Metodologías de detección de *outliers* en datos espaciales, temporales y espacio-temporales

Isabel Blasco Fernández*

Recibido el 9 de febrero de 2018; aceptado el 30 de marzo de 2018

Resumen

En la obtención de conjuntos de datos se pueden detectar registros con valores mucho mayores o menores a los usuales. Dichos registros, denominados *outliers*, pueden ser correctos, o ser el resultado de errores de captura o del procesado de los datos. El estudio y la detección de *outliers* ayuda a revelar información valiosa de los conjuntos de datos, así como a depurar las bases de datos de información que pueda ser errónea. En este trabajo se describen distintos métodos de detección de *outliers* propuestos recientemente y aplicados al marco espacial y espacio-temporal, junto con los resultados de su aplicación. Entre las propuestas se abordan métodos puramente espaciales, puramente temporales y otros mixtos que han demostrado su valía en ensayos controlados. Entre otros se considerarán: 1) emplear un algoritmo que tiene en cuenta los datos temporales y combina las ventajas del *clustering* y la aproximación basada en la densidad; 2) comparar el valor obtenido por un sensor con el valor esperado calculado de forma incremental, al tener en cuenta la correlación temporal de los datos que muestran una correlación espacial en el pasado reciente; 3) plantear un algoritmo que detecta *outliers* en grandes bases de datos espacio-temporales a partir del uso conjunto de la información espacial, no espacial y valores temporales; 4) detectar *outliers* en bases de datos temporales a partir de reglas de asociación extraídas del comportamiento normal de los objetos de un conjunto de datos, de forma que se definen relaciones entre los atributos y el tiempo; 5) detectar los *outliers* en las propiedades globales del conjunto de datos, en contraste con la mayoría de métodos existentes que lo aplican de forma local y 6) asignar, para el caso específico de los *outliers* espaciales, diferentes pesos para diferentes vecinos a la hora de calcular el objeto central, determinando el peso a través de relaciones espaciales tales como la distancia. A partir de los diferentes

* Ingeniero Técnico Forestal y Master en Ingeniería Geodésica y Cartografía por la Universidad Politécnica de Madrid, correo electrónico: isablascofer@gmail.com.

métodos expuestos, se evidenciará la importancia de aplicar un método u otro en función de los conjuntos de datos a analizar, así como la necesidad de continuar avanzando en la mejora de la eficacia, fiabilidad y rapidez de los métodos de detección de *outliers*.

Palabras clave: *Outlier, Espacio-temporal, Clúster*.

Resumo

Na obtenção de conjuntos de dados é possível detectar registros com valores muito maiores ou menores aos usuais. Estes registros, denominados *outliers*, podem ser corretos, ou ser o resultado de erros de captura ou de processamento dos dados. O estudo e a detecção de *outliers* ajuda a revelar informação valiosa dos conjuntos de dados, assim como a depurar as bases de dados de informação que possa ser errônea. Neste trabalho se descrevem distintos métodos de detecção de *outliers* propostos recentemente e aplicados ao marco espacial e espaço-temporal, junto com os resultados de sua aplicação. Entre as propostas se abordam métodos puramente espaciais, puramente temporais e outros mistos que têm demonstrado seu valor em ensaios controlados. Entre outros, se considerou: 1) empregar um algoritmo que tem em conta os dados temporais e combina as vantagens do *clustering* e a aproximação baseada em densidade; 2) comparar o valor obtido por um sensor com o valor esperado calculado de forma incremental, ao se ter em conta a correlação temporal dos dados que mostram uma correlação espacial no passado recente; 3) criar um algoritmo que detecta *outliers* em grandes bases de dados espaço-temporais a partir do uso conjunto da informação espacial, não espacial e valores temporais; 4) detectar *outliers* em bases de dados temporais a partir de regras de associação extraídas do comportamento normal dos objetos de um conjunto de dados, de forma que se definem relações entre os atributos o tempo; 5) detectar os *outliers* nas propriedades globais do conjunto de dados, em contraste com a maioria dos métodos existentes que se aplicam de forma local e 6) alocar, para o caso específico dos *outliers* espaciais, diferentes pesos para diferentes vizinhos na hora de calcular o objeto central, determinando o peso através de relações espaciais tais como a distância. A partir dos diferentes métodos expostos, se evidenciará a importância de se aplicar um método ou outro em função dos conjuntos de dados a analisar, assim como a necessidade de continuar avançando na melhora da eficácia, confiabilidade e rapidez dos métodos de detecção de *outliers*.

Palavras chave: *Outlier, Espaço-temporal, Cluster*.

Abstract

While you are recording datasets it is possible to notice values which are substantially larger or smaller than usual. Such registers, named outliers hereinafter, might be correct,

or might arise after a data recording error or a processing error. Research on outlier detection might provide useful information about the dataset, while being helpful in the database cleansing operation. In this paper we describe different methodologies recently proposed for outlier detection, intended for spatial and spatio-temporal dataset, as well as case results. Among them, we distinguish between those purely spatial, purely temporal, and mixed ones that have proven its value in controlled experiments. Among others we will discuss: 1) use an algorithm that considers temporal data while combines the advantages of clustering and approximations based upon kernel densities 2) compare the measured value with its expected value calculated in a incremental fashion considering temporal correlation, thus showing a spatial correlation in the recent past 3) propose a new algorithm for outlier detection for large spatio-temporal databases using spatial and aspatial information as well as temporal information 4) detect oultiers in tempral databases using association rules extracted from the normal behaviour and linking temporal evolution of the atributes 5) detect the outliers based upon the global properties of the dataset unlike most of the prevailing methods which have a local span and 6) assign, for the specific case of spatial outliers, different weights for different neighbors in order the estimate the central value, and estimating the weights as functions of the distance. After analyzing the different methods described, the importance of selecting one method or another will be shown. It is necessary to prosecute the research on the reliability and speed of the outlier detection algorithms.

Key words: *Outlier, Temporal-space, Clúster.*

Introducción

La identificación de casos que, por alguna razón, no encajan bien con el resto del conjunto de datos —*outliers* (Schuber, Wojdanewski, Zimer y Kriegel, 2012), ha recibido recientemente gran atención en muchos ámbitos de la ciencia. Aunque los *outliers* son, por definición, poco frecuentes, su importancia es elevada en comparación con otros eventos (Yang, Latecki and Pokrajac, 2009). En los últimos años, los avances tecnológicos en la recopilación de datos han facilitado la obtención masiva de información, por lo que la detección de valores inusuales o valores erróneos está adquiriendo progresivamente un mayor interés en el análisis de los conjuntos de datos. A modo de ejemplo, cuando se parte de unos datos brutos para realizar una investigación, estos pueden estar contaminados por información errónea que puede ser consecuencia del mal estado del instrumento de medida. Si de esa información no se extraen los *outliers*, los resultados del estudio podrían quedar falseados. Esto explica por qué la detección de *outliers* es un área de investigación muy activa con nuevos métodos propuestos cada año, sobre la base de diferentes metodologías como el razonamiento estadístico (Hadi, Rahmatullah and Werner, 2009), las distancias (Angulli and Pizzuti, 2001; Knorr *et al.*, 2000; Orair *et al.*, 2010; Ramaswamy *et al.*, 2000;

Vu and Gopalkrishnan, 2009; Zhang *et al.*, 2009) o las densidades (Breunig *et al.*, 2000;Vries *et al.*, 2010; Keller *et al.*, 2012; Kriegel *et al.*, 2009).

A continuación, se presentan seis métodos de detección de *outliers* desarrollados en los últimos años, así como ejemplos de su aplicación, dejando a un lado los métodos tradicionales descritos por ejemplo en (Ben-Gal, 2005). Por último, se efectúa una reflexión final sobre los métodos presentados y, en general, sobre los *outliers.*

Métodos novedosos de detección de outliers

En el presente trabajo se exponen seis métodos desarrollados en los últimos años por diferentes autores para detectar *outliers* (Yang *et al.*, 2009; Cheng *et al.*, 2009; Ap-pice *et al.*, 2014; Birant *et al.*, 2006; Bruno *et al.*, 2010; Yufeng, 2006). En primer lugar, se describe el modelo STARIMA; después se describe el método SWOD; a continuación se propone el algoritmo de detección de *outliers* ST; luego se propone el algoritmo TOD, seguimos con una variante del algoritmo EM; y finalmente se presen-ta un algoritmo de detección de *outliers* por ponderación espacial.

Modelo STARIMA (*Space-Time Autoregressive Integrated Moving Average*)

STARIMA (Cheng *et al.*, 2009) es un modelo espacio-temporal dinámico que expresa cada observación en el tiempo *t* y la ubicación *i* como una combinación lineal ponderada de las observaciones anteriores y sus observaciones vecinas desplazadas tanto en el espacio como en el tiempo. STARIMA se define con la ecuación 1.

$$z_i(t) = \sum_{k=1}^{p}\sum_{h=0}^{m_k} \phi_{kh} W^{(h)} z_i(t-k) - \sum_{l=1}^{q}\sum_{h=0}^{n_l} \theta_{lh} W^{(h)} z_i(t-l) + \varepsilon_i(t) \quad (1)$$

Donde p es el orden autorregresivo, q el orden de media móvil, m_k es el orden espacial del término k-enésimo autorregresivo, n_l es el orden espacial del término l-enésimo medida móvil, Φ_{kh} es el parámetro autorregresivo del desfase temporal k y del desfase espacial h, θ_{lh} es el parámetro medio de movimiento en el desfase temporal l y en el desfase espacial h, $W^{(h)}$ es la matriz de pesos N × N para el orden espacial h, y $\varepsilon_i(t)$ es un error aleatorio normal distribuido en un tiempo t y en una ubicación i.

Por su diseño STARIMA captura la estructura de autocorrelación del espacio-tiempo lineal de los datos de series espacio-temporales. Además cuenta con un operador de desfase que es la representación de la dependencia del espacio-tiempo, lo que indica que cada observación $z_i(t)$ en una ubicación i y en un momento t determinado, está influenciado tanto por las series de tiempo anteriores de esa ubicación como por las series temporales de sus vecinos espaciales.

La dependencia espacio-temporal se mide por la función de autocorrelación espacio-temporal (ST-ACF) y el espacio-tiempo con la función de autocorrelación parcial (ST-FAP). A partir del cálculo de la ST-ACF y ST-FAP, se está en condiciones de definir el tiempo de desfase (vecino temporal) y el desfase espacial (vecino espacial).

El orden autorregresivo *p* y el orden movimiento medio *q* del modelo STARIMA se eligen provisionalmente después de examinar la autocorrelación espacio-temporal (STACF) y las funciones parciales de autocorrelación espacio-temporal (ST-FAP).

El análisis está basado en un semivariograma que es capaz de describir las variaciones espaciales y temporales, así como determinar la magnitud de la dependencia espacial y el rango de la correlación espacial entre los datos. Los valores dentro del rango estarán espacialmente autocorrelacionados (se pueden considerar vecinos espaciales), mientras que los casos de fuera se considerarán independientes.

Para comprobar la eficacia del procedimiento, los autores utilizaron los datos medios anuales de temperaturas en China con una serie de 52 años, de 1951 a 2002, medidas en 192 estaciones. La primera zona espacial se determinó en 1 550 km utilizando un modelo de semivariograma isotrópico. Mediante el uso de autocorrelación espacio-temporal y las funciones de autocorrelación parcial de STARIMA, el desfase espacial se determina como 1 y el desfase temporal se determina en 2.

Se realizó una investigación de la relación entre año y temperatura para cada estación, obteniéndose una regresión lineal que los autores utilizan para ajustar los datos. Si el residuo para un determinado año es superior a tres veces la desviación estándar de los residuos, este valor se señala como *outlier*, como se puede ver en la Figura 1.

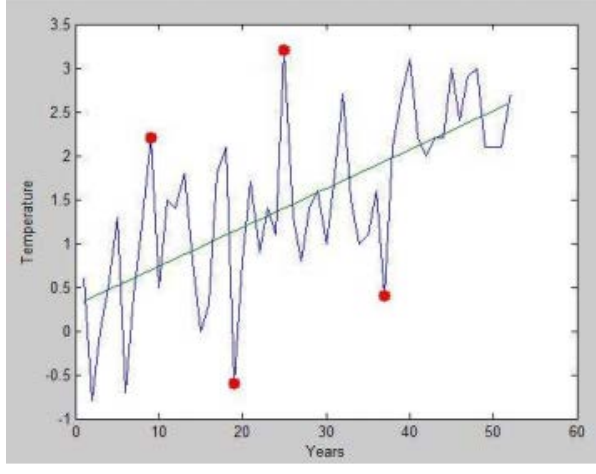

Figura 1. Posibles outliers detectados por el modelo (marcados en rojo). La línea verde muestra la regresión lineal de la estación. Los años de 1 a 51 son los correspondientes a los años 1951 y 2002.
Fuente: Cheng *et al.*, 2009.

El siguiente paso es validar los *outliers*. Para ello se construye el modelo del vecino espacio-temporal (STN) incluyendo observaciones de dos años anteriores cuya distancia a la estación sea inferior a 1 550 km. Después se construye el modelo STN de cada posible *outlier* comparando su temperatura con la $\mu + k\sigma$ y con la $\mu - k\sigma$ de los valores de temperatura de su STN, siendo μ la media y σ la desviación estándar de los valores de temperatura de STN. El parámetro *k* determina la credibilidad del *outlier*. A medida que k aumenta, los valores extremos detectados serán más distantes en comparación con los valores de *k* menores. El *outlier* se valida como STO (*outlier* espacio-temporal) si su valor de temperatura es superior a $\mu + k\sigma$ o inferior a $\mu - k\sigma$. En la Figura 2 se observa un ejemplo de los resultados.

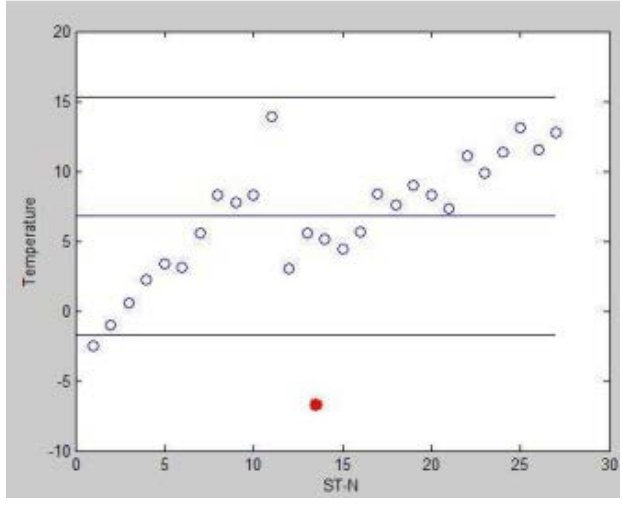

Figura 2. Muestra la verificación de un outlier cuando k es 2. La línea azul horizontal es la medida del STN del valor posible del outlier (punto rojo). Las líneas negras muestran los límites superior e inferior cuando se añade la desviación estándar k veces y se resta y suma a la media.
Fuente: Cheng *et al.*, 2009.

Los autores señalan que una de las ventajas del modelo STARIMA es que su definición se basa en resultados empíricos. Y aunque no lo mencionen, el método presume una distribución normal para la discrepancia entre el dato y su estimador.

SWOD (*Sliding Window Outlier Detector*)

Es un método semi-supervisado para detectar y clasificar los *outliers* que tiene en cuenta la información espacial y temporal. Utiliza modelos que predicen los próximos datos y compara restando los datos previstos y los datos reales para alertar de la existencia de *outliers*. Para ello en cada sensor *u* se compara cada observación $z_u(t_i)$ con su

predicción $\check{z}_u(t_i)$. Si $|| z_u(t_i) - \check{z}_u(t_i)|| > \delta+\varepsilon$, u es considerado un *outlier*, siendo δ el umbral de similitud definido por el usuario y ε es la amplitud del intervalo de confianza de la serie de errores residuales mantenida a lo largo del tiempo de la ventana deslizante para el modelo seleccionado.

El proceso del método SWOD (Appice *et al.*, 2014) aparece representado en la Figura 3. En él se puede ver que realiza una agrupación, que está construida de forma incremental con el modelo de ventana deslizante y se aplica para hacer la predicción de los datos en cada fila (que representa el tiempo t_i). Los datos que se alejan mucho de las predicciones se etiquetan como *outliers*. Estos son clasificados a su vez como cambios o anomalías, siendo estas últimas corregidas.

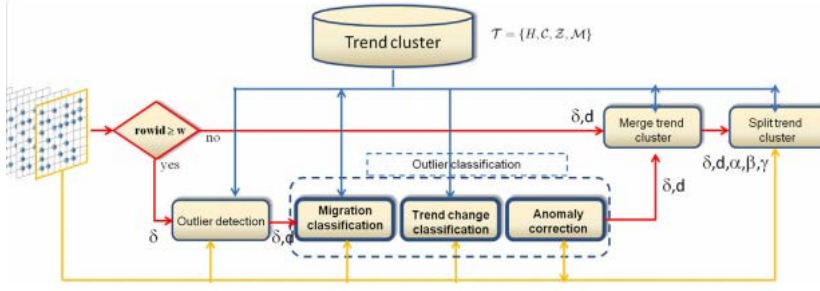

Figura. 3. Esquema del proceso del método SWOD.
Fuente: Appice *et al.,* 2014.

En la Figura 4 se muestra un ejemplo de clasificación de *outliers* y de cómo las anomalías se van intentando recolocar de un grupo de datos a otro.

El modelo de clúster se actualiza de acuerdo con una estrategia de aprendizaje progresivo que opera con cada anomalía corregida en una fila y con cada sensor que se reasigna a una agrupación.

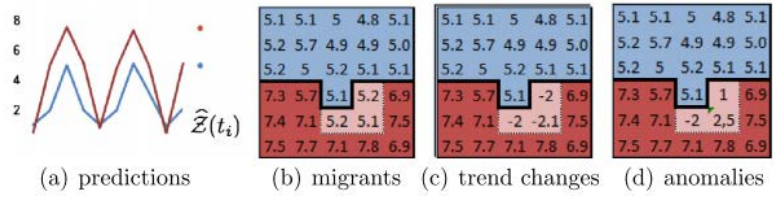

Figura 4. Clasificación de *outliers* según el modelo SWOD. (a) ž (t_i) es la predicción sugerida por el modelo para la serie histórica (azul). (b) los datos correlacionados del área punteada (rosa) migran hacia la zona azul. (c) los datos correlacionados del área punteada (rosa) cambian de acuerdo con el modelo actualizado. (d) los datos no correlacionados en el área punteada son anomalías.
Fuente: Appice *et al.*, 2014.

Los autores han ensayado el modelo SWOD con datos artificiales (para evaluar tanto la precisión del proceso de clasificación de *outliers* en comparación con otros métodos, como la capacidad de pronóstico del modelo) y con datos reales a partir de redes de varios tamaños. Los métodos utilizados para comparar fueron: 1) TSA (*Time Series Analysis*), que identifica los datos que se alejan significativamente de las predicciones calculadas con un modelo de predicción (ARIMA o el modelo suavizado de predicción) y aprende sensor por sensor(detalles en Sharma *et al.*, 2010); 2) método basado en reglas heurísticas a la restricción de datos (detalles en Ramanathan *et al.*, 2006); 3) SSA (*Segmented Sequence Analysis*), que compara el modelo lineal a trozos de un nuevo dato del segmento con un modelo (temporal y espacial) de referencia (detalles en Yao *et al.*, 2010); y 4) método de detección de anomalías y cambios (A+C), que utiliza ARIMA para detectar valores atípicos y una prueba paramétrica con base en la tasa de error, para filtrar los falsos *outlier* que pertenecen a un periodo de cambio (detalles en Pechenizkiy *et al.*, 2009).

Tras realizar las comprobaciones, los resultados de clasificación revelan que SWOD tiene un desempeño muy superior a otros modelos, independientemente de la regularidad y tamaño de la red utilizada. Además contribuye a comprobar empíricamente la idea analizar conjuntamente la correlación espacial y temporal, puede filtrar con eficiencia falsas anomalías sin disminuir la capacidad de detectar las verdaderas.

ST-Outlier Detection Algorithm

Los autores (Birant *et al.*, 2006) proponen un enfoque en tres pasos para detectar *outliers* espacio-temporales en grandes bases de datos. Ellos son: agrupación o *clustering*, comprobación espacial de vecinos y comprobación temporal de vecinos. El *clustering* es una forma de detectar posibles *outliers*, siendo éstos los objetos no situados dentro de ningún grupo. El algoritmo de *clustering* satisface tres requisitos importantes: 1) realizar concentraciones de forma arbitraria; 2) tener un buen rendimiento en las grandes bases de datos y 3) contar con algunas heurísticas para determinar los parámetros de entrada. Tras ejecutar el algoritmo (que aparece en la Figura 5), pueden aparecer puntos que se marquen como *outliers*, si no aparecen al algoritmo no detectaría ningún *outlier* con este procedimiento.

En la segunda fase, durante la detección de vecinos espaciales, se comprueban tanto los *outliers* detectados, como los grupos diferenciados en el paso anterior, utilizando la definición 1.

Definición 1: dada una base de datos con n objetos *D={o1,o2,…,on}*, se supone que el objeto *o* se detecta como un *outlier* potencial en la agrupación. El valor medio de los vecinos espaciales de *o* dentro de un radio *Eps1* se define como:

$$A \stackrel{\text{def}}{\equiv} \frac{o_{\text{neigh.1}} + o_{\text{neigh.2}} + \ldots + o_{\text{neigh.m}}}{m} \qquad (2)$$

```
Algorithm ST_Outlier_Detection(D,Eps1,Eps2,MinPts,Δε)

 // Clustering Part
 Cluster_Label = 0

 For i=1 to n                          //(i)
  If o  is not in a cluster Then       //(ii)
   X=Retrieve_Neighbors(o ,Eps1,Eps2)//(iii)

  If |X| < MinPts Then
   Mark o  as outlier                  //(iv)
  Else       //construct a new cluster (v)
   Cluster_Label = Cluster_Label + 1

   For j=1 to |X|                      //(vi)
    Mark all objects in X with current Cluster_Label
   End For

   Push(all objects in X)              //(vii)

   While not IsEpmty()
    CurrentObj = Pop()
    Y= Retrieve_Neighbors(CurrentObj, Eps1, Eps2)

    If |Y| >= MinPts Then
     ForAll objects o in Y             //(viii)
      If (o is not marked as outlier or
          it is not in a cluster) and
         |Cluster_Avg() - o.Value| <= Δε Then
       Mark o with current Cluster_Label
       Push(o)
      End If
     End For
    End If
   End While
  End If
 End If
End For

 Checking_Spatial_Neighbors()
 Checking_Temporal_Neighbors()
End Algorithm
```

Figura 5. Algoritmo de detección de *outliers.*
Fuente: Birant *et al.*, 2006.

Donde *m* es el número de vecinos espaciales dentro de un radio *Eps1* y la desviación estándar para los objetos se define como la raíz de *V*, donde

$$V \stackrel{\text{def}}{=} \frac{(o_{\text{neigh1}}-A)^2 + (o_{\text{neigh2}}-A)^2 + \ldots + (o_{\text{neigh m}}-A)^2}{m} \qquad (3)$$

El objetivo *o* es clasificado como *outlier* si está dentro del intervalo [L,U], donde $L=A-k_0\cdot\sigma$, $U=A+k_0\cdot\sigma$, y $k_0>1$ es algún valor preseleccionado.

El tercer paso es la comprobación de los vecinos temporales de los *outliers* detectados en el paso anterior. Dos objetos serán vecinos temporales si sus unidades de tiempo son consecutivas, como días consecutivos en el mismo año o en el mismo día en años consecutivos. Para apoyar los aspectos temporales, los *outliers* se comparan con otros objetos de la misma área, pero en diferentes momentos, filtrando primero sólo los vecinos temporales y sus correspondientes valores. Si el valor marcado como *outlier* no tienen diferencias significativas con sus vecinos temporales, se descarta que sea un *outlier*. En caso contrario se confirma que es un *outlier*.

La fórmula utilizada es similar a la expresada en la definición 1, utilizando los vecinos temporales en vez de los vecinos espaciales.

Para probar el método, los autores utilizaron datos de altura de ola entre 1992 y 2002 de cuatro mares: el Mar Negro, el Mar de Mármara, el Mar Egeo y el este del Mar Mediterráneo. Aplicando el proceso resumido anteriormente, se obtuvo una región con altura de ola extrema en 1998 (Figura 6). Al compararlo con otros años, se comprobó que los valores de altura de olas registrados habitualmente en esa región no eran demasiado altos, por lo que se verificó esa región como *outlier*.

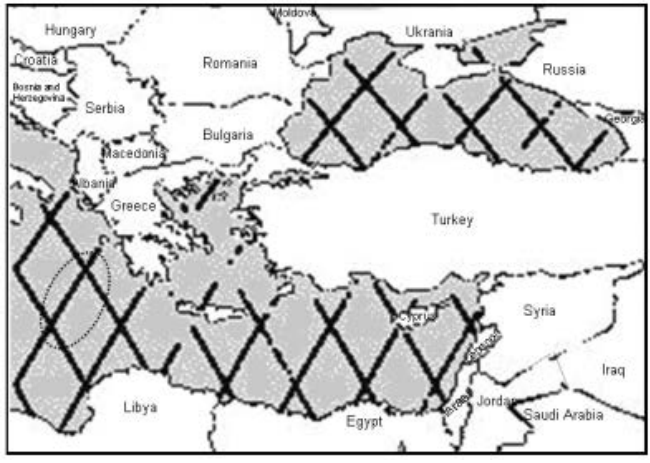

Figura 6. La región marcada con un óvalo punteado se corresponde con el área detectada con altura de ola extrema en 1998.
Fuente: Birant *et al.*, 2006.

TOD (Temporal Outlier Detection)

Es un método basado en reglas de asociación (técnica de minería de datos utilizada para la detección de patrones locales en los sistemas de aprendizaje no supervisado) que deducen el comportamiento normal de los objetos mediante la extracción de reglas de frecuencia en un conjunto de datos (Bruno y Garza, 2010). Para incluir la información del tiempo, se define el concepto de reglas de asociación temporal (extensión de las reglas de asociación que consideran el tiempo de retraso entre el antecesor y el sucesor) que se combinan para generar dependencias cuasi-funcionales temporales para definir las relaciones de los atributos con el tiempo. Dada una dependencia cuasi-funcional temporal, es posible descubrir los *outliers* mediante la consulta de las reglas de asociación temporal almacenadas previamente.

El método deduce las reglas para detectar los *outliers* directamente de los datos. La estructura principal del método TOD se describe en el algoritmo 1 (Figura 7):

Algorihtm 1. TOD: temporal outlier detection algorithm

Input: temporal dataset *K*, sliding window size *w*, length of extracted quasi-functional dependency *l*, minimum dependency degree value *Pthreshold*
Output: set of outliers *O*
1: /* *Extract temporal association rules of length l and discover l-length temporal quasi-functional dependencies analyzing extracted association rules* */
2: *TARs* = mine_temporal_association_rules(*K*,*w*,*l*)
3: *TQFDs* = mine_temporal_quasi-functional_dependencies(*TARs*,*Pthreshold*,*l*)
4: /* *Highlight outliers by considering mined temporal quasi-functional dependencies* */
5: **for all** *t* in *TQFDs* **do**
6: *LowConfAssRules* = select_low_confidence_rules(*t*)
7: *O* = *O* ⋃ matching_data(*K*,*LowConfAssRules*)
8: **end for**
9: **return** *O*

Figura 7. Algoritmo con la estructura principal del método TOD.
Fuente: Bruno y Garza, 2010.

Primero se extrae el conjunto de reglas de asociación temporales de la base de datos temporal denominada *k* con un tamaño de ventana deslizante *w* y una longitud máxima de *l* (algoritmo 1, línea 2, Figura 7). Después, se utilizan las reglas asociación de minearía, para extraer las dependencias cuasi-funcionales temporales (algoritmo 1, línea 3, Figura 7). Cada una de estas dependencias extraídas se utiliza para descartar un conjunto de *outliers* (algoritmo1, líneas 5-8, Figura 7). En particular, para cada dependencia, se extraen las reglas de asociación temporales que representan los valores extremos (algoritmo 1, línea 6, Figura 7). A continuación, se realiza un análisis del conjunto de datos que corresponden a una de las reglas seleccionadas se incluyen en el conjunto de *outliers* (algoritmo 1, línea 7, Figura 7)

El algoritmo 2 (Figura 8) muestra con más detalle cómo por cada dependencia cuasi-funcional temporal, se selecciona el conjunto de reglas asociado a los *outliers*.

Algorithm 2. select_low_confidence_rules

Input: temporal quasi-functional dependency *t*
Output: low confidence association rules AR_{low_conf}
1: $AR_{low\_conf} = \emptyset$
2: /* *Create one group of rules for each antecedent* */
3: *rule_groups*=group by antecedent the rules in *t.AssociationRules*. Create one group for each possible antecedent.
4: **for all** group *g* in *rule_groups* **do**
5: $max\_confidence = Max_{\forall r \in g}(r.confidence)$
6: **for all** rule $r \in g$ **do**
7: **if** *r.confidence* < *max_confidence* **then**
8: $AR_{low\_conf} = AR_{low\_conf} \bigcup r$
9: **end if**
10: **end for**
11: **end for**
12: **return** AR_{low}

Figura 8. Algoritmo 2.
Fuente: Bruno y Garza, 2010.

En su trabajo, los autores ensayan el método con datos de mercado que incluye información temporal de los datos históricos diarios de los precios de las acciones. Los experimentos se realizaron en una serie de conjuntos de datos generados por la variación del número de días y fijando uno a uno el número de acciones. Antes de extraer la reglas de asociación, todos los atributos continuos del conjunto de datos de tiempo se discretizan mediante distintos algoritmos.

Para analizar en detalle la escalabilidad de método TOD, se realizan experimentos variando tanto las características del número de datos (el número de registros y el número de atributos), como los valores de los parámetros de TOD (el tamaño de la ventana deslizante y la longitud de la pauta). En concreto se realiza el tiempo de ejecución cuando varía: 1) el número de registros; 2) el número de atributos; 3) el tamaño de la ventana deslizante y 4) la longitud de dependencia.

Con ello se demuestra la eficacia del método TOD y, en particular, el rendimiento de TOD al variar el grado de dependencia. Los resultados expuestos por los autores no son precisamente sintéticos, ya que van comprobando en cuatro apartados las variaciones indicadas en el párrafo anterior, exponiendo multitud de tablas y gráficas. Por ello, y para no alargar aún más el artículo, únicamente se enumeran las pruebas que realizan y la conclusión final.

Modelo de detección de outliers de forma global

Es una técnica para detectar *outliers* basada en una variante del algoritmo EM (algoritmo de esperanza-maximación citado en Yang *et al.*, 2009). El enfoque no hace ninguna suposición sobre la distribución de datos y es supervisado. Utiliza el *outlier factor*:

$$OF_k = \frac{1}{F_k} \tag{4}$$

Siendo F_k la suma de los productos de:

$$F_k = z_k(t_h) = \sum_{j=1}^{n} s_{kj}\pi_j(t_h) \tag{5}$$

Donde $s_{kj}\pi_j(t)$ representa como de fuerte es la influencia del punto x_k que es influido por el punto x_j siendo s_{kj} la fuerza de conexión y $\pi_j(t)$ la medición de la importancia de j.

El *outlier factor* que proponen se basa en las propiedades globales del conjunto de datos, lo que contrasta con la mayoría de los enfoques existentes, que lo hacen de forma local.

Para ilustrar las ventajas del algoritmo propuesto frente a otros, se tomó un conjunto de datos (Figura 9) que contiene 41 puntos espaciados en el cluster1, 104 puntos densos en el cluster2 y cuatro valores extremos A,B,C y D (marcados en rojo). Los métodos con los que se compara son COF (detalles en Tang, *et al.*, 2002), LOF (detalles en Breunig *et al.*, 2000) y LOCI (detalles en Papadimitriou *et al.*, 2003). Los métodos LOF y COF, no fueron capaces de detectar los *outliers* A y B, mientras que el tercer método (LOCI) si fue capaz de encontrar los cuatro puntos, al igual que el método propuesto.

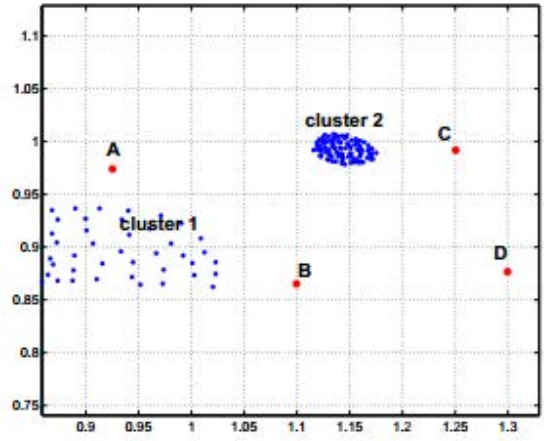

Figura 9. Primer grupo de datos para demostrar las ventajas del método. Los *outliers* son los puntos en rojo (A, B, C y D).
Fuente: Yang, *et al.*, 2009.

El segundo grupo de datos utilizado como ejemplo se obtuvo mediante la digitalización de unas imágenes de microbiología de una lámina de silicio. El objetivo es detectar y eliminar los *outliers* con el fin de obtener una aproximación paramétrica apropiada de la curva representada. La imagen contiene tres *outliers*, los puntos a, b y c representados en la Figura 10.

Para este segundo grupo de datos se aplicaron los métodos LOF, COF y LOCI, sin obtener los resultados esperados, los que si se obtienen con el método propuesto por los autores.

Los autores también emplearon para probar su método conjuntos de datos anteriormente utilizados por otros investigadores. Los resultados del método propuesto en comparación con otros métodos se muestran en la
Figura 11. La mejora para la detección con los datos COIL 200 resulta obvia, pero todavía queda por debajo del 60% (Tabla 1). El principal problema parece ser el número de atributos. Pasa algo similar con el conjunto de datos Satimage. Las mejoras más significativas se obtienen en la detección de *outliers* de las mamografías y en las

azoteas (como se pude ver en la Figura 11 es mayor el área bajo la curva negra que se corresponde al método de estudio, lo que indica que hay mayor cantidad de verdaderos positivos que con los otros métodos comparados).

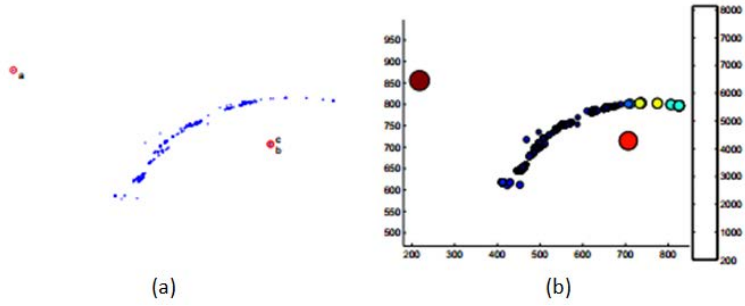

Figura 10. Segundo grupo de datos. (a) Los puntos a, b y c representan *outliers* (los autores no dan información de los ejes). (b) Los resultados obtenidos por el método propuesto, (*outliers* en tonos rojos).
Fuente: Yang *et al*., 2009.

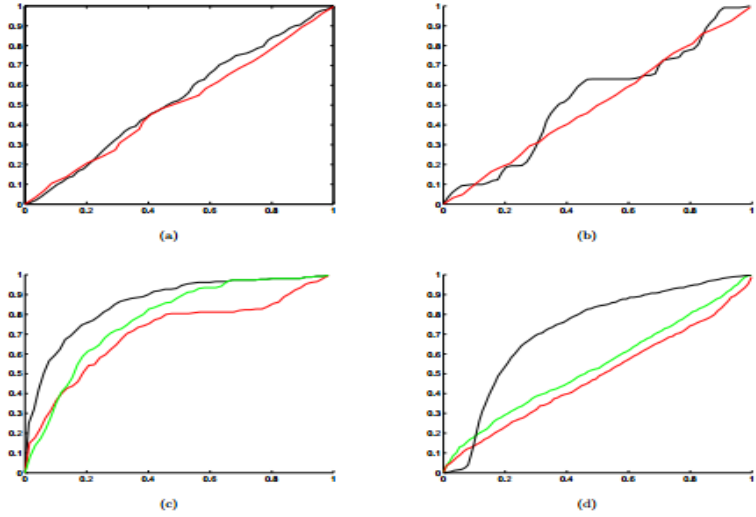

Figura 11. Curvas ROC (que representan los verdaderos positivos en ordenadas y los falsos positivos en abscisas) para la comparación de los resultados de los tres métodos. La curva negra representa el método prepuesto, la curva verde representa el método COF y la roja el método LOF. Para las bases de datos (a) COIL 200, (b) Satimage, (c) mamografías (Mamography), y (d) azoteas (rooftop). Cuanto más alta va la curva, mayor cantidad de verdaderos positivos obtiene el método, por lo que sus resultados son mejores. Fuente: Yang *et al*., 2009.

Tabla 1
Áreas debajo de las curvas (AUC) de la Figura 11
Cuanto mayor sea el área, más efectivo es el método, ya que hay mayor cantidad de verdaderos positivos

	AUC		
Data sets	*LOF*	*COF*	*Proposed Approach*
COIL200	0.499	0.505	0.529
Satimage	0.497	0.503	0.533
Mamography	0.710	0.780	0.862
Rooftop	0.538	0.498	0.722

Fuente: Yang *et al.*, 2009.

Como se ha visto, este método proporciona buenos resultados tanto para datos sintéticos como para datos usados por otros autores.

Algoritmos de detección de outliers por ponderación espacial

Los autores (Yufeng *et al.*, 2006) proponen dos algoritmos de detección de *outliers* que utilizan propiedades espaciales para representar el impacto de los vecinos. Se basan en que la mayoría de los algoritmos de detección de *outliers* espaciales se fundamentan únicamente en los atributos no espaciales del objeto estudiado y de sus vecinos. Pero en muchas aplicaciones, los objetos espaciales no pueden simplemente abstraerse como puntos aislados, ya que tienen otras propiedades como diferente ubicación, área, curvas de nivel, volumen, etc. que desempeñan papeles importantes a la hora de determinar el impacto espacial y no deben ser ignorados. Por ello proponen un método de detección de *outliers* que asigna diferentes pesos a los diferentes vecinos, determinando ese peso según las relaciones espaciales tales como la distancia.

El primer algoritmo propuesto (Algoritmo 1, Figura 12), llamado "enfoque del valor de *z* ponderado" tiene cuatro parámetros de entrada. *X* es un conjunto de *n* objetos que contienen atributos espaciales, tales como la ubicación, límites y área. Los atributos no espaciales están contenidos en *Y*. *k* es un valor fijo para cada objeto, que puede generalizarse sustituyéndolo por un *k* dinámico. Y *m* es el número de *outliers* solicitados, que generalmente no debe ser superior al 5 %.

El segundo algoritmo propuesto (algoritmo 2, Figura 12), el "algoritmo de diferencia promedio" es una variante del primer algoritmo. A diferencia él, se basa en la media ponderada de la diferencia absoluta entre x_i y cada uno de sus vecinos. La idea principal es comparar un objeto con cada uno de sus vecinos, uno por uno, en lugar de la obtención de la media de todos sus vecinos antes de la comparación. La razón radica en que el promedio simple de los vecinos ignora su varianza.

Algorithm 1 : Weighted z value approach

Input:
X is a set of n spatial objects;
Y is the set of attribute values for X;
k is the number of neighbors;
m is the number of requested spatial outliers;
Output:
O_s is a set of m outliers

for(i=1; i ≤ n ; i++) {
/* *calculate the neighbor hood relationship* */
$NN_k(x_i)$ = GetNeighbors(X, x_i);
/* calculate the weighted average of all x_i's neighbors */
$NbrAvg(x_i) = 0$;
for each $x_j \in NN_k(x_i)$ {
$weight$ = getWeight($NN_k(x_i)$, x_j)
$NbrAvg(x_i) = NbrAvg(x_i) + y_j * weight$
}
$Diff(x_i) = y_i - NbrAvg(x_i)$
}
/* calculate the standardized $Diff(x_i)$ as the outlierness factor */
$\mu = getMean(Diff)$
$\sigma = getStd(Diff)$
for (each $x_i \in X$) {
$OF(x_i) = |\frac{Diff(x_i)-\mu}{\sigma}|$
}
O_s = getTopMOutliers(OF, m)

Algorithm 2 : AvgDiff Algorithm

Input:
X is a set of n spatial objects;
Y is the set of attribute values for X;
k is the number of neighbors;
m is a number of requested spatial outliers;
Output:
O_s is the set of m outliers

for(i=1; i ≤ n ;i++) {
/* *calculate the neighbor hood relationship* */
$NN_k(x_i)$ = GetNeighbors(X, x_i);
/* calculate the weighted average difference between x_i and */
/* its neighbors */
$AvgDiff(x_i) = 0$;
for each $x_j \in NN_k(x_i)$ {
$diff = |y_i - y_j|$
$weight$ = getWeight($NN_k(x_i)$, x_j)
$AvgDiff(x_i) = AvgDiff(x_i) + diff * weight$
}
}
for (each $x_i \in X$) {
$OF(x_i) = AvgDiff(x_i)$
}
O_s = getTopMOutliers(OF, m)

Figura 12. Algoritmos 1 o "enfoque del valor de z ponderado" y algoritmo 2 o "diferencia promedio".
Fuente: Yufeng *et al.,* 2006.

Los experimentos se llevaron a cabo con datos reales sobre el virus del Nilo Occidental proporcionados por los Centros para el Control y Prevención de Enfermedades (CDC). El conjunto de datos incluye el registro del número de casos de infección del virus en aves silvestre, mosquitos y los casos veterinarios a nivel de condado de Estados Unidos entre enero de 2002 y diciembre de 2003.

Tras aplicar los algoritmos se detectan 30 *outliers* principales, que representan alrededor del 1% del total de los 3 109 condados. Ambos algoritmos tienen resultados similares, identificando los mismos 22 condados en los 30 principales *outliers*. La variación de la clasificación (en los ocho condados no coincidentes en ambos algoritmos) es causada por los diferentes mecanismos utilizados para calcular la diferencia promedio de la zona.

Conclusiones

Los trabajos presentados confirman el intenso interés en el desarrollo de métodos de detección de *outliers*. Las principales características de estos métodos recientes son:

Modelo *STARIMA*: utilizado para detectar *outliers* en series de datos espacio-temporales. Una de las ventajas del modelo STARIMA es que su definición se basa en resultados empíricos. Además ofrece buenos resultados para observaciones a largo plazo. Una desventaja es que las distintas densidades pueden sesgar la desvia-

ción estándar y la media, por lo que el modelo puede clasificar como *outliers* objetos que no debería detectar. Esto hace que el modelo pueda ser poco robusto.

Método SWOD: utilizado para detectar *outliers* espacio-temporales. Según los resultados experimentales, es más preciso que otros métodos lo que deja ver la eficiencia de analizar conjuntamente el espacio y el tiempo para detectar *outliers*. Como desventajas, cabe reseñar que los autores no ofrecen los resultados en casos verdaderos al no conocer la realidad sobre del terreno de los datos utilizados, por lo que necesita ser evaluado para comprobar si sus resultados son realmente tan precisos como indican los autores.

ST-Outlier Detection Algorithm: utilizado para detectar *outliers* en grandes bases de datos espacio-temporales. Con el ejemplo propuesto por los autores, destaca el hecho de que detecta toda un área como *outlier*, y no una serie de puntos como suele pasar, por lo que este método necesita ser probado con más bases de datos para comprobar si los resultados son realmente tan buenos como indican los autores.

Método TOD (Temporal Outlier Detection): utilizado para detectar *outliers* mediante técnicas de minería de datos con aprendizaje no supervisado aplicado a datos temporales. Según lo indicado por los autores, es un método eficaz, siendo particularmente alto el rendimiento al variar el grado de dependencia, como muestran los autores en sus aplicaciones prácticas. El problema del modelo es que está formulado para datos temporales y no tienen en cuenta el espacio.

Modelo de detección de *outliers* de forma global: este método es utilizado para detectar outliers basándose en las propiedades globales del conjunto de datos. Proporciona buenos resultados para datos sintéticos y también para datos ampliamente usados por los investigadores en este campo, mejorando sus resultados.

Algoritmos de detección de *outliers* por ponderación espacial: utilizado para detectar outliers teniendo en cuenta la disposición espacial de los datos. Su idea principal es realizar un filtrado de los datos para extraer un conjunto de *outliers* candidatos a ser investigados por expertos, pudiéndolos ofrecer dando un orden de precedencia. Esta es una ventaja muy interesante a la hora de decidir qué datos se van a revisar y con qué preferencia.

Este trabajo pone de manifiesto que existe un creciente interés por el desarrollo de nuevo métodos de detección de *outliers*, que cada vez aportan mejores resultados cuando son aplicados sobre el tipo de datos para los que fueron desarrollados. Pero aún queda mucho por investigar, pues no existe un método óptimo de detección de *outliers*. Únicamente existen métodos que los detectan con mayor o menor efectividad.

Se estima indispensable continuar avanzando con el desarrollo de los métodos expuestos de cara a optimizar los resultados que proporcionan y facilitar su aplicación en los conjuntos de datos para los que resultan aptos.

Bibliografía

Angiulli, F.; Pizzuti, C., (2002). "Fast outlier detection in high dimensional spaces", Proc. PKDD, pp. 15-26.

Appice, A.; Guccione, P.; Malerba, D. and Ciampi, A., (2014). "Dealing with temporal and spatial correlations to classify outliers in geophysical data streams", Information Sciences, pp. 1-19.

Ben-Gal, I., (2005). "Outlier detection", Data Mining and Knowledge Discovery Handbook. Springer US, pp. 131-146.

Birant, B. and Kut, A., (2006). "Spatio-Temporal Outlier Detection in Large Databases", CTI. Journal of Computing and Information Technology, 14 (4), pp. 292-297.

Breunig, M.; Kriegel, H.; Ng, R. and Sander, J., (2000). "LOF: Identifying density-based local outliers", Proc. SIGMOD, pp. 93-104.

Bruno, G. and Garza, P., (2010). "TOD: Temporal outlier detection by using quasi-functional temporal dependencies", Data & Knowledge Engineering, 69 (6), June, pp. 619-639.

Breunig, M.; Kriegel, H.; Ng, R. and Sander, J., (2000). "LOF: Identifying Density Based Local Outliers", The ACM SIGMOD Conference, pp. 27-39.

Cheng, T.; Anbaroğlu, B., (2009). "Spatio-Temporal Outlier Detection in Environmental Data", International Conference on Spatial Information Theory Aber Wrac'h, France, pp. 17-24.

Hadi, A.; Rahmatullah, A. and Werner, M., (2009). Detection of outliers. WIREs Comp. Stat., 1(1), pp. 57-70.

Keller, F.; Müller, E. and Böhm, K., (2012). "HiCS: high contrast subspaces for density-based outlier ranking", Proc. ICDE, pp. 83-92.

Knorr, E.; Ng, R. and Tucanov, V., (2000). "Distancebased outliers: Algorithms and applications", VLDB J., 8 (3-4), pp. 237-253.

Kriegel, H.; Kröger, P.; Schubert, E. and Zimek, A., (2009). "LoOP: local outlier probabilities", Proc. CIKM, pp. 1649-1652.

Orair, G.; Teixeira, C.; Wang, Y.; Meira W. and Parthasarathy, S., (2010). "Distance-based outlier detection: Consolidation and renewed bearing", PVLDB, pp. 1469-1480.

Papadimitriou, S.; Kitagawa, H.; Gibbons, P. and Faloutsos, C., (2003). LOCI: Fast Outlier Detection Using the Local Correlation Integral, ICDE, pp. 237-243.

Pechenizkiy, M.; Bakker, J.; Zliobaite, I.; Ivannikov, A. and Kärkkäinen, T., (2009). "Online mass flow prediction in cfb boilers with explicit detection of sudden concept drift", ACM SIGKDD Explorations Newsletter, pp. 109-116.

Ramaswamy, S.; Rastogi, S. and K. Shim, K., (2000). "Efficient algorithms for mining outliers from large data sets", Proc. SIGMOD, pp. 427-438.

Ramanathan, N.; Balzano, L.; Burt, M.; Estrin, D.; Kohler, E.; Harmon, T.; Harvey, C.; Jay, J.; Rothenberg, S. and Srivastava, M., (2006). "Rapid Deployment with Confidence: Calibration and Fault Detection in Environmental Sensor Networks", Technical Report CENS, pp.119-125.

Schuber, E.; Wojdanowski, R.; Zimek, A. and Kriegel, H.P., (2012). "On Evaluation of Outlier Rankings and Outlier Scores", 2012 SIAM International Conference on Data Mining, pp. 1047-1058.

Sharma, A.B.; Golubchik, L. and Govindan, R. (2010). Sensor faults: detection methods and prevalence in real-world datasets, ACM Trans. Sensor Netw. 6, pp. 1-39.

Tang, J.; Chen, Z.; Fu, A. and Cheung, D., (2002). "Enhancing Efectiveness of Outlier Detections for Low Density Patterns", Advances in Knowledge Discovery and Data Mining, pp. 535-548.

Vu N. and V. Gopalkrishnan, (2009). "Efficient pruning schemes for distance-based outlier detection", Proc. ECML PKDD, pp. 160-175.

Vries, T.; Chawla, S. and Houle, M., (2010). "Finding local anomalies in very high dimensional space", Proc. ICDM, pp. 128-137.

Yang, X.; Latecki, L.J.; Pokrajac, D. (2009). "Outlier Detection with Globally Optimal Exemplar-Based GMM", 2009 SIAM International Conference on Data Mining, pp. 145-154.

Yao, Y.; Sharma, A.; Golubchik, L. and Govindan, R., (2010). "Online anomaly detection for sensor systems: a simple and efficient approach", Performance Evaluation, 67, pp. 1059-1075.

Yufeng, K.; Chang-Tien, L. and Dechang, C., (2006). "Spatial Weighted Outlier Detection", The 2006 SIAM International Conference on Data Mining, pp. 614-618.

Zhang, K.; Hutter, M. and Jin, H., (2009). "A new local distancebased outlier detection approach for scattered realworld data", Proc. PAKDD, pp. 813-822.