On Local Spatial Outliers

Pei Sun
University of Sydney
School of Information Technologies
Sydney, NSW, Australia
psun2712@it.usyd.edu.au

Sanjay Chawla
University of Sydney
School of Information Technologies
Sydney, NSW, Australia
chawla@it.usyd.edu.au

Abstract

We propose a measure, Spatial Local Outlier Measure (SLOM) which captures the local behaviour of datum in their spatial neighborhood. With the help of SLOM we are able to discern local spatial outliers which 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 supresses the reporting of outliers in highly unstable areas, where data is 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 which show that our approach is novel and scalable to large data sets.

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 which 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 meterologist 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 Osscillation 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 Figure 1(Reprinted from [6]), 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 degrees above normal and that 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.

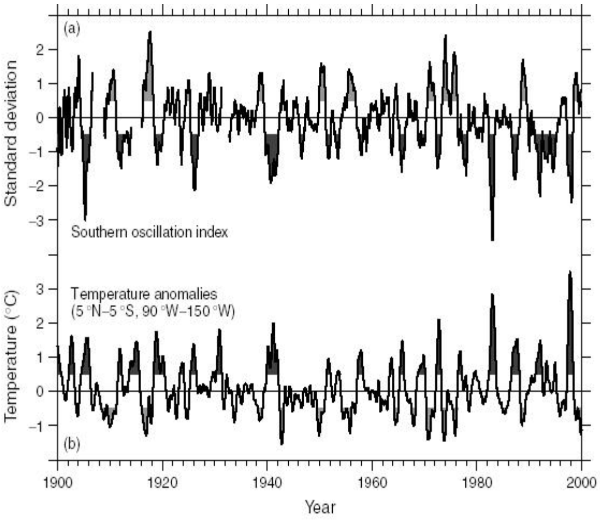


Figure 1. The relationship between the Southern Oscillation Index(SOI) and sea surface temperature. High tempeature anomolies correspond to El Niño and low to La Niña. The relationship was discovered by Sir Gilbert Walker and clearly shows how outlier detection can provide penetrating insights about the underlying phenomenon, global weather patterns in this case [6]

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

The classic definition of an outlier is due to Hawkins [2] who defines *"an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a diferrent mechanism."* Several different approaches has been taken in order to operationalize this definition. For example, it is standard to use variations of the Chebyshev's inequality,

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

where μ and σ 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 [3] were the first to propose the definition of distance-based outlier, which was free of any distributional assumptions and was readily generalizable to multi-dimensional dataset. 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 lie 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 μ and standard deviation δ, and t ∈ 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 distribution including the Possion.

For spatial data, both statistical and data mining approaches have to be modified because of the qualitative difference between spatial and non-spatial dimensions. The attributes which comprise the non-spatial dimensions intrinsically characterize the data while the spatial dimensions provide a locational index to the object and are not instrinsic to the object. However, the physical neighborhood plays a very important role in analysis of spatial data. For example, in Figure 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. [7] proposed the following definition of spatial outlier: "*A spatial outlier is a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighborhood.*"

A spatial neighborhood 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 non-spatial attributes.

There are two types of spatial outlier: multi-dimensional space-based outliers and graph-based outliers. The only difference between them is that they use different spatial neighborhood definitions. Multi-dimensional space-based outliers use Euclidean distances to define spatial neighborhoods, while graph-based outliers use graph connectivity.

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 neighborhood of o. Now, a Chebyshev inequality like approach, can be undertaken in order to identify those points o which are candidate outliers. Indeed this is the state of the art [9, 8, 4, 5].

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 Figure 2. Clearly this point is a local outlier as it is forms a local maxima in its neigborhood 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.

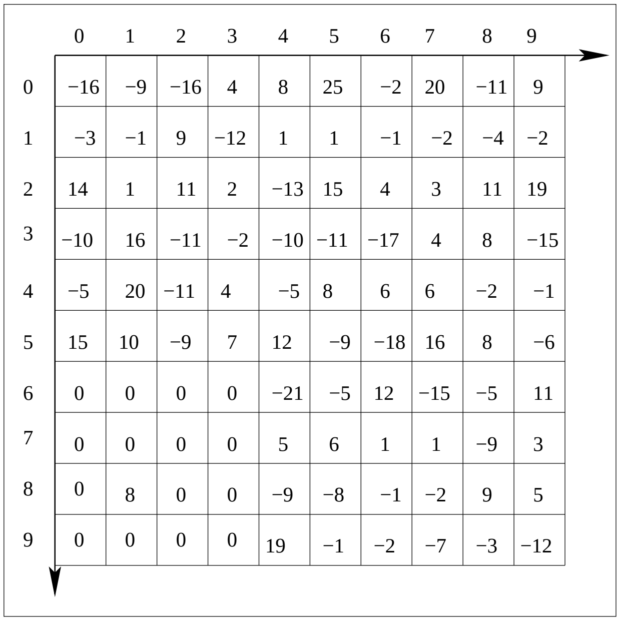
Thus clearly an approach is needed which can efficiently capture spatial local outliers. In fact our method will go further and associate a SLOM score with each data point. The SLOM defines the "degree of outlierness" of each point very much along the lines proposed by Breunig et. al. [1]. However besides the qualitative difference between spatial and non-spatial attributes, spatial data exhibits spatial autocorrelation (non-independence) and heteroscedasticity(non-constant variance) both of which must be factored into SLOM.

1.1 Problem Definition

Given: A large spatial database with multi-dimensional non-spatial attributes.
Design: A measure which 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 neighbors.
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.



	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

Figure 2. Original data matrix

1.2 Key Insights and Contributions

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

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

  where $dist(o, p)$ is a definition of (euclidean) distance between the non-spatial components of o and p and $N(o)$ are the neighboring points of o.

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

  First, define $maxd(o) = max\{dist(o, p)|p \in N(o)\}$ as the maximum non-spatial distance between o and its neighbors. Then define

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

  Now notice that for the point 8 (location (8,1)) in Figure 2, $\tilde{d}(o) = d(o)$ but for points in the neighborhood 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 neighborhood or in other words stretches the difference between an outlier and its neighbors.

  Thus $\tilde{d}$ stretches the difference between an outlier and its neighbors compared to d.

  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 [10]. The trimmed mean is less sensitive to outliers like the median but retains some of the averaging behavior of the mean.

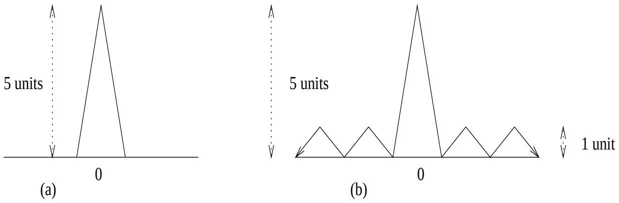


Figure 3. Both (a) and (b) have the same $\tilde{d}$ value however the the β values in (a) is higher than (b) because of the instability around (b)

- The second insight that underpins our approach is that outliers which 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 which can be deterministically bounded. In particular we have defined a statistic β which captures the net oscillation with respect to the average value around o (details in Section 2). For example, Figure 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 Figure 3(a) is higher than Figure 3(b).

- 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 Section 2, we introduce a series of definitions which will culminate in the definition of the Spatial Local Outlier Measure (SLOM). Along the way we will explain how each component of SLOM address spatial autocorrelation and heteroscedasticity. In Section 3, we analyze 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 Section 4, we report the results of our experiments on synthethic and real data sets. In Section 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 which can capture both spatial autocorrelation and heteroscedasticity (non-constant variance). We have already defined $\tilde{d}$ in Section 1 which factors out the effect of spatial autocorrelation and now define β which penalizes for oscillating behavior around a potential outlier.

For an object o, we can define its SLOM value as $\frac{\sum_{p \in N(o)} \frac{\tilde{d}(o)}{\tilde{d}(p)}}{|N(o)|}$, just like the method used in [1]. 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 neighborhood.

- Let $N_+(o)$ denote the set of all the objects in o's neighborhood and o itself and $avg(N_+(o)) = \frac{\sum_{p \in N_+(o)} \tilde{d}(p)}{|N_+(o)|}$.

- 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 neighborhood 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 neighborhood, if all neighbors 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 $avg(N_+(o))$. We calculate $\beta(o)$ using the following pseudo-code.

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

While step one to four are self-explanatory, we explain step 5 and 6. There are two reasons why we divide $\beta(o)$ by $|N_+(o)|-2$ in step 5. First, we need to correct for boundary terms where the number of neighbors is fewer than that in the interior. The second motivation is that for a local region like that in Figure 2 where the data value 8 at location $o = (8,1)$ is surrounded by constant values, $\beta(o) = 1$, the highest value β can assume.

However, if we have stopped at step 5 then β cannot distinguish between the two cases shown in Figure 3. In order to do that we divide $\beta(o)$ by $1 + avg\{\tilde{d}(p)|p \in N(o)\}$. This allows us to penalize the situation where large values of $\tilde{d}$ exist around the point o. However in order to bound this term we have to normalize the original data so that the maximum value that denominator can assume is $1 + \sqrt{d}$, where d is the dimensionality of the non-spatial attributes. Thus in Figure 3(a) and (b) the β values are 1 and 0.5 respectively.

- We are ready to define SLOM. For a point o,

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

A high value of SLOM indicates that the point is 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 β term is analogous to the standard deviation of the first derivative of a smooth random variable.

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

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

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

Proof: The value of $\tilde{d}(o)$ is between 0 and $\sqrt{d}$, and 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 [3] the key step is a method to search for nearest neighbors. 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 neighborhood 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 neighbors($k \leq 8$ for a 2D grid), the calculation of SLOM for the full data set involves the following steps.

1. The first step is to normalize 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 catalog. Thus this step can be done in one database pass and the computational cost is $O(Nd)$, where d is the number of non-spatial dimensions.

2. To compute $\tilde{d}(o)$ we need to find the spatial neighbors of each object and calculate the distance between them. The cost of a single k-NN query using an R-tree is $O(klogN)$ and the cost of computing the nonspatial distance is $O(kd)$. Thus the cost of this step is $O(NklogN + kdN)$.

3. After the computation of $\tilde{d}(o)$, we need to compute $\beta(o)$. This involves another round of nearest neighbor queries followed by a summation to compute the neighborhood average of $\tilde{d}(o)$. The cost of this step is thus $O(NklogN + dN)$.

4. To compute the SLOM we multiply the $\tilde{d}$ and β 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 $O(NlogN)$.

6. Thus the final cost of the whole operation is $O(NklogN + kdN)$.

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 neighbor search, finding the k nearest neighbors 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).

	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

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

The first option is that we store the neighborhood 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 neighborhood information and store it into a table beforehand. When computing $\tilde{d}$ and β, we visit this table instead of the original table that stores the spatial information. Since spatial data enjoys slow updates this is a very attractive option and can result in huge savings in the running time as we can amortize the cost of creating the neighborhood 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 supress the reporting of *global outliers* in unstable areas. Intuitively a point is a global outlier, if its determination, that it is an outlier, depends upon 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 comparsion with points in its neighborhood.

2. Compare the SLOM approach with the the family of

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

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

	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

Figure 5. The SLOM matrix

	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

Figure 6. Result from the SLZ method

methods to discover spatial outliers proposed in [9, 8, 4].

3. Test how the running time changes as we vary the number of nearest neighbors 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 which involve spatial relationships. We used Oracle9i to store all spatial and non-spatial 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 neighbors. We use the following SQL statement to generate the neighborhood 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 a R-tree index is created on it to speed up the processing of this query.

If the spatial objects are points, and the neighborhood is defined to be the k nearest neighbors, then the following SQL statement can be use to generate the neighborhood information.

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

4.1 Result on a Synthetic Dataset

We have created a synthetic data set consisting of one non-spatial attribute in order to explain and compare our method with the prototype method proposed in [8, 9, 5]. We

County Name	SLOM value	Area	Pop. den.	Neighbor	area	Pop. den.
Yukon-Koyukuk,AK	0.2896	157094.25	0.05	Bethel Census Area,AK	41080.34	0.33
				Wade Hampton,AK	17121.14	0.34
				Southeast Fairbanks,AK	25989.64	0.23
				Fairbanks North S. B.,AK	7361.16	10.56
				Matanuska-Susitna B.,AK	24689.41	1.61
				Nome Census Area,AK	23008.59	0.36
				North Slope B.,AK	87845.38	0.07
				Northwest Arctic B.,AK	35856.31	0.17
Philadelphia,PA	0.1884	135.11	11735.78	Burlington,NJ	804.63	490.99
				Delaware,PA	184.19	2973.23
				Montgomery,PA	483.06	1403.78
				Bucks,PA	607.54	890.76
				Camden,NJ	222.29	2261.98
				Gloucester,NJ	324.81	708.36
Suffolk,MA	0.1831	58.51	11347.16	Essex,MA	497.98	1345.59
				Norfolk,MA	399.54	1542.01
				Middlesex,MA	823.40	1698.40
Bronx,NY	0.1633	42.02	28645.75	Bergen,NJ	234.16	3524.80
				Nassau,NY	286.72	4489.89
				New York,NY	28.37	52428.34
				Westchester,NY	432.81	2021.36
				Queens,NY	109.38	17842.16
Northwest A. B.,AK	0.1489	35856.31	.17	Nome Census Area,AK	23008.59	0.36
				Yukon-Koyukuk,AK	157094.25	0.05
				North Slope B.,AK	87845.38	0.07

Table 2. Top five outliers and their neighbors

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 neighborhood 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 one hundred spatial objects organized as a 10×10 matrix. We used a Gaussian generator to produce the values of non-spatial attribute, and they are listed in Figure 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 percent) and SLOM are listed in Table 1. The $\tilde{d}$, SLOM and the SLZ matrices are shown in Figures 4, 5 and 6 respectively. The objects located at position (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 SLZ. This means they are local but not global outliers. Again their SLOM values are 0.17 and 0.25 respectively.

4.2 Result on a Real Dataset

The real data set that we have used is from the U.S Census Bureau and consists of spatial and non-spatial information about all the counties in the United States. In order to make the results easily comprehensible we have selected two non-spatial attributes: area and population density. The informaton about the top five outliers counties and their neighbors is listed in Table 2. Not surprisingly the top 5 outliers consist of counties which 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 neighboring 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 to the Bronx even though it has a bigger area and smaller popula-

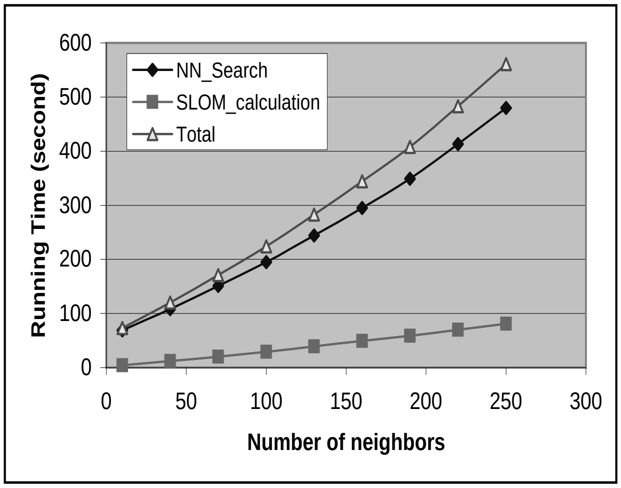


Figure 7. The break-up of the total running time into NN search and SLOM value calculation as a function of the number of nearest neighbors

tion density again because its neighborhood is relative more stable.

4.3 Break-up of the total running time

The total running time of our algorithm mainly consists of two part: the time to search the nearest neighbors (NN_Search) and the time to calculate the SLOM values ($SLOM\_Calculation$). The break-up is shown in Figure 7 from which it is clear that most of the running time is consumed by the nearest neighbor 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 (non-constant variance). The effects of spatial autocorrelation are factored out by a new measure $\tilde{d}$ which reduces the effects of outliers on its neighbors. The variance of a neighborhood is captured by $\beta(o)$ which quantifies the oscillation and instability of an area around o. The use of β, 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 neighbor 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).

References

[1] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA*, pages 93–104. ACM, 2000.

[2] D. Hawkins. *Identification of Outliers*. Chapman and Hall, London, 1980.

[3] E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, pages 392–403. Morgan Kaufmann, 1998.

[4] C.-T. Lu, D. Chen, and Y. Kou. Algorithms for spatial outlier detection. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA*, pages 597–600. IEEE Computer Society, 2003.

[5] C.-T. Lu, D. Chen, and Y. Kou. Detecting spatial outliers with multiple attributes. In *Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2003), 3-5 November 2003, Sacramento, California, USA*, pages 122–128. IEEE Computer Society, 2003.

[6] M. McPhadden. El nino and la nina: Causes and global consequences. In *Encyclopedia of Global Environmental Change*, pages 353–370, 2002.

[7] S. C. Shashi Shekhar. *Spatial Databases: A Tour*. Prentice Hall, 2003.

[8] S. Shekhar, C.-T. Lu, and P. Zhang. 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, August 26-29, 2001, San Francisco, CA, USA. ACM, 2001*, pages 371–376, 2001.

[9] S. Shekhar, C.-T. Lu, and P. Zhang. A unified approach to detecting spatial outliers. *GeoInformatica*, 7(2):139–166, 2003.

[10] R. Wilcox. *Applying Contemporary Statistical Techniques*. Elseiver Science, 2003.