011) |
| | | Pondera, Montana | (0.001, 0.116, 0.005, 0.005) |
| | | Toole, Montana | (0.001, 0.025, 0.005, 0.007) |
| Thurston, Nebraska (0.001, 0.464, 0.002, 0.009) | 0.4453 | Monona, Iowa | (0.001, 0.003, 0.002, 0.003) |
| | | Woodbury, Iowa | (0.022, 0.018, 0.020, 0.028) |
| | | Burt, Nebraska | (0.001, 0.009, 0.003, 0.010) |
| | | Cuming, Nebraska | (0.001, 0.001, 0.003, 0.002) |
| | | Dakota, Nebraska | (0.005, 0.019, 0.034, 0.062) |
| | | Dixon, Nebraska | (0.001, 0.002, 0.001, 0.001) |
| | | Wayne, Nebraska | (0.005, 0.003, 0.006, 0.002) |
| Petersburg, Virginia (0.836, 0.002, 0.012, 0.013) | 0.4384 | Chesterfield, Virginia | (0.151, 0.002, 0.028, 0.012) |
| | | Dinwiddie, Virginia | (0.413, 0.002, 0.005, 0.006) |
| | | Prince Georges, Virginia | (0.337, 0.004, 0.034, 0.040) |
| | | Colonial H., Virginia | (0.009, 0.002, 0.035, 0.010) |

**Table 4** Top five outliers flagged by SLZ method and their neighbours on the same data set

| County name (Value in percentage) | Chi-square | Neighbours | Values (Percentage) |
|---|---|---|---|
| Menominee, Wisconsin (0.000, 0.942, 0.000, 0.015) | 268.70 | Langlade, Wisconsin | (0.001, 0.007, 0.002, 0.005) |
| | | Oconto, Wisconsin | (0.001, 0.007, 0.002, 0.004) |
| | | Shawano, Wisconsin | (0.001, 0.050, 0.003, 0.004) |
| Shannon, South Dakota (0.001, 1.000, 0.001, 0.019) | 224.82 | Cherry, Nebraska | (0.001, 0.030, 0.003, 0.004) |
| | | Dawes, Nebraska | (0.007, 0.042, 0.013, 0.012) |
| | | Sheridan, Nebraska | (0.001, 0.082, 0.004, 0.010) |
| | | Bennett, South Dakota | (0.003, 0.488, 0.001, 0.011) |
| | | Custer, South Dakota | (0.002, 0.026, 0.003, 0.007) |
| | | Fall River, South Dakota | (0.004, 0.065, 0.006, 0.017) |
| | | Jackson, South Dakota | (0.001, 0.448, 0.003, 0.005) |
| | | Pennington, South Dakota | (0.018, 0.076, 0.018, 0.022) |
| Buffalo, South Dakota (0.000, 0.820, 0.000, 0.002) | 171.76 | Brule, South Dakota | (0.001, 0.074, 0.003, 0.006) |
| | | Hand, South Dakota | (0.001, 0.001, 0.004, 0.003) |
| | | Hyde, South Dakota | (0.001, 0.036, 0.001, 0.004) |
| | | Jerauld, South Dakota | (0.000, 0.002, 0.005, 0.000) |
| | | Lyman, South Dakota | (0.001, 0.305, 0.001, 0.005) |
| Sioux, North Dakota (0.001, 0.797, 0.005, 0.008) | 163.79 | Adams, North Dakota | (0.001, 0.003, 0.000, 0.000) |
| | | Emmons, North Dakota | (0.000, 0.001, 0.001, 0.001) |
| | | Grant, North Dakota | (0.000, 0.010, 0.002, 0.003) |
| | | Morton, North Dakota | (0.001, 0.019, 0.003, 0.003) |
| | | Campbell, South Dakota | (0.001, 0.002, 0.000, 0.000) |
| | | Corson, South Dakota | (0.000, 0.512, 0.001, 0.012) |
| | | Perkins, South Dakota | (0.002, 0.015, 0.001, 0.004) |
| Rolette, North Dakota (0.003, 0.703, 0.002, 0.005) | 152.59 | Bottineau, North Dakota | (0.001, 0.008, 0.003, 0.002) |
| | | Pierce, North Dakota | (0.000, 0.005, 0.005, 0.000) |
| | | Towner, North Dakota | (0.001, 0.015, 0.002, 0.001) |

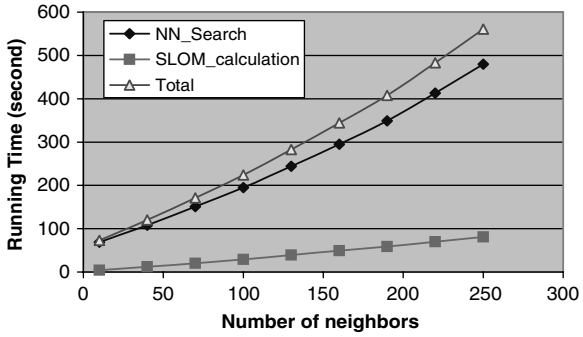



**Fig. 7** The break-up of the total running time into NN search and SLOM value calculation as a function of the number of nearest neighbours

### 4.4 Break-up of the total running time

The total running time of our algorithm mainly consists of two parts: the time to search the nearest neighbours (NN_Search) and the time to calculate the SLOM values

(SLOM_Calculation). The break-up is shown in Fig. 7, from which it is clear that most of the running time is consumed by the nearest-neighbour search.

## 5 Summary and future work

We have proposed a new measure, spatial local outlier measure (SLOM), which captures both spatial autocorrelation and spatial heteroscedasticity (nonconstant variance). The effects of spatial autocorrelation are factored out by a new measure, $\tilde{d}$, which reduces the effects of outliers on its neighbours. The variance of a neighbourhood is captured by $\beta(o)$, which quantifies the oscillation and instability of an area around $o$. The use of $\beta$ instead of standard deviation was motivated by a desire to deterministically bound a variance-like measure. We have compared our approach with the current state-of-the-art methods and have shown that SLOM is sharper in detecting local outliers. Local outliers may be more interesting than global outliers because they are likely to be less known and therefore more surprising. Another novel feature of our approach is related to system integration. The spatial data never leaves the database and we use an R-tree index to carry out nearest neighbour queries directly in the database.

For future work, we would like to apply our method to large climate databases and discover potentially useful patterns like the Southern Oscillation Index (SOI).

**Acknowledgements** Sanjay Chawla is partially supported by an ARC Discovery Research Grant. Pei Sun gratefully acknowledges a CMCRC top-up scholarship.

## Appendix

For object $o$, let

1. $N(o)$ be the neighbourhood of $o$

2. $\max\ d(o) = \max\{\text{dist}(o, p) \mid p \in N(o)\}$
3. $\min\ d(o) = \min\{\text{dist}(o, p) \mid p \in N(o)\}$
4. $\text{sum}(o) = \sum_{p \in N(o)} \text{dist}(o, p)$
5. $\text{sum}(p) = \sum_{q \in N(p)} \text{dist}(p, q)$

**Theorem 3** *Assume every point in the dataset has the same number (n) of neighbours. For any point $p \in N(o)$, if*
*1.* $\max d(p) = \text{dist}(o, p)$, *i.e.* $\max d(p) \geq \min d(o)$
*2.* $\tilde{d}(o) - \tilde{d}(p) > \max d(o) - \min d(o)$
*then* $\tilde{d}(o) - \tilde{d}(p) > d(o) - d(p)$.

*Proof*

$$
\begin{aligned}
&(\tilde{d}(o) - \tilde{d}(p)) - (d(o) - d(p)) \\
&= \left(\frac{\text{sum}(o) - \max d(o)}{n-1} - \frac{\text{sum}(p) - \max d(p)}{n-1}\right) - \left(\frac{\text{sum}(o)}{n} - \frac{\text{sum}(p)}{n}\right) \\
&= \left(\frac{\text{sum}(o) - \max d(o)}{n-1} - \frac{\text{sum}(p) - \max d(p)}{n-1}\right) \\
&\quad - \left(\frac{\text{sum}(o) - \max d(o)}{n} - \frac{\text{sum}(p) - \max d(p)}{n}\right) - \frac{\max d(o) - \max d(p)}{n} \\
&= \frac{1}{n}\left(\frac{\text{sum}(o) - \max d(o)}{n-1} - \frac{\text{sum}(p) - \max d(p)}{n-1}\right) - \frac{\max d(o) - \max d(p)}{n} \\
&= \frac{1}{n}((\tilde{d}(o) - \tilde{d}(p)) - (\max d(o) - \max d(p)))
\end{aligned}
$$

From the condition 1 and 2, we have

$$
\begin{aligned}
&(\tilde{d}(o) - \tilde{d}(p)) - (\max d(o) - \max d(p)) \\
&\geq (\tilde{d}(o) - \tilde{d}(p)) - (\max d(o) - \min d(o)) > 0.
\end{aligned}
$$

Then we have $(\tilde{d}(o) - \tilde{d}(p)) - (d(o) - d(p)) > 0$
i.e. $\tilde{d}(o) - \tilde{d}(p) > d(o) - d(p)$. □

**Theorem 4** *Assume every point in the dataset has the same number (n) of neighbours. For any point, $p \in N(o)$, if*
*1.* $\max d(p) = rmdist(o, p)$, *i.e.* $\max d(p) \geq \min d(o)$
*2.* $\frac{\min d(o)}{\max d(o)} > \frac{\tilde{d}(p)}{\tilde{d}(o)}$
*then* $\frac{\tilde{d}(o)}{\tilde{d}(p)} > \frac{d(o)}{d(p)}$.

*Proof*

$$
\begin{aligned}
\frac{\frac{d(p)}{\tilde{d}(p)}}{\frac{d(o)}{\tilde{d}(o)}} &= \frac{\frac{\frac{\text{sum}(p)}{n}}{\frac{\text{sum}(p) - \max d(p)}{n-1}}}{\frac{\frac{\text{sum}(o)}{n}}{\frac{\text{sum}(o) - \max d(o)}{n-1}}} \\
&= \frac{\frac{\text{sum}(p)}{\text{sum}(p) - \max d(p)}}{\frac{\text{sum}(o)}{\text{sum}(o) - \max d(o)}} \\
&= \frac{\frac{\text{sum}(p) - \max d(p) + \max d(p)}{\text{sum}(p) - \max d(p)}}{\frac{\text{sum}(o) - \max d(o) + \max d(o)}{\text{sum}(o) - \max d(o)}} \\
&= \frac{\left(1 + \frac{\max d(p)}{\text{sum}(p) - \max d(p)}\right)}{\left(1 + \frac{\max d(o)}{\text{sum}(o) - \max d(o)}\right)}.
\end{aligned}
$$

From the conditions 1 and 2, we have

$$\frac{\max d(p)}{\max d(o)} \geq \frac{\min d(o)}{\max d(o)} > \frac{\tilde{d}(p)}{\tilde{d}(o)}$$

$$\frac{\max d(p)}{\max d(o)} > \frac{\tilde{d}(p)}{\tilde{d}(o)} = \frac{(\text{sum}(p) - \max d(p))/(n-1)}{(\text{sum}(o) - \max d(o))/(n-1)}$$

$$\frac{\max d(p)}{\text{sum}(p) - \max d(p)} > \frac{\max d(o)}{\text{sum}(o) - \max d(o)}.$$

Then

$$\frac{\frac{d(p)}{\tilde{d}(p)}}{\frac{d(o)}{\tilde{d}(o)}} > 1, \text{ i.e.}$$

$$\frac{\tilde{d}(o)}{\tilde{d}(p)} > \frac{d(o)}{d(p)}.$$

□

## References


1. Aggarwal CC, Yu PS (2001) Outlier detection for high dimensional data. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Santa Barbara, California, USA
2. Angiulli F, Pizzuti C. (2002) Fast outlier detection in high dimensional spaces. In: Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)
3. Bay SD, Schwabacher M (2003) Mining distance-based outliers in near linear time with randomisation and a simple pruning rule. In: Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
4. Breunig MM, Kriegel HP, Ng RT, Sander J (2000) LOF: Identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104. Dallas, Texas, USA
5. Hawkins D (1980) Identification of outliers. Chapman and Hall, London
6. Knorr EM, Ng RT (1998) Algorithms for mining distance-based outliers in large datasets. In: Proceedings of 24th International Conference on Very Large Data Bases, pp. 392–403. New York City
7. Lu CT, Chen DC, Kou YF (2003a) Algorithms for spatial outlier detection. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), pp. 597–600. Melbourne, Florida
8. Lu CT, Chen DC, Kou YF (2003b) Detecting spatial outliers with multiple attributes. In: Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2003), pp 122–128. Sacramento, California
9. McPhadden M (2002) El Nino and La Nina: Causes and global consequences. Encyclopedia of Global Environmental Change, pp. 353–370
10. Papadimitriou S, Kitagawa H, Gibbons PB, Faloutsos C (2003) LOCI: Fast outlier detection using the local correlation integral. In: Proceedings of the 19th International Conference on Data Engineering, pp. 315–328. Bangalore, India
11. Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large datasets. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 427–438. Dallas, Texas
12. Shekhar S, Chawla S (2003) Spatial databases: A tour. Prentice Hall
13. Shekhar S, Lu CT, Zhang PS (2001) Detecting graph-based spatial outliers: Algorithms and applications (a summary of results). In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 371–376, San Francisco

14. Shekhar S, Lu CT, Zhang PS (2003) A unified approach to detecting spatial outliers. GeoInformatica, **7**(2), 139–166
15. Wilcox R (2003) Applying contemporary statistical techniques. Elsevier Science

**Sanjay Chawla** is a Senior Lecturer in the School of Information Technologies at the University of Sydney. His research interests span the area of data mining and spatial database management. He is a co-author of the textbook "Spatial Databases: A Tour", which is published by Prentice Hall. His research work has appeared in leading publications, including IEEE Transaction on Knowledge and Data Engineering and GeoInformatica. He received his Ph.D. in Mathematics from the University of Tennessee, USA.

**Pei Sun** is currently a Ph.D. student in the School of Information Technology, Sydney University, Australia. His research interests include data mining and spatial database. He received his M.E. degree from the University of New South Wales, Sydney, Australia, in 2002 and a B.E. degree from Beijing Forestry University, China, in 1990.